UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Mining user behavior in location-based social networks

Jorge Carlos Valverde Rebaza

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Jorge Carlos Valverde Rebaza

Mining user behavior in location-based social networks

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Alneu de Andrade Lopes

USP – São Carlos
September 2017

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

V135m

VALVERDE REBAZA, JORGE CARLOS
Mining user behavior in location-based social networks / JORGE CARLOS VALVERDE REBAZA; orientador ALNEU DE ANDRADE LOPES. -- São Carlos, 2017.
221 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2017.

1. COMPUTAÇÃO APLICADA. 2. REDES COMPLEXAS. 3. REDES SOCIAIS. I. DE ANDRADE LOPES, ALNEU, orient. II. Título.

Jorge Carlos Valverde Rebaza

Mineração do comportamento de usuários em redes sociais baseadas em localização

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Alneu de Andrade Lopes

USP – São Carlos
Setembro de 2017

Dedicado a mis padres, Jorge y Nidia,
a mis hermanos, Cleydee, Joel y Luis,
y a mi amada Nathalia.

ACKNOWLEDGEMENTS

First of all, I would like to thank God for blessing me along this hard journey. I am extremely grateful to my parents, Jorge and Nidia, whom not only supported me by all means but also encouraged me to chase my dreams even at the expense of being far away from them. I would like to thank my sister, Cleydee, and my brothers, Joel and Luis, for all their love and companionship. It is an honour to be part of this family.

My special thanks to my sweet love, Nathalia, who I discovered when I was wandering around the urban jungle of São Carlos sometime in the middle of my PhD. She is by my side since then. Her strength and patience have guided me during this journey, her smile and loving embrace make my days much happier. I am very lucky to have her love.

I would like to express my deep gratitude to my advisor, Prof. Alneu de Andrade Lopes, for his continuous support, motivation, and knowledge. During these years he has patiently monitored and directed my work, allowing for the expression and development of my skills. His experience in research, and mainly his friendship, helped me to I become a researcher with the ability to navigate in the landscape of academia in a manner that I love and appreciate. Thank you so much.

I owe thanks for the good moments and the collaboration in this work to Prof. Dra. Maria Carolina Monard, Alan Valejo, Newton Spolaôr, Thiago Faleiros, Merley Conrado, Roberta Sinoara, Ricardo Puma, Diego Minatel, João Antunes, Vinícius Ferreira, Vinícius de Souza, Rafael Giusti, Brett Drury, Rafael Rossi, Lilian Berton, Diego Silva and the rest of past and present members at LABIC who have made the lab not only a good place to work, but also a fun place to be.

I was also lucky to have received great support from people outside Brazil. I want to thank Dr. Alípio Jorge for receipt me at LIAAD, in Porto (Portugal), and to Dr. Pascal Poncelet and Dr. Mathieu Roche for opened the doors for me at LIRMM, in Montpellier (France). I am also especially grateful to Dr. Antonio Lossio, by his friendship, support, collaboration and good times in the different places where the live took us.

Finally, I would like to thank to Instituto de Ciências Matemáticas e de Computação (ICMC) - University of São Paulo (Brazil), University of Porto (Portugal), and University of Montpellier (France), for providing academic structures that enabled the development of this work. Also, I would like to acknowledge the financial support from the National Council for Scientific and Technological Development (CNPq), grant No $151836/2013-2$, and from São Paulo Research Foundation (FAPESP), grant No $2013/12191-5$.

“Science has not yet taught us if madness is
or is not the sublimity of the intelligence.”
Edgar Allan Poe

RESUMO

VALVERDE-REBAZA, J. C. **Mineração do comportamento de usuários em redes sociais baseadas em localização**. 2017. 221 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Redes sociais *online* (OSNs) são plataformas Web que oferecem serviços para promoção da interação social entre usuários. OSNs que adicionam serviços relacionados à geolocalização são chamadas redes sociais baseadas em localização (LBSNs). Um dos maiores desafios na análise de LBSNs é a predição de links. A predição de links refere-se ao problema de estimar a probabilidade de conexão futura entre pares de usuários que não se conhecem. Grande parte das pesquisas que focam nesse problema exploram o uso, de maneira isolada, de informações sociais (e.g. amigos em comum) ou de localização (e.g. locais comuns visitados). Porém, algumas pesquisas mostraram que a combinação de diferentes fontes de informação pode influenciar o incremento da acurácia da predição. Motivado por essa lacuna, neste trabalho foram desenvolvidos diferentes métodos para predição de links combinando diferentes fontes de informação. Assim, propomos sete métodos que usam a informação relacionada à participação simultânea de usuários en múltiples grupos sociais: *common neighbors within and outside of common groups* (WOCG), *common neighbors of groups* (CNG), *common neighbors with total and partial overlapping of groups* (TPOG), *group naïve Bayes* (GNB), *group naïve Bayes of common neighbors* (GNB-CN), *group naïve Bayes of Adamic-Adar* (GNB-AA), e *group naïve Bayes of Resource Allocation* (GNB-RA). Devido ao fato que a presença de grupos sociais não está restrita a alguns tipo de redes, essas propostas podem ser usadas nas diversas OSNs existentes, incluindo LBSNs. Também, propomos oito métodos que combinam o uso de informações sociais e de localização: *Check-in Observation* (ChO), *Check-in Allocation* (ChA), *Within and Outside of Common Places* (WOCP), *Common Neighbors of Places* (CNP), *Total and Partial Overlapping of Places* (TPOP), *Friend Allocation Within Common Places* (FAW), *Common Neighbors of Nearby Places* (CNNP), e *Nearby Distance Allocation* (NDA). Tais propostas são para uso exclusivo em LBSNs. Os resultados obtidos indicam que nossas propostas são tão competitivas quanto métodos do estado da arte, podendo até superá-los em determinados cenários. Ainda mais, devido a que na maioria dos casos nossas propostas são computacionalmente mais eficientes, seu uso resulta mais adequado em aplicações do mundo real.

Palavras-chave: Redes Sociais, Redes Sociais baseadas em Localização, Predição de Links, Recomendação de Amizade, Análise do Comportamento de Usuários.

ABSTRACT

VALVERDE-REBAZA, J. C. **Mining user behavior in location-based social networks**. 2017. 221 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Online social networks (OSNs) are Web platforms providing different services to facilitate social interaction among their users. A particular kind of OSNs is the location-based social network (LBSN), which adds services based on location. One of the most important challenges in LBSNs is the link prediction problem. Link prediction problem aims to estimate the likelihood of the existence of future friendships among user pairs. Most of the existing studies in link prediction focus on the use of a single information source to perform predictions, i.e. only social information (e.g. social neighborhood) or only location information (e.g. common visited places). However, some researches have shown that the combination of different information sources can lead to more accurate predictions. In this sense, in this thesis we propose different link prediction methods based on the use of different information sources naturally existing in these networks. Thus, we propose seven new link prediction methods using the information related to user membership in social overlapping groups: *common neighbors within and outside of common groups* (WOCG), *common neighbors of groups* (CNG), *common neighbors with total and partial overlapping of groups* (TPOG), *group naïve Bayes* (GNB), *group naïve Bayes of common neighbors* (GNB-CN), *group naïve Bayes of Adamic-Adar* (GNB-AA) and *group naïve Bayes of Resource Allocation* (GNB-RA). Due to that social groups exist naturally in networks, our proposals can be used in any type of OSN. We also propose new eight link prediction methods combining location and social information: *Check-in Observation* (ChO), *Check-in Allocation* (ChA), *Within and Outside of Common Places* (WOCP), *Common Neighbors of Places* (CNP), *Total and Partial Overlapping of Places* (TPOP), *Friend Allocation Within Common Places* (FAW), *Common Neighbors of Nearby Places* (CNNP) and *Nearby Distance Allocation* (NDA). These eight methods are exclusively for work in LBSNs. Obtained results indicate that our proposals are as competitive as state-of-the-art methods, or better than they in certain scenarios. Moreover, since our proposals tend to be computationally more efficient, they are more suitable for real-world applications.

Keywords: Social Networks, Location-based Social Networks, Link Prediction, Friendship Recommendation, User Behavior Analysis.

LIST OF FIGURES

LIST OF CHARTS

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

Many of large-scale natural and man-made phenomena are characterized by having a complex structure. Elements or components from these structures are related to each other in a non-random way. The capacity of recognise and use patterns is inherent of living organisms but notable efforts have been displayed by the scientific community to provide efficient frameworks to extract and quantify reproducible regularities of phenomena (HUETTEL; MACK; MCCARTHY, 2002; NEWMAN, 2010).

Methods to extract patterns from complex data have a long history. The first data analysis efforts have been performed via direct data manipulation, i.e. by the manual application of some techniques, such as Bayes' theorem (1700s) and regression analysis (1800s), over data. In this context, the study of networks began with the foundations of the *graph theory*, when Leonhard Euler in 1736 proposed its negative resolution to the seven bridges of Königsberg problem. However, the advance and popularization of computer technology over the past century has allowed the dramatically increased of data collection, storage, and manipulation ability. Moreover, some studies estimate that in 2020 the digital data size will contain about 40 trillion of gigabytes (GANTZ; REINSEL, 2012; ZANIN *et al.*, 2016).

With increasing data size, direct data analysis has been replaced progressively with indirect through automated data processing. Therefore, *data mining* (DM) was born aiming on apply automated data processing methods for recognition, analysis and modeling data patterns through a variety of simbolic and bio-inspired computational techniques, such as neural networks, cluster analysis, genetic algorithms, decision trees and decision rules, support vector machines, and deep learning. DM is an interdisciplinary subfield of Computer Science and its reach is not limited by the type (e.g. text, image and video) nor by size or complexity of data (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; HAND; SMYTH; MANNILA, 2001; BISHOP, 2006; ZAKI; JR., 2014). When the data processing implies a challenge in itself due to the volume of data, DM is called *big data*. In big data, a variety of efforts are performed to effectively

identify and analyze patterns from an absurdly large amount of data (MAYER-SCHÖNBERGER; CUKIER, 2013; MARR, 2015).

Curiously, in parallel to the appearance and growth of DM, have surged network studies with important discoveries. Some of the most highlighted discoveries related to the study of networks are: the analysis of *random networks* performed by Erdös and Rényi (1959), the *concept of separation in six degrees* performed by Milgram (1967) and which was the first seed for the study of the *small-world networks* formally defined later by Watts and Strogatz (1998), and the study of the *scale-free networks* by Barabási and Albert (1999). Despite the importance of these discoveries, it was only over the past two decades that the *complex network theory* has emerged as a strong discipline aiming to understand the structure and the different processes characterizing *complex systems*[1]. Due to that complex network theory drifting away from analyzing small graphs to large-scale graphs, it has received much attention in a wide range of disciplines from computer scientists, physicists, mathematicians, biologists, and others (COSTA *et al.*, 2007; NEWMAN, 2010; JAMAKOVIĆ, 2008; ZANIN *et al.*, 2016; SILVA; ZHAO, 2016).

Complex network analysis and data mining have similar objectives since both aim to extract (or synthesise) relevant information from data, corresponding to all or part of a complex system. Furthermore, both research issues focus their efforts to develop new models and methods capable of perform an adequate data analysis. Given the similarity in general purposes and sometimes even in procedures, it may be possible that both approaches can be used interchangeably and/or in combination (ZANIN *et al.*, 2016). However, traditional data processing methods assume that data instances are independent and identically distributed (IID); therefore, their naïve application over network data can lead to inappropriate analysis and conclusions (JENSEN, 1999).

To deal with this problem, a new research field called *link mining* (LM) emerged with the goal of finding patterns from complex network data. LM focuses on exploiting, and explicitly consider, the relationships (or links) naturally existing among the instances of network data (GETOOR; DIEHL, 2005). Moreover, LM encompasses a wide range of tasks, being the *link prediction* one of the most popular and multidisciplinary of them. Link prediction addresses the problem of predicting the likelihood of existence of missing or future links between disconnected instances, based on the known network information (LIBEN-NOWELL; KLEINBERG, 2007; LÜ; ZHOU, 2011; MARTÍNEZ; BERZAL; CUBERO, 2016).

Link prediction has a range of applications in different domains (SRINIVAS; MITRA, 2016). For instance, it is used in bioinformatics, to discover genetic or protein-protein interactions (KOTERA *et al.*, 2012; SINGH-BLOM *et al.*, 2013; MENCHE *et al.*, 2015); security, to identify groups of terrorist or criminals (HASAN *et al.*, 2006; CLAUSET; MOORE; NEWMAN,

[1] A complex system is composed by a large number of interacting components whose collective behavior can not be described as a simple combination of the behavior of each individual component (NEWMAN, 2003).

2008; ANIL *et al.*, 2015); and information retrieval, to predict words, topics or documents in very large collections of documents (ARNOLD; COHEN, 2009; ITAKURA *et al.*, 2011; LI *et al.*, 2016). However, despite the importance and attractiveness of these applications, the most popular are those implemented as services in online social networking sites, for instance, the recommender systems to suggest items to be purchased, new friends to meet, content to be appreciated, places to be visited, among others (LI; CHEN, 2009; CHILUKA; ANDRADE; POUWELSE, 2011; ESSLIMANI; BRUN; BOYER, 2011; WEI *et al.*, 2013; BAHABADI; GOLPAYEGANI; ESMAEILI, 2014; LI *et al.*, 2014; BARBIERI; BONCHI; MANCO, 2014; AHMED; ELKORANY, 2015; LIAO *et al.*, 2016; PEROZZI *et al.*, 2016; RAFAILIDIS; CRESTANI, 2016).

Online social networking sites, also called *online social networks* (OSNs), are web platforms implementing a social network structure made up of people connected by one or more specific types of interdependency, such as friendship, common interests, preferences, etc. The use of OSNs has increased in recent years, becoming part of the daily life of millions of people around the world (KUMAR; NOVAK; TOMKINS, 2006; MISLOVE, 2009; ESLAMI *et al.*, 2014b). Moreover, the increasing availability of positioning technology, e.g. GPS and Wi-fi, has led to some OSNs to implement different services based on location to attract more attention of users. OSNs offering services related to locations are called *location-based social network* (LBSN) (ZHENG; ZHOU, 2011; GRABOWICZ *et al.*, 2014).

In the context of OSNs, and consequently in LBSNs, the dynamics of the link creation process differs from that of other types of networks, mainly because users of these networks constantly establish new friendships. Therefore, the study of link creation process on these networks can lead us to, besides explain the network evolution process over time, better understand *user behavior* (KOSSINETS; WATTS, 2006; CRANSHAW *et al.*, 2010; ALLAMANIS; SCELLATO; MASCOLO, 2012; AHMED; ELKORANY, 2015).

In this scenario, link prediction algorithms emerge as an alternative for mining and analyzing user behavior. Furthermore, link prediction algorithms also can be useful for discovering new or hidden behaviors. Despite there are a variety of previous work focusing on use link prediction algorithms for mining and understanding user behavior (ESSLIMANI; BRUN; BOYER, 2011; CHO; MYERS; LESKOVEC, 2011; CHIKHAOUI *et al.*, 2014; ZOU; XIE; SHA, 2015; CHORLEY; WHITAKER; ALLEN, 2015), most of these studies do not use many of information sources naturally existing and available in conventional OSNs as well as in LBSNs.

1.1 Motivation

As previously mentioned, there are several studies showing that link prediction can be used as an effective framework for understanding user behavior. After analyzing different OSNs and LBSNs, these studies have shown the different behaviors that lead to a user to establish a new friendship with other user. For instance, the continuous interaction with common friends,

the similar preferences and the frequent visits at same places.

Considering this scenario, we have found two gaps in the current state-of-the-art research. The first gap is related to the link prediction problem in general. Several link prediction methods use a variety of information sources available in OSNs. Some of the information sources commonly used are: i) homophily (MCPHERSON; SMITH-LOVIN; COOK, 2001; CHANG *et al.*, 2014; FARALLI; STILO; VELARDI, 2015), ii) social ties (LÜ; ZHOU, 2010; SOCIEVOLE; RANGO; MARANO, 2013; BACKSTROM; KLEINBERG, 2014; XU *et al.*, 2017) and iii) communities (ZHELEVA *et al.*, 2008; SOUNDARAJAN; HOPCROFT, 2012; VALVERDE-REBAZA; LOPES, 2012a; HOSEINI; HASHEMI; HAMZEH, 2012; CANNISTRACI; ALANIS-LOBATO; RAVASI, 2013; KEMAL; TSUYOSHI *et al.*, 2014; MALLEK *et al.*, 2015; DAMINELLI *et al.*, 2015; WU *et al.*, 2016; DING *et al.*, 2016; MA *et al.*, 2016; KUANG; LIU; YU, 2016; CAIYAN; CHEN; LI, 2016; BISWAS; BISWAS, 2017). Because some researches have shown that *community information* can drastically improve the link prediction accuracy (FENG; ZHAO; XU, 2012; LIU *et al.*, 2013), most of the research efforts have been directed to better explore this information source. However, despite the accurate results obtained by link prediction methods using community information, there are no methods using *social group information*.

Social groups are entities naturally existing in OSNs, i.e. social groups are created by users, the same who voluntary and explicitly declare their membership to one or more of them. Therefore, social groups are a more sensible choice when compared to communities, since no additional cost is add to the algorithm to identify such communities. Moreover, the usage of social group information in link prediction opens new challenges. For instance, existing link prediction methods using community information consider the fact that a user belongs only to a single community, but in OSNs the users participate of one or more social groups; therefore, it will be necessary to redesign existing methods or to propose new ones. Moreover, due to the possibility that some groups are sharing the same users, link prediction methods using social group information should consider the presence of overlapping groups.

The second research gap is related specifically to the link prediction problem in LBSNs. Similar to conventional OSNs, in the literature also exists several link prediction methods to work on LBSNs (XIAO *et al.*, 2010; CRANSHAW *et al.*, 2010; CHO; MYERS; LESKOVEC, 2011; YU *et al.*, 2011; MENGSHOEL *et al.*, 2013; PHAM; SHAHABI; LIU, 2013; XIAO *et al.*, 2014; BAYRAK; POLAT, 2014; ZHANG; PANG, 2015; BAYRAK; POLAT, 2016; KYLASA; KOLLIAS; GRAMA, 2016). However, to the establishment of a future friendship between a pair of disconnected users, most of the existing methods are limited to the exclusive use of location information, e.g. frequency of visits at specific places, number of visits at similar places and geographical distance between frequently visited places. Therefore, to use other information sources and/or combine them with location information could be interesting to improve the link prediction accuracy in LBSNs.

In this work, both of the research gaps mentioned above are investigated. First, we investigate the challenges of link prediction using community information in real-world, large-scale social networks. Therefore, we propose a strategy combining a fast and efficient community detection algorithm with fast and accurate link prediction methods based on community information for friendship recommendation in a time frame comparable to state-of-the-art link prediction methods. After, we introduce the link prediction problem considering the participation of users in multiple social groups, as well as the presence of overlapping groups. For this, we propose new network properties to extract characteristics related to the presence of overlapping groups. Among these properties, we highlight the *overlapping groups clustering coefficient*. Based on these properties, we proposed seven new link prediction methods: *common neighbors within and outside of common groups* (WOCG), *common neighbors of groups* (CNG), *common neighbors with total and partial overlapping of groups* (TPOG), *group naïve Bayes* (GNB), *group naïve Bayes of common neighbors* (GNB-CN), *group naïve Bayes of Adamic-Adar* (GNB-AA) and *group naïve Bayes of Resource Allocation* (GNB-RA).

By establishing a parallel between the participation of users in groups and visitations of users to places, we combine location with social strength information to propose eight new friendship prediction methods for LBSNs: *Check-in Observation* (ChO), *Check-in Allocation* (ChA), *Within and Outside of Common Places* (WOCP), *Common Neighbors of Places* (CNP), *Total and Partial Overlapping of Places* (TPOP), *Friend Allocation Within Common Places* (FAW), *Common Neighbors of Nearby Places* (CNNP) and *Nearby Distance Allocation* (NDA).

1.2 Objectives

The main objective of this work is to contribute to better understanding of user behavior in LBSNs by exploring, mainly, the different ways used by the users to establish new friendships. This understanding can be provided by link prediction, a mining task capable of explain the likelihood of formation or dissolution of links between all node pairs in a network.

Given the research gaps mentioned above, the contribution are threefold:

- *Theoretical Development*. Contribute to the theoretical foundation in both complex networks and social network analysis fields. This objective can be achieved by formulate new network properties, new or more efficient notation, as well as by new findings related to ways in which users establish friendships. This objective also can be achieved by contributing with new surveys and categorizations of different aspects (e.g. problem statement, existing methods and challenges) related to the research topics addressed in this thesis.

- *Algorithmic Development*. Design new algorithms to efficiently explore the influence of user behavior in the establishment of new friendships in OSNs, including LBSNs. Such algorithms must consider different aspects of a variety of information sources (e.g. social

strength, social group membership, geographical distance and frequency of visitations at places). Moreover, the algorithms should take into account the restrictions related to the information sources used.

- *Application Development*. Evaluate and analyze the performance of developed algorithms in networking data corresponding to large-scale OSNs, including LBSNs. Different aspects have to be considered in the performance analysis of developed algorithms to ensure its use in real-world applications.

1.3 Hypothesis

The research hypothesis of this work is that the use of both social and location information enable better understanding of the behaviors and interests of users in location-based social networks, specially those related to the friendship establishment process. Thus, the combination of different aspects of such information sources (e.g. social strength, social group membership, geographical distance and frequency of visitations at places) can be a useful resource for mining and analyzing user behavior.

1.4 Thesis Organization

This work is divided in three more parts. In Chapter 2 we present the main theoretical fundamentals and previous work on complex networks, link prediction and location-based social networks, which are the main topics tackled in this thesis. In Chapter 3 we introduce all our contributions. In this same chapter, we begin by discussing how to use link prediction methods based on community information is possible: i) obtain accurate friendship predictions, and ii) perform such predictions in a time frame comparable to state-of-the-art methods. After, we introduce a new research issue, which consists in take into account information related to user membership in one or more social groups. Then, we establish the challenges of this issue, as well as we propose new network properties and new link prediction methods based on social group information. After, we extend these findings to the domain of LBSNs and present new friendship prediction methods combining different aspects of location information (e.g. frequency of visits, information gain and geographical distance) with social information. Finally, in Chapter 4 we present the conclusions of our investigation and point out the directions for some future work.

CHAPTER

2

BACKGROUND AND RELATED WORK

The work in this thesis tackled three different although intertwined topics: complex networks, link prediction and location-based social networks. Therefore, in this chapter we will discuss the main theoretical fundamentals and previous work on these topics. In Section 2.1 we will introduce some properties, models, categories, and other concepts related to complex network research issue. Afterwards, in Section 2.2, we will present the link prediction problem, as well as a detailed survey on existing link prediction methods and applications. Considering that link prediction is one of the most used frameworks to understand user behavior in both the traditional online social networks and location-based social networks, in Section 2.3 we will provide a systematic review of main concepts, properties and challenges of location-based social networks, as well as the different link prediction approaches for this specific kind of network. Finally, we conclude in Section 2.4 by presenting an overview of the research topics addressed in this chapter.

2.1 Complex Networks

Complex networks are systems of either natural and/or man-made phenomena in which the interactions between their components give rise to intricate networks. They emerge in a wide range of disciplines in the natural and social sciences, such as Internet, the World Wide Web, power grids, communication systems, and genetic, social, and epidemiological networks. Therefore, it is only logical to infer that networks are virtually everywhere.

In this section we will discuss and explore the main fundamentals of the analysis of complex networks. We will focus on the basic concepts used to develop the main scope of this thesis. Section 2.1.1 provides an overall idea of concepts and perspectives related to the new, interdisciplinary research field of Complex Networks. Since complex networks and graphs share similar definitions, Section 2.1.1 includes the basic notations of graph theory and network

representation. Sections 2.1.3, 2.1.4, and 2.1.5 show the main complex network models, measurements, and types, respectively. Section 2.1.6 concludes with a comprehensive list of recent trends in complex network research.

2.1.1 Complex Networks and Complex Network Science

For centuries, *reductionism* seemed to be the best approach to understand natural phenomena, dissecting them into their smaller segments and analyzing each part in isolation, disregarding their interactions (POLKINGHORNE, 2002). Later, an opposite model took place, stating that to better understand a phenomenon we need to consider all its multiple objects as well as their interactions (BALL, 2005). One of the main properties of this model, called system, is the fact that only the *collective behavior* of an ensemble of interacting objects allow the phenomenon to take place (VICSEK, 2002).

To study this new kind of phenomenon properly, a new kind of science was required. The *science of complexity* arose to reveal the principles that govern the ways in which the interactions and behaviors of inter-connected parts of a complex system take place. A *complex system* is basically defined as one whose pattern and behavior is hard to describe and understand by dissecting it into different units, while disregarding their interactions (WALDROP, 1993; VICSEK, 2002; ZWEIG, 2007). This definition can include systems that are not generally considered "complex"; thus Newman (2003) points out that the fundamental property of complex systems is the fact that they are composed by a large number of interacting components whose collective behavior can not be described as a simple combination of the behavior of each individual component.

Complex networks have emerged as a way to illustrate complex systems. Their different entities are represented by vertices, and the interactions between such entities are represented by edges between these vertices (ZWEIG, 2007; NEWMAN, 2010; SILVA; ZHAO, 2016; BARABÁSI, 2016). While some researchers see complex networks simply as a special, very useful structure to observe and analyze the interaction patterns between entities (NEWMAN, 2003), others believe they play a more important role (as a kind of 'skeleton' of the complex system); in this case, its understanding becomes key to understand processes in the whole system (STROGATZ, 2001; BARABÁSI, 2005).

Regardless of which of these perspectives is correct, complex networks have become the object of study of the *complex network science*, a research field formally accepted in the realm of *complex systems science* (ONCE-CS, 2006). Thereby, all research related to the complex network science can be grouped under three perspectives:

1. *Complex Network Analysis*: uses formalisms and methods to try to understand and explain the structure of real-world networks, aiming to differentiate them from established graph models.

2. *Complex Network Models*: intends to model topological properties related to real-world networks, and therefore characterize them in an easier way.

3. *Processes on Complex Networks*: analyze the results of a process or algorithm on a specific network structure.

This section also provides a short overview of the most important findings in complex network analysis and models due to their relevance for the understanding of this thesis. In later chapter, we will develop different processes in complex networks as part of the outcomes of the contributions of this thesis.

2.1.2 Basic Concepts

In order to establish a well defined and coherent network terminology, in this section we review the main concepts to be used. These concepts have been widely studied, reason why we will not offer details (BOLLOBÁS, 1998; ZWICK, 2001; COSTA *et al.*, 2007; NEWMAN, 2010; YE; WU; WANG, 2010; SILVA; ZHAO, 2016; BARABÁSI, 2016).

Considering that a network is modeled with the mathematical object called *graph*, we will use these two terms interchangeably. Similarly, the data relationships that make up a graph are called structure, topology, or anatomy of the network.

Graph Definitions

In the following, we present the formal definitions of different graph topologies.

Definition 2.1. Graph. A graph G is composed by a pair (V,E), where V is a nonempty and finite set of vertices, also called nodes, and E is a set of edges, also called links, between the vertices. Therefore, $E \subseteq \{(x,y) \mid x,y \in V\}$.

There is important to note that some special graphs can be grouped as follows:

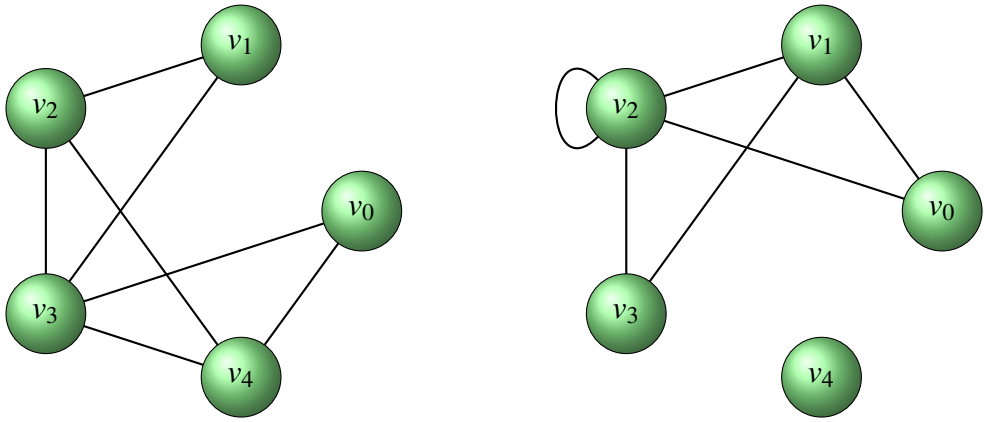
1. *Graph with no self-loops*: when all the edges in E are irreflexive, i.e. $\forall x \in V, (x,x) \notin E$. This leads to the graph being called free of self-loops.

2. *Graph with self-loops*: when edges connect back the same vertices without leaving them, i.e. $\exists x \in V, (x,x) \in E$.

***Remark* 2.1.** The sizes of the sets V and E are denoted by $|V|$ and $|E|$, and represent, respectively, the total number of vertices and edges of the graph.

In Figure 1a, we portrayed a graph G which vertex set is $V = \{v_0, v_1, v_2, v_3, v_4\}$ and the edge set is $E = \{(v_0,v_3),(v_0,v_4),(v_1,v_2),(v_1,v_3),(v_2,v_3),(v_2,v_4),(v_3,v_4)\}$. Therefore, the size

Figure 1 – An example of a graph (a) with no self-loops and (b) with self-loops.



(a) (b)

Source: Elaborated by the author.

of vertex set is $|V| = 5$, and the size of the edge set is $|E| = 7$. We can check that no edge is connecting back the same vertex, so this graph is into the group of graphs with no self-loops.

In Figure 1b we plotted another graph with the presence of the edge (v_2, v_2), which turns it into one with self-loops. We also observe that vertex v_4 is not connected with any other vertex. This type of vertex is called as *singleton* or *isolated*.

Definition 2.2. Complete Graph. A complete graph is a graph where their edges connect all the pairs of vertices.

It is also possible that a complete graph can be with or without self-loops. Figure 2a shows an example of complete graph with no self-loops, where every pair of vertices is connected by an edge.

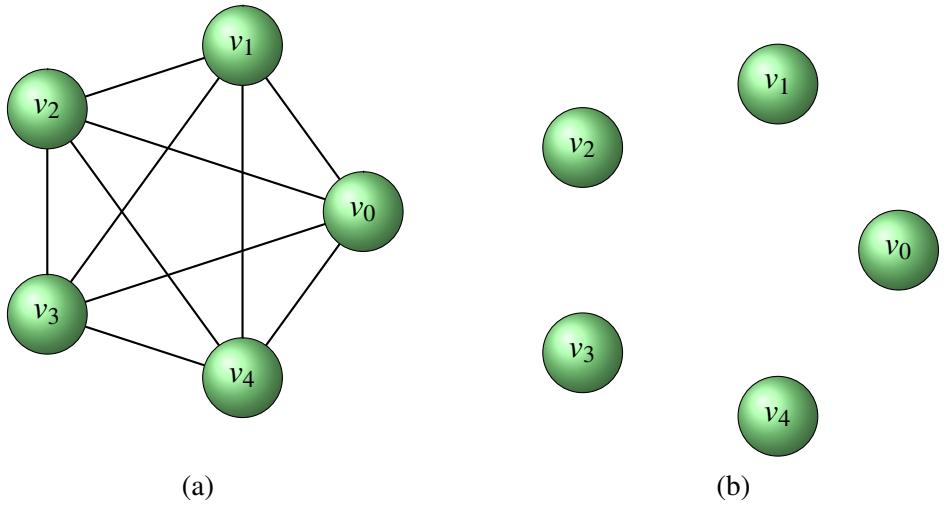
Definition 2.3. Null Graph. A null graph is a graph without edges, i.e. $E = \varnothing$.

Despite the set of edges of a null graph is empty, it meets Definition 2.1. Figure 2b illustrates an example of null graph.

Definition 2.4. Undirected Graph. When the edges between the vertices are symmetric, i.e. $\forall(x,y) \in E \Rightarrow (y,x) \in E$, the graph is called undirected. In other terms, when there is an edge linking vertices x to y, so there will be a link from y to x.

Commonly, edges of undirected graphs are drawn with no arrows in their endpoints. Thus, for an edge $(x,y) \in E$, here is assumed the existence of its opposite $(y,x) \in E$. Examples of undirected graphs are the graphs showed in Figures 1a, 1b and 2b.

Figure 2 – An example of a (a) complete graph and a (b) null graph.



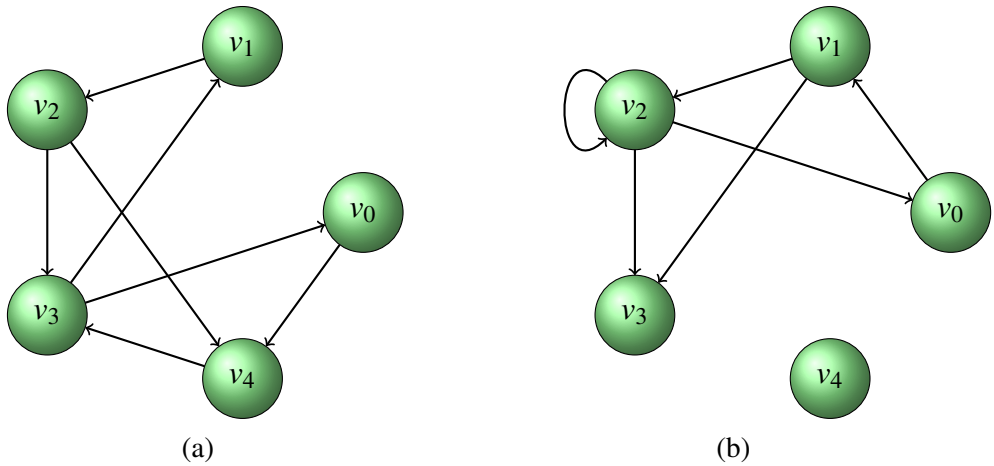
(a) (b)

Source: Elaborated by the author.

Definition 2.5. Directed Graph. The edges of a directed graph, also called digraph, not necessarily are symmetric, i.e. $\exists(x,y) \in E \mid (y,x) \notin E$. In other terms, there exists at least an arbitrary edge linking x to y, with the absence of the opposite link.

Edges of directed graphs are drawn with arrows indicating the linkage direction. Figure 3a illustrates a direct graph where the directness of the edges implies that there exists the edge $(x,y) \in E$ such that $(y,x) \notin E$. The presence of sel-loops in this type of graphs is possible too, as showed in Figure 3b.

Figure 3 – An example of a directed graph (a) with no self-loops and (b) with self-loops.



(a) (b)

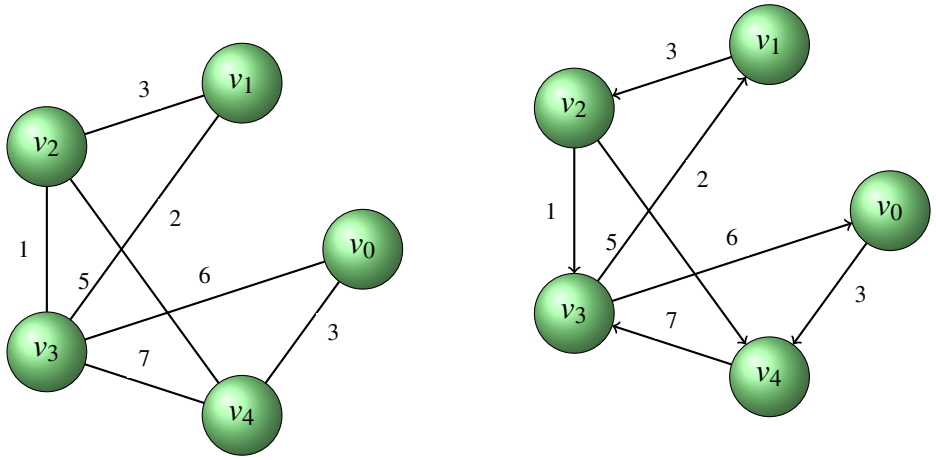
Source: Elaborated by the author.

Definition 2.6. Weighted Graph. A weighted graph G is defined as a triple $G = (V,E,\mathbf{W})$, where besides the sets of nodes and edges, V and E, respectively, there is also $\mathbf{W}$, representing

a matrix that carries the edge weights. Therefore, the weight of the edge linking vertices x to y is represented by the entry $\mathbf{W}_{x,y} = w$, considering that $w > 0$ and $\forall(x,y) \in E$. If $(x,y) \notin E \Rightarrow \mathbf{W}_{x,y} = 0$.

Weighted graphs are drawn with values associated to each edge. Often, when no edge weight is specified, it is assumed that the weight is unitary. Therefore, the graph showed in Figure 1a can be assumed as a weighted graph where all its edges have unitary weights. Whilst, the same undirected graph is plotted in Figure 4a but considering different weight values for its edges. In Figure 4b is showed the graph with the same weights but considering directed edges.

Figure 4 – An example of a (a) weighted undirected graph and a (b) weighted directed graph.



Source: Elaborated by the author.

***Remark* 2.2.** When $\mathbf{W}$ is a binary matrix, then the weighted graph is reduced to a simple graph according to the Definition 2.1.

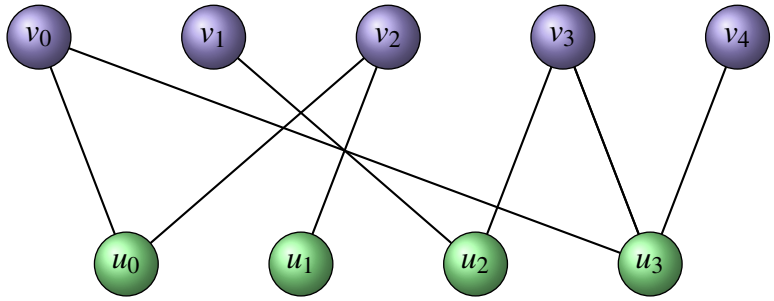
Definition 2.7. Bipartite Graph. A bipartite graph, or bigraph, is a graph in which its vertex set V can be divided into two disjoint non-empty subsets V_1 and V_2, i.e. $V = V_1 \cup V_2$, in such a way that $(x,y) \in E \Rightarrow x \in V_1 \wedge y \in V_2$. Therefore, there is no link between pairs of vertices in the same subsets V_1 or V_2.

***Remark* 2.3.** Note that a bipartite graph cannot have self-loops.

***Remark* 2.4.** A bipartite complete graph is one where $\forall(x,y) \in V_1 \times V_2, (x,y) \in E$.

When it is necessary modelling of relations between two different classes of objects, bipartite graphs often arise naturally. Figure 5 depicts a bipartite graph with $V = \{v_0, v_1, v_2, v_3, v_4, u_0, u_1, u_2, u_3\}$, where $V_1 = \{v_0, v_1, v_2, v_3, v_4\}$ and $V_2 = \{u_0, u_1, u_2, u_3\}$. Note that the only exists edges between vertices of different subsets.

Figure 5 – An example of a non-weighted and undirected bipartite graph.



Source: Elaborated by the author.

***Remark* 2.5.** A k-partite graph is a graph whose set of vertices can be divided into k disjoint non-empty subsets. Therefore, a bipartite graph is a 2-partite graph due to that $k = 2$, i.e. a bipartite graph is a particular instance of a k-partite graph.

For further details about bipartite graphs and k-partite graphs, one can refer to Asratian, Denley and Häggkvist (1998).

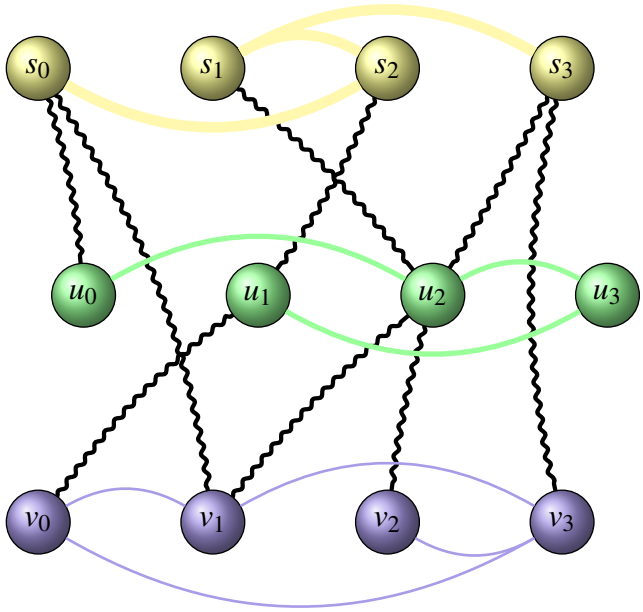
Definition 2.8. Heterogeneous Graph. A heterogeneous graph is a graph allowing multiple types of both nodes and edges. Therefore, given the graph $G = (V,E)$ as stated in Definition 2.1, there exists a node type mapping function $\phi : V \Rightarrow \mathscr{A}$ and an edge type mapping function $\psi : E \Rightarrow \mathscr{R}$, where each node $x \in V$ belongs to one particular object type $\phi(x) \in \mathscr{A}$, and each edge $e = (x,y) \in E, \forall x,y \in V$ belongs to a particular relation $\psi(e) \in \mathscr{R}$.

***Remark* 2.6.** If the number of node types and the number of edges types is unitary, i.e. $|\mathscr{A}| = 1$ and $|\mathscr{R}| = 1$, the graph is a simple graph according Definition 2.1, but if $|\mathscr{A}| > 1$ or $|\mathscr{R}| > 1$, the graph is heterogeneous.

When it is necessary modelling different types of relations between objects belonging to different classes, heterogeneous graphs arise naturally. Figure 6 shows an example of heterogeneous graph with three types of nodes, $|\mathscr{A}| = 3$, being that we color the first node type, $\phi_1 = \{s_0, s_1, s_2, s_3\}$, with yellow, the second node type, $\phi_2 = \{v_0, v_1, v_2, v_3\}$, with green, and the third node type, $\phi_3 = \{v_0, v_1, v_2, v_3\}$, with blue. Therefore, $V = \phi_1 \cup \phi_2 \cup \phi_3$.

Similarly, the heterogeneous graph depicted in Figure 6 has four types of edges, $|\mathscr{R}| = 4$, being that we illustrate the first edge type, $\psi_1 = \{(s_0, s_2), (s_1, s_2), (s_1, s_3)\}$, which represent the relations between first type nodes, with yellow, thicker and continuous lines; the second edge type $\psi_2 = \{(u_0, u_2), (u_1, u_3), (u_2, u_3)\}$, which represent the relations between second type nodes, with green, less thick and continuous lines; the third edge type, $\psi_3 = \{(v_0, v_1), (v_0, v_3), (v_1, v_3), (v_2, v_3)\}$, which represent the relations between third type nodes, with blue, fine and continuous lines; and the fourth edge type, $\psi_4 = \{(s_0, u_0), (s_0, v_1), (s_1, u_2), (s_2, u_1), (s_3, u_2), (s_3, v_3), (u_1, v_0), (u_2, v_1), (u_2, v_2)\}$, which represent the relations between nodes belonging to different types, with black and snake lines. Therefore, $E = \psi_1 \cup \psi_2 \cup \psi_3 \cup \psi_4$.

Figure 6 – An example of a non-weighted and undirected heterogeneous graph with three different types of nodes and four different types of edges.



Source: Elaborated by the author.

Based on heterogeneous graphs, different graph configurations, such as: multigraph, multilayer graph, multidimensional graph, and others, have been studied in the literature. For further details refer to Kivela *et al.* (2014).

Connectivity

In this section, we briefly define common concepts related to graph connectivity, which are used throughout this thesis.

Definition 2.9. Adjacent vertices. Two vertices $x \in V$ and $y \in V$ are called adjacent if they share a common edge.

***Remark* 2.7.** In undirected graphs, if x is adjacent to y, then y must be adjacent to x as well.

***Remark* 2.8.** In directed graphs, if x is adjacent to y, then y not necessary is adjacent to x. Specifically, if $(x,y) \in E$ and $(y,x) \notin E$, then x is adjacent to y, but the opposite does not hold.

For instance, in the undirected graph shown in Figure 1a, vertices v_1 and v_2 are adjacent to each other. In contrast, vertex v_0 is not adjacent to v_1. On the other hand, in the directed graph portrayed in Fig 3a, vertex v_2 is adjacent to v_1, but the converse is not true.

Definition 2.10. Neighborhood of a vertex. The neighborhood of a vertex $x \in V$ in a graph G is the set of vertices adjacent to x. The neighborhood is denoted by $\Gamma(x)$ and is formally given by $\Gamma(x) = \{y : (x,y) \in E\}$.

For instance, in the undirected graph depicted in Figure 1a, the neighborhood of vertex v_3 is $\Gamma(v_3) = \{v_0, v_1, v_2, v_4\}$. On the other hand, in the directed graph exhibited in Fig 3a, $\Gamma(v_3) = \{v_0, v_1\}$.

Definition 2.11. Degree of a vertex. The degree, valency or connectivity of a vertex is the total number of adjacent vertices it has. The degree of a vertex $x \in V$ is denoted by k_x, and we can formally define it as the cardinality of its neighborhood set, i.e. $k_x = |\Gamma(x)|$.

***Remark* 2.9.** The feasible values of k_x are within the discrete-valued interval $\{0, \ldots, |V|-1\}$ if self-loops are not allowed, and in $\{0, \ldots, |V|\}$ if self-loops are permitted.

***Remark* 2.10.** When $k_x = 0$, then x is called as singleton or isolated vertex.

***Remark* 2.11.** When k_x assumes relative large values than the remainder of the vertices in the network, it is called *hub*.

For instance, in the undirected graph depicted in Figure 1b, vertex v_4 is a singleton, for $k_{v_4} = 0$. On the other hand, in Figure 1a, the vertex v_3 is considered a hub due to that it has a degree relative large in relation to the remainder vertices of the graph.

For directed graphs, due to that the distinctions in the edge endpoints must be taken into consideration, is necessary extend some connectivity definitions.

Definition 2.12. In-degree and out-degree of a vertex. In a directed graph, the notion of vertex degree can be extended to the in-degree, k_x^{in}, and out-degree, k_x^{out}, being that for each one of them is important to consider the direction of the connectivity, so that $k_x^{in} = |\{y : (y,x) \in E\}|$, and $k_x^{out} = |\{y : (x,y) \in E|$. Furthermore, the in-degree and out-degree of a directed graph provide the total degree of a vertex, i.e. $k_x = k_x^{in} + k_x^{out}$.

***Remark* 2.12.** The feasible discrete-valued interval for k_x^{in} and k_x^{out} is $\{0, \ldots, |V|-1\}$ if self-loops are not allowed, and $\{0, \ldots, |V|\}$ if loops are present. Therefore, k_x may assumes the values $\{0, \ldots, 2(|V|-1)\}$ when self-loops are not permitted and $\{0, \ldots, 2|V|\}$ when loops are allowed.

***Remark* 2.13.** Note that $k_x^{out} = |\Gamma(x)|$.

For instance, in the directed graph showed in Figure 3a, $k_{v_3}^{in} = |\{v_2, v_4\}| = 2$, $k_{v_3}^{out} = |\{v_0, v_1\}| = 2$, and $k_{v_3} = 2+2 = 4$. Note that $k_{v_3}^{out} = |\Gamma(v_3)|$.

Definition 2.13. Average graph degree. The average degree of a graph, $\langle k \rangle$, is the average of degrees of all the vertices in the graph, as stated in Equation 2.1.

$$\langle k \rangle = \frac{1}{|V|} \sum_{x \in V} k_x \tag{2.1}$$

For instance, in the undirected graph exhibited in Figure 1a, the average degree is $\langle k \rangle = \frac{1}{5}(k_{v_0}, k_{v_1}, \ldots, k_{v_4}) = \frac{1}{5}(2+2+3+4+3) = 2.8$.

Definition 2.14. Average in-degree and out-degree. In a directed graph, the average in-degree and out-degree have the same numerical values and are calculated as stated in Equations 2.2 and 2.3, respectively.

$$\langle k^{in} \rangle = \frac{1}{|V|} \sum_{x \in V} k_x^{in} \tag{2.2}$$

$$\langle k^{out} \rangle = \frac{1}{|V|} \sum_{x \in V} k_x^{out} \tag{2.3}$$

For instance, in the directed graph showed in Figure 3a, the average in-degree is given by $\langle k^{in} \rangle = \frac{1}{5}(k_{v_0}^{in} + k_{v_1}^{in} + \ldots + k_{v_4}^{in}) = \frac{1}{5}(1+1+1+2+2) = 1.4$, whilst the average out-degree is given by $\langle k^{out} \rangle = \frac{1}{5}(k_{v_0}^{out} + k_{v_1}^{out} + \ldots + k_{v_4}^{out}) = \frac{1}{5}(1+1+2+2+1) = 1.4$.

Definition 2.15. Strength. In an undirected weighted graph, the strength of a vertex $x \in V$, indicated by s_x, is the sum of the weights of all connection of x with its neighbors as stated in Equation 2.4.

$$s_x = \sum_{y \in \Gamma(x)} \mathbf{W}_{x,y} \tag{2.4}$$

For instance, in the undirected weighted graph plotted in Figure 4a, the strength of vertex v_3 is $s_{v_3} = 6+5+1+7 = 19$.

Definition 2.16. In-strength and out-strength. In a directed graph, the concept of vertex strength is extended to in-strength, s_x^{in}, and out-strength, s_x^{out}, as stated in Equations 2.5 and 2.6, respectively.

$$s_x^{in} = \sum_{y \in \Gamma(x)} \mathbf{W}_{y,x} \tag{2.5}$$

$$s_x^{out} = \sum_{y \in \Gamma(x)} \mathbf{W}_{x,y} \tag{2.6}$$

For directed graph, $s_x = s_x^{in} + s_x^{out}$.

For instance, in the directed weighted graph supplied in Figure 4b, for vertex v_3, $s_{v_3}^{in} = 1+7 = 8$, $s_{v_3}^{out} = 5+6 = 11$, and $s_{v_3} = s_{v_3}^{in} + s_{v_3}^{out} = 19$.

Definition 2.17. Regular Graph. A graph is called regular if all the vertices of the graph have the same degree. In particular, if the degree of each vertex is k, the graph is said to be k-regular.

***Remark* 2.14.** G is a complete graph, with $|V|$ vertices, if it is $(|V|-1)$-regular.

In Figure 7a is illustrated a 2-regular graph that is not complete. In Figure 2a we observe a 4-regular graph that is complete.

Paths and Cycles

In this section, we define some of the main concepts related to graph transversal.

Definition 2.18. Walk. Given a set of vertices $v_0, v_1, \ldots, v_{k-1} \in V$, $k \geq 2$, a walk $\mathscr{W}$ is an ordered sequence of edges: $W = \{(v_0, v_1), (v_1, v_2), \ldots, (v_{k-2}, v_{k-1})\}$, such that $\forall k \in \{2, \ldots, k\}$: $(v_{k-2}, v_{k-1}) \in E$. In this case, v_0 and v_{k-1} are called the origin and destination of the walk, respectively. There is important to note that vertices can be revisited in the same walk.

***Remark* 2.15.** A walk is called closed if v_0 and v_{k-1} are the same vertex and open otherwise.

***Remark* 2.16.** A walk consisting of a single vertex is called a trivial walk.

Definition 2.19. Trail. A trail is a walk in which no edge is repeated. Trails can also be classified as open and close according Remark 2.15.

***Remark* 2.17.** A closed trail is called tour or circuit.

Definition 2.20. Walk Length. The length of a walk W is the number of edges that the walk traverses.

To introduce the graph transversal concepts, we use the undirected graph showed in Figure 7b, $W_1 = \{(v_3, v_2), (v_2, v_1), (v_1, v_6), (v_6, v_0), (v_0, v_1)\}$ is an open walk. In contrast, $W_2 = \{(v_3, v_2), (v_2, v_1), (v_1, v_6), (v_6, v_0), (v_0, v_1), (v_1, v_2), (v_2, v_3)\}$ is a closed walk. There are no trivial walks due to the observed graph has no self-loops. Also, $W_3 = \{(v_5, v_7), (v_7, v_0)\}$ is an open trail, whilst $W_4 = \{(v_5, v_7), (v_7, v_0), (v_0, v_5)\}$ is a closed trail or a tour. The walk lengths are $|W_1| = 5$, $|W_2| = 7$, $|W_3| = 2$, $|W_4| = 3$.

Definition 2.21. Path. A path P is a non-trivial walk in which all vertices (except possibly the first and last) are distinct.

***Remark* 2.18.** A path is always a walk.

Definition 2.22. Cycle. A cycle is a closed path.

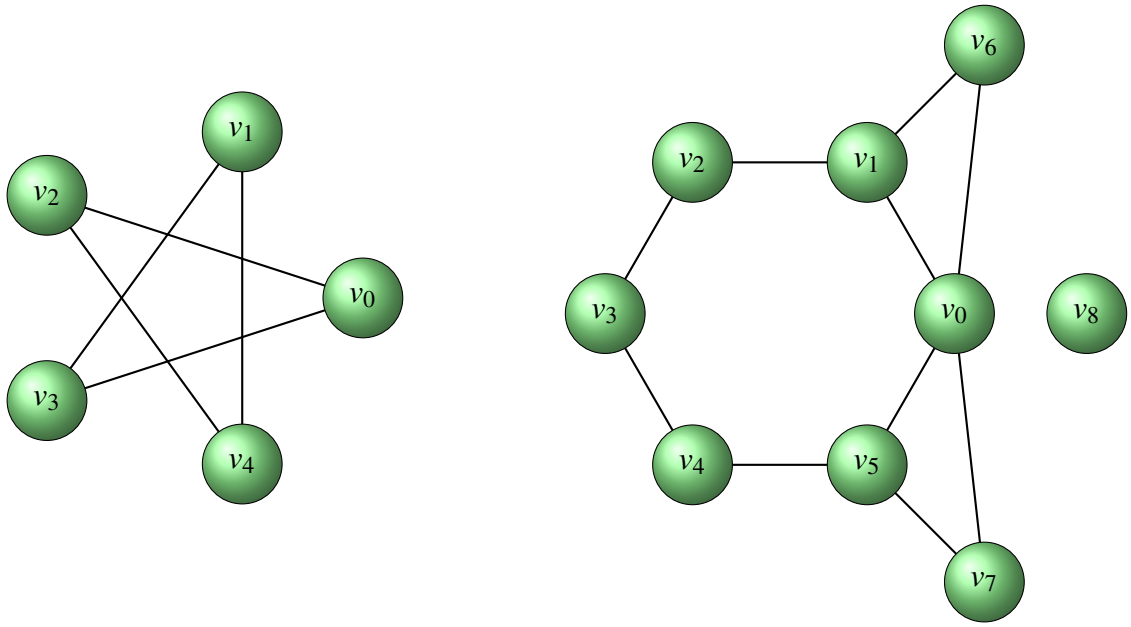
For instance, in the graph of Figure 7b, $P_1 = \{(v_3, v_4), (v_4, v_5), (v_5, v_0)\}$ is a path, and $P_2 = \{(v_3, v_4), (v_4, v_5), (v_5, v_0), (v_0, v_1), (v_1, v_2), (v_2, v_3)\}$ is a cycle. Note that $P_3 = \{(v_5, v_7), (v_7, v_0), (v_0, v_6), (v_6, v_1), (v_1, v_0), (v_0, v_5), (v_5, v_7)\}$ is a walk and tour but not a cycle.

Definition 2.23. Walk or Path Distance. The distance d of the walk W is given by:

$$d(W) = \sum_{i=1}^{k-1} |(v_{i-1}, v_i)|, \tag{2.7}$$

where $|(v_{i-1}, v_i)|$ is the edge linking the $(i-1)$-th and i-th vertices, i.e. $|(v_{i-1}, v_i)| = \mathbf{W}_{v_{i-1}, v_i}$.

Figure 7 – An example of a 2-regular graph that is not complete and (b) an illustrative undirected graph to introduce the graph transversal concepts.



(a) (b)

Source: Elaborated by the author.

Definition 2.24. Shortest Path. The shortest path, also called geodesic path, between $x \in V$ and $y \in V$, denoted by $P^s_{x,y}$, is given by the path starting with x and ending in y with the smallest distance, as stated in Equation 2.8.

$$P^s_{x,y} = \min_{W_{x \to y}} d(W_{x \to y}). \tag{2.8}$$

$W_{x \to y}$ represents a walk starting with x and ending in y.

Definition 2.25. Distance between vertices. The distance $d_{x,y}$ between two vertices x and y is always their shortest path distance, i.e. $d_{x,y} = |P^s_{x,y}|$.

***Remark* 2.19.** Note that $d_{x,y}$ is always a path.

***Remark* 2.20.** The distance from any vertex to itself is 0.

***Remark* 2.21.** If there is no path from x to y, then $d_{x,y} = \infty$.

For instance, in the graph plotted in Figure 7b, the distance between v_2 and v_3 is $d_{v_2,v_3} = 1$, since the shortest path from v_2 to v_3 is $P^2_{v_2,v_3} = \{(v_2, v_3)\}$. The distance from vertex v_8 to itself is $d_{v_8,v_8} = 0$, and the distance from v_0 to v_8 is $d_{v_0,v_8} = \infty$.

Subgraphs

In this section we present some concepts related to subgraphs.

Definition 2.26. Reachability. A vertex $y \in V$ is reachable from $x \in V$ if $d_{x,y}$ is finite. Alternatively, x reaches y if there is at least one walk that starts from x and ends at y.

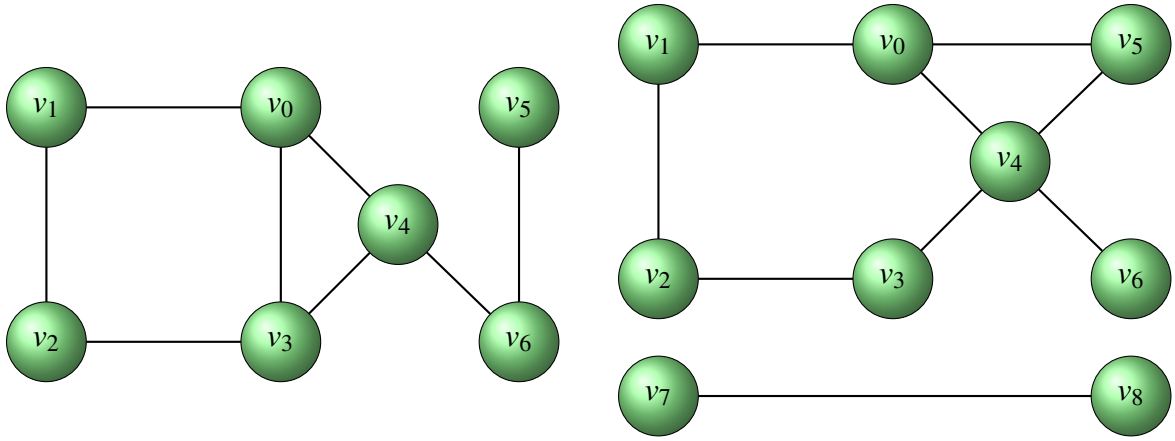
Definition 2.27. Connectedness. A graph is connected if, for every pair of vertices $x \in V$ and $y \in V$, y is reachable from x or x is reachable from y.

Definition 2.28. Strong Connectedness. A graph is strongly connected if, for every pair of vertices $x \in V$ and $y \in V$, y is reachable from x and x is reachable from y.

***Remark* 2.22.** In undirected graphs, connectedness implies strong connectedness, but in directed graphs this fact is not true.

For instance, in the undirected graph illustrated in the Figure 8a, the graph is strongly connected since all the pair of vertices are mutually reachable. In contrast, the graph depicted in Figure 8b is neither strongly connected nor connected, since v_3 and v_7 are mutually non-reachable.

Figure 8 – Illustrative graphs for exemplifying subgraph concepts. In (a) an undirected graph with one component, and in (b) an undirected graph with two components.



(a) (b)

Source: Elaborated by the author.

Definition 2.29. Graph Component. The subgraph G_C of G is a component if:

- G_C is connected;
- All of the proper subsets of G_C are not connected.

Alternatively, a subgraph G_C is a graph component if any two of its vertices are reachable at least from one to another, and if its vertices are connected to no additional vertices in the remainder of the graph.

***Remark* 2.23.** A connected graph has always a single component.

For instance, in Figure 8a there is a single component that is the graph itself. In contrast, in Figure 8b, two components exists, $G_1 = \{v_0, v_1, v_2, v_3, v_4, v_5, v_6\}$ and $G_2 = \{v_7, v_8\}$.

Definition 2.30. Clique. The clique in an undirected graph is a subset of vertices such that every two vertices in the subset are connected by an edge. Therefore, cliques are subgraphs or graphs that are complete.

For instance, in Figure 8a we observe the presence of a clique comprising the vertices $\{v_0, v_3, v_4\}$, while in Figure 8b, other clique is observed in the subset $\{v_0, v_4, v_5\}$.

Graph Representation

A non-weighted graph $G = (V, E)$ or a weighted graph $G = (V, E, \mathbf{W})$ are frequently represented by two data structures: adjacency list and adjacency matrix. In this section, we briefly define these network representations since they are used throughout this thesis (SILVA; ZHAO, 2016; YAVEROGLU, 2013; NEWMAN, 2010).

Definition 2.31. Adjacency List. The adjacency list of a non-weighted graph $G = (V, E)$ is a $|V|$ dimensional array A, where each element of the array A_x corresponds to a vertex $x \in V$ and is linked to the set of vertices adjacent to x. For a weighted graph $G = (V, E, \mathbf{W})$, an extra list of edge weights should be kept for each vertex.

Definition 2.32. Adjacency Matrix. The adjacency matrix of a non-weighted graph $G = (V, E)$ is a $|V| \times |V|$ matrix $\mathbf{A}$, where $\mathbf{A}_{x,y}$ is a non-zero value when vertices x and y are connected, and equal to zero otherwise. For a weighted graph $G = (V, E, \mathbf{W})$, the weighted edges can be encoded in the value of $\mathbf{A}_{x,y}$.

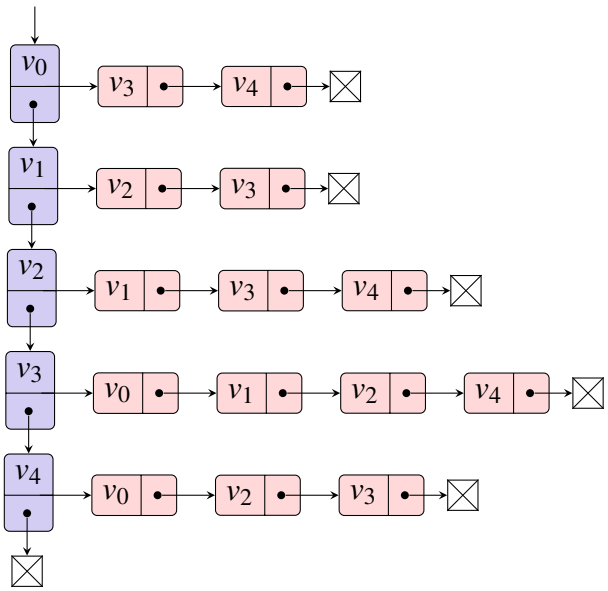
***Remark* 2.24.** If the graph is undirected, then $\mathbf{A}$ is symmetric. This fact implies that if $\mathbf{A}_{x,y} = 1$, then $\mathbf{A}_{y,x} = 1$.

***Remark* 2.25.** If the graph is directed, then $\mathbf{A}$ can not be symmetric.

For instance, the adjacency list and adjacency matrix for the graph portrayed in Figure 1a, are showed in Figures 9a and 9b, respectively.

Both representations have their own representations. Based on Yaveroglu (2013), in Chart 1 we summarize the worst-case of space and time complexities for the performance of basic operations in computer memory when the two graph representations are used. Therefore, from this chart we can observe that the adjacency list representation potentially provides greater memory efficiency than adjacency matrices. This advantage is maintained by the adjacency list representation operations related to adding a vertex to or delete it from the network, since the size of the matrix changes and the whole matrix needs to be allocated again. On the contrary, adding, deleting or searching an edge can be best performed by using adjacency matrices, due

Figure 9 – Examples of the (a) adjacency list A and (b) adjacency matrix $\mathbf{A}$, for the graph previously showed in Figure 1a.



(a)

	v_0	v_1	v_2	v_3	v_4
v_0	0	0	0	1	1
v_1	0	0	1	1	0
v_2	0	1	0	1	1
v_3	1	1	1	0	1
v_4	1	0	1	1	0

(b)

Source: Elaborated by the author.

Note – In Figure 9a, the last element of the dimensional array A is linked to a null element. The same occurs with the last element of set of vertices of all the elements of A. The image is inspired in the source code publicly shared by Wdvorak: <https://commons.wikimedia.org/wiki/File:Adjacencylist_linkedlistof_linkedlists_directedgraph.svg> (Accessed August 31, 2017).

to the existence and the weight of an edge can be directly changed from the relevant matrix element.

Chart 1 – Comparison between adjacency list and adjacency matrix representations with respect to their space and time complexities for basic graph operations.

Operation	**Adjacency List**	**Adjacency Matrix**
Storage	$O(\|V\|+\|E\|)$	$O(\|V\|^2)$
Add Vertex	$O(1)$	$O(\|V\|^2)$
Add Edge	$O(1)$	$O(1)$
Delete Vertex	$O(\|E\|)$	$O(\|V\|^2)$
Delete Edge	$O(\|E\|)$	$O(1)$
Search Edge	$O(\|V\|)$	$O(1)$

Source: Adapted from Yaveroglu (2013).

The space allocated for the adjacency matrix can be used more effectively by using

different types of information about the graph. For example, the diagonal elements of the adjacency matrix of a graph with no self-loops are all equal to 0 as showed in Figure 9b. Therefore, the space allocated for the diagonal elements can be used for representing other node-specific information.

Definition 2.33. Degree Matrix. The degree matrix **D**, is a diagonal matrix containing information about the degree of each vertex $x \in V$, where:

$$\mathbf{D}_{x,y} := \begin{cases} k_x, & if \ \ x = y \\ \\ 0, & otherwise. \end{cases} \tag{2.9}$$

Definition 2.34. Laplacian Matrix. The Laplacian matrix **L**, also called admittance, Kirchhoff or discrete Laplacian matrix, measures to what extent a graph differs at one vertex from its values at nearby vertices. The standard combinatorial Laplacian matrix is computed as stated in Equation 2.10.

$$\mathbf{L} = \mathbf{D} - \mathbf{A}. \tag{2.10}$$

For instance, for the graph plotted in Figure 1a, we shown the respective degree matrix and Laplacian matrix in Figures 10a and 10b, respectively.

Figure 10 – Illustrative figures of the (a) degree matrix and (b) Laplacian matrix, for the graph previously showed in Figure 1a.

	v_0	v_1	v_2	v_3	v_4
v_0	2	0	0	0	0
v_1	0	2	0	0	0
v_2	0	0	3	0	0
v_3	0	0	0	4	0
v_4	0	0	0	0	3

(a)

	v_0	v_1	v_2	v_3	v_4
v_0	2	0	0	-1	-1
v_1	0	2	-1	-1	0
v_2	0	-1	3	-1	-1
v_3	-1	-1	-1	4	-1
v_4	-1	0	-1	-1	3

(b)

Source: Elaborated by the author.

For further details about the Laplacian matrix, one can refer to Chung (1997) and Newman (2010).

2.1.3 Network Models

Aiming to understand different topological properties related to real-world networks, several network models have been proposed. As examples of most popular categories of network models, one can list: random networks (ERDÖS; RÉNYI, 1959), small-world networks (WATTS; STROGATZ, 1998), and scale-free networks (BARABÁSI; ALBERT, 1999). Other network

models are based on specific structures present in real-world network, such as: core-periphery networks (BORGATTI; EVERETT, 2000) and random clustered networks (GIRVAN; NEWMAN, 2002). In this section, we introduce briefly only the three most popular network models (SILVA; ZHAO, 2016; BARABÁSI, 2016; NEWMAN, 2010; JAMAKOVIĆ, 2008).

Random Networks

The random network model, developed by Erdös and Rényi (1959), is one of the most studied network models. This model generates random networks consisting of $|V|$ vertices and $|E|$ edges. Starting from V vertices completely disconnected (no edges in the network), the network is built considering a gradual addition of $|L|$ edges randomly created, without self-loops. Each edge is created according to a probability $p > 0$.

In general, if an arbitrary edge is present in a random network with probability p, so it is absent with probability $1-p$. Furthermore, there are $\binom{|V|-1}{k}$ ways of choosing k vertices over $|V|-1$ in total, and p^k denotes the joint probability of these k vertices to have exactly k connected vertices, then $\binom{|V|-1}{k}p^k$ provides the probability of these k vertices to have exactly k other interconnected vertices. However, in this analysis, it should be imposed that there are no more edges beyond these k vertices, i.e. for the reminiscent quantity of vertices, $|V|-1-k$, the complementary probabilistic event of existing edges $(1-p)^{(|V|-1-k)}$ must happen. Therefore, the degree distribution of a random graph is governed by the Equation 2.11.

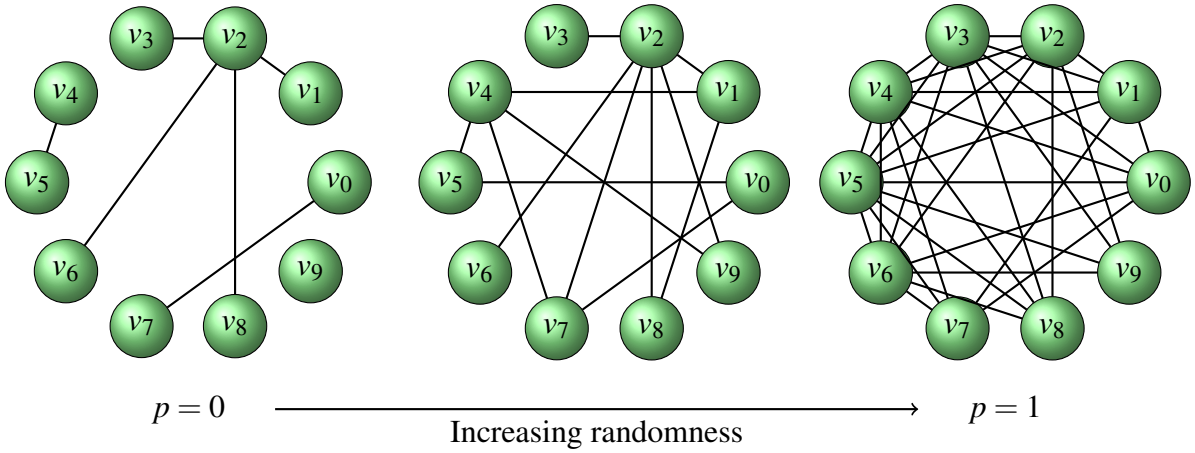
$$p(k) = \binom{|V|-1}{k} p^k (1-p)^{(|V|-1)-k}. \tag{2.11}$$

Given that $V \to \infty$ and $p \ll 1$, the degree distribution asymptotically approximates to a *Poisson distribution* with parameter *Poisson*(λ), i.e. $(|V|-1)p \approx \lambda$. Moreover, the average shortest path length $\langle \ell \rangle$ is small in random networks, increasing proportionally to the logarithm of the network size, i.e. $\langle \ell \rangle \sim \frac{\ln(|V|)}{ln(\langle k \rangle)}$, where $\langle k \rangle$ is given by the average value of the Poisson distribution, i.e. $\langle k \rangle = \lambda = (|V|-1)p$.

Other important property of the random networks is the phase transition, i.e. a random network emerges from a low edge density or low p value for which there are few edges and many small components to a high edge density or high p value for which an extensive fraction of all nodes are joined together in a single giant component. In Figure 11, we display an example of the phase transition for a network with $|V| = 10$ and with probability p varying from 0 to 1.

Many other properties of the random networks are known analytically in the limit of large network size, as was shown by Erdös and Rényi in a series of publications (ERDÖS; RÉNYI, 1959; ERDÖS; RÉNYI, 1960; ERDÖS; RÉNYI, 1961), and many other interesting properties have been discussed by other authors (BOLLOBÁS, 2001; NEWMAN, 2003; COSTA *et al.*, 2007).

Figure 11 – An example of the phase transition process in a random network. Increasing edge probability p implies that the network moves from a low edge density for which the are few edges and many small components to a high edge density network with an extensive fraction of all vertices connected in a single giant component.



Source: Elaborated by the author.

Small-World Networks

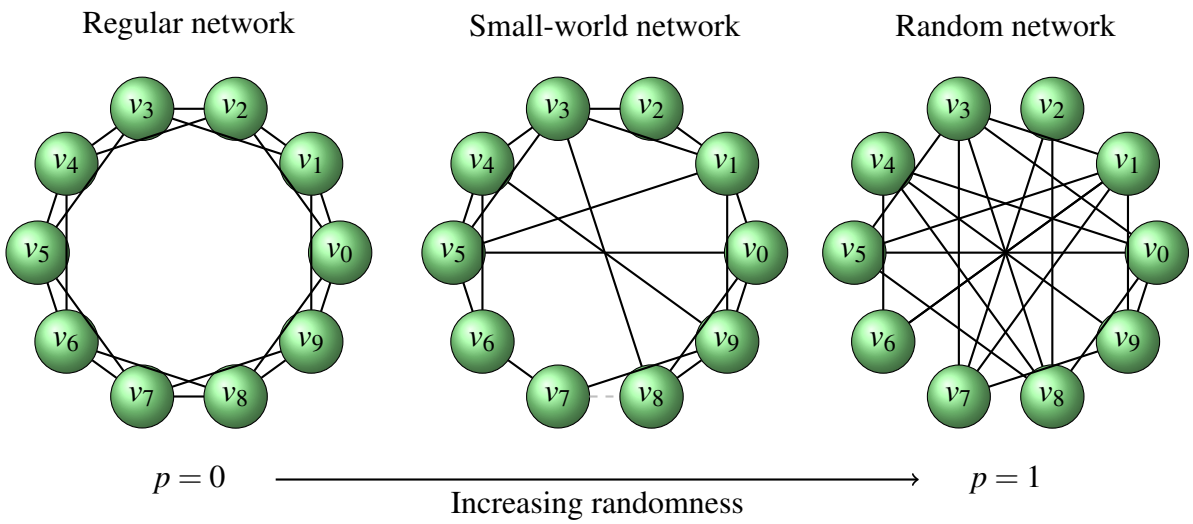
Several real-world networks exhibit the small-world property, i.e. despite the large size of a network there is a relatively short path between any two vertices. This characteristic is found, for example, in social networks, where virtually everyone in the world can be reached by a short chain of people (MILGRAM, 1967; DODDS; MUHAMAD; WATTS, 2003).

There are different realizations of the small-world model but the original model as proposed by Watts and Strogatz (1998) is by far the most widely studied. It starts by comprising $|V|$ vertices so that each vertex connects to its k nearest neighbors, totalizing $2k$ connections. After that, given an arbitrary vertex $x \in V$, an edge (x,y) belonging to its $2k$ connections is randomly selected considering that $y \in V$. The selected edge is randomly relocated, such that the destination from vertex y is switched to another vertex $z \in V$, $y \neq z$, with probability p.

In Figure 12, we illustrate the small-world network formation process, initializing from a 4-regular graph, i.e. each vertex from the network is connected to its $k = 2$ nearest neighbors establishing a total of $2k = 4$ connections. This network is regular due to that when $p = 0$, no rearrangements are performed and, therefore, the network continues to be regular. As p increases, but still remains small, the property of small-world becomes apparent. When $p = 1$, the network turns out to similar, though not identical, to a random network.

One of the most important properties of small-world networks is the locally clustering property: even for small p values, the small-world network becomes a locally clustered network in which two arbitrary vertices are connected by a small number of intermediate edges. For further details about small-world network properties, one can refer to Barthelemy and Amaral (1999), Telesford *et al.* (2011), and Watts (2003).

Figure 12 – An example of the formation process of a small-world network. Increasing the rewiring probability p implies that the network moves from a regular network (4-regular) to a random network.



Source: Elaborated by the author.

Scale-Free Networks

Scale-free model arose due to that the observation that in many real world networks some vertices act as "highly connected hubs", i.e. a few amount of vertices connected to a large number of neighbors in terms of a magnitude larger than the average value. This fact is clearly in contrast to random and small-world networks where every vertex has roughly the same number of neighbors.

With this observation, Barabási and Albert (1999) proved that the degree distribution of scale-free networks obeys a power-law, as follows:

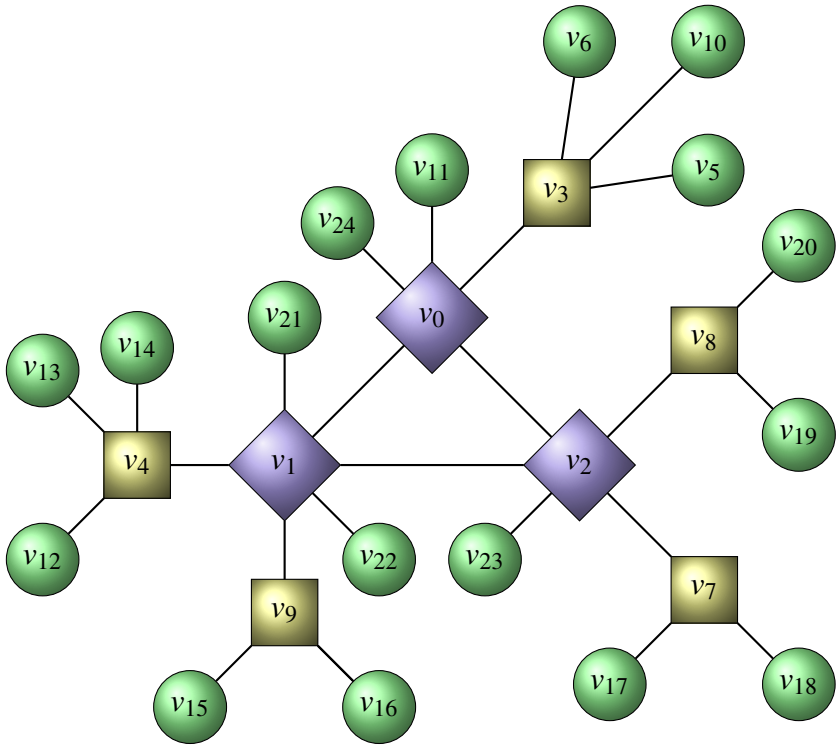
$$p(k) \sim k^{-\gamma}. \tag{2.12}$$

where γ is called the scaling exponent. Note that, by setting a fixed value for γ, as the degree k grows, the number of vertices that have degree k decreases. Therefore, it is expected that $p(k)$ will have a large value for small values of k and a small value for large values of k, which is consistent with the observation of the presence of hubs.

One of the most important properties of scale-free networks is its strongly correlates with the network robustness to failure. In a scale-free network topology, it turns out that major hubs are closely followed by smaller ones. In turn, these smaller hubs are followed by other vertices with an even smaller degree and so on until we reach peripheral or terminal vertices. Figure 13 depicts an example of scale-free network where is possible to observe few vertices with large degree (hubs), while the great majority has small degree (terminal vertices).

Hubs are both a strength and a weakness of scale-free networks. If failures occur at

Figure 13 – An example of a scale-free network. The hubs, corresponding to the vertices with large degrees, are in diamond format.



Source: Elaborated by the author.

random and the majority of vertices are those with small degree, the chance that a hub would be affected is almost negligible. Even if a hub-failure occurs, the network generally does not lose its connectedness, due to the remaining hubs. On the other hand, if we choose a few major hubs and take them out of the network, the network is turned into a set of rather isolated graphs.

The formation of scale-free networks happens due to other of its property: *the preferential attachment of vertices*. This property implies that the number of vertices in the network increases over time. Preferential attachment means that the more connected a vertex is, the more likely it will be connected to new vertices. Based on this property, Barabási and Albert (1999) proposed an algorithm to generate scale-free networks. The network begins with an initial connected network of $|V_0|$ vertices. New vertices are added to the network one at a time considering that each new vertex x must be connected to $|V| < |V_0|$ exiting vertices with a probability p_x. This probability, which is proportional to the number of edges that the existing vertices already have, is defined as stated in Equation 2.13.

$$p_x = \frac{k_x}{\sum_{y \in V} k_y}. \tag{2.13}$$

Therefore, heavily connected vertices or hubs tend to quickly accumulate even more links, while vertices with only a few links are unlikely to be chosen as the destination for new links. The new vertices have a “preference” to attach themselves to the already heavily connected

vertices. Such preference leads to the "the rich get richer" effect.

For further details about other properties of scale-free networks, one can refer to Callaway *et al.* (2000), Cohen *et al.* (2001), Ravasz and Barabási (2003), and Choromanski, Matuszak and Miekisz (2013).

2.1.4 Complex Network Measures

Centrality measures quantify the intuitive notion we have that in most networks some vertices are more prominent than others. In this section, we review some of the main network centrality measures proposed in the complex network literature (FREEMAN, 1977; WATTS; STROGATZ, 1998; GIRVAN; NEWMAN, 2002; CLAUSET; NEWMAN; MOORE, 2004; NEWMAN, 2002; NEWMAN, 2003; NEWMAN, 2004; NEWMAN, 2006; NEWMAN, 2010; KOSCHÜTZKI *et al.*, 2005; OPSAHL; PANZARASA, 2009; WANG; FLEURY, 2013; SILVA; ZHAO, 2016; BARABÁSI, 2016).

Degree-based Measures

Some measures related to degree and degree-correlation, which are used throughout this thesis, will be briefly presented in this section.

Definition 2.35. Universal Set. The universal set U is the set containing the total possible edges of a graph. Therefore, if the graph is undirected the size of the universal set is $|U| = \frac{|V|(|V|-1)}{2}$ if self-loops are not allowed, and $|U| = \frac{|V|^2}{2}$ if the graph allows self-loops. If the graph is directed, then $|U| = |V|(|V|-1)$ if self-loops are not allowed, and $|V|^2$ in opposite.

Definition 2.36. Degree Heterogeneity. The degree heterogeneity H of a graph measures how different (or similar) are its vertices according to their degrees. Graphs whose vertices have very different values of each other, tend to have a high degree heterogeneity. Therefore, the degree heterogeneity can be calculated as stated in Equation 2.14.

$$H = \frac{\langle k^2 \rangle}{\langle k \rangle^2}. \tag{2.14}$$

***Remark* 2.26.** The higher the H value, the larger the degree heterogeneity.

Definition 2.37. Density. The network density D measures how strong the vertices of a graph are connected. It is defined as the fraction of actual connections over the total possible connections according to the Equation 2.15.

$$D = \frac{|E|}{|U|}. \tag{2.15}$$

***Remark* 2.27.** The density assumes values in the interval $[0,1]$. When $D = 0$, the network is empty. Conversely, when $D = 1$ the network is a complete or maximal clique.

***Remark* 2.28.** If the density is near 0, the network is classified as sparse, and dense, otherwise.

Definition 2.38. Network Assortativity. The network assortativity is the degree of correlation among vertices properties. It captures the preference of vertices to attach to others that are similar or different in terms of the given property. Regarding the degree centrality, the assortativity coefficient r is essentially the Pearson correlation coefficient of degree between pairs of connected vertices. Therefore, considering that u_e and v_e are the degrees of the two vertices at the endpoints of the e-th edge, the network assortativity r is calculated as stated in Equation 2.16.

$$r = \frac{|E|^{-1}\sum_{e\in E} u_e v_e - (\frac{|E|^{-1}}{2}\sum_{e\in E}(u_e+v_e))^2}{\frac{|E|^{-1}}{2}\sum_{e\in E}(u_e^2+v_e^2) - (\frac{|E|^{-1}}{2}\sum_{e\in E}(u_e+v_e))^2}. \tag{2.16}$$

Positive values of r indicate a correlation between vertices of similar degree, while negative values indicate relationships between vertices of different degrees. Network assortativity lies between -1 and 1. When $r = 1$, the network have perfect assortative mixing patterns, while at $r = -1$ the network is completely disassortative.

Distance and Path Measures

Some measures related to calculation of paths and distances between vertices, which are used throughout this thesis, will be briefly presented in this section.

Definition 2.39. Diameter. The diameter T of a graph is defined as the largest distance between all the different pairs of vertices in the graph, as stated in Equation 2.17.

$$T = \max_{x,y\in V \,\wedge\, x\neq j} d_{x,y}. \tag{2.17}$$

***Remark* 2.29.** For a non-weighted network, the feasible values of T are $[0, |V|-1]$.

Definition 2.40. Vertex Eccentricity. The eccentricity of a vertex $x \in V$, ρ_x, is the largest distance from x to any other vertex $y \in V \setminus \{x\}$, as stated in Equation 2.18.

$$\rho_x = \max_{y\in V\setminus\{x\}} d_{x,y}. \tag{2.18}$$

Definition 2.41. Radius. The network radius, ζ, is its minimum eccentricity, as stated in Equation 2.19.

$$\zeta = \min_{x\in V} \rho_x. \tag{2.19}$$

Definition 2.42. Average Shortest Path Length. The average shortest path length of a graph, denoted by $\langle \ell \rangle$, is the average of the sum of the distances calculated between all the pairs of different vertices in the graph, as stated in Equation 2.20.

$$\langle \ell \rangle = \frac{1}{|U|} \sum_{x\in V \,\wedge\, y\in V \,\wedge\, x\neq j} d_{x,y}. \tag{2.20}$$

Structural Measures

Some measures related to the topological structure of the networks, which are used throughout this thesis, will be briefly presented in this section.

Definition 2.43. Clustering Coefficient. The clustering coefficient quantifies the degree to which nodes in a network tend to cluster together, i.e. nodes tend to create tightly knit groups characterized by a relatively high density of connections. Therefore, the local clustering coefficient of a node quantifies how close its neighbors are to being a clique (complete graph). The local clustering coefficient of node $x \in V$ is given by:

$$CC_x = \frac{2|e_x|}{k_x(k_x-1)}, \tag{2.21}$$

where $|e_x|$ is the number of edges shared by the direct neighbors of node x, i.e. the number of triangles formed by node x and any of its two neighbors. The feasible values of local clustering coefficient are $[0,1]$.

Definition 2.44. Global Clustering Coefficient. The global clustering coefficient is the average of the local clustering clustering coefficient of all nodes in the network, as stated in Equation 2.22.

$$CC = \frac{1}{|V|}\sum_{x\in V} CC_x. \tag{2.22}$$

The global clustering coefficient tells how well-connected the neighborhood of the node is, and their feasible values are $[0,1]$. If the neighborhood is fully connected, the global clustering coefficient is 1 and a value close to 0 means that there are hardly any triangular connections in the neighborhood.

Definition 2.45. Modularity. The network modularity quantifies how good a particular division of a network is, i.e. it measures the strength of division of a network into modules, which also are called groups, clusters or communities. Therefore, for non-weighted networks, the modularity is calculated as follows:

$$Q = \frac{1}{2|E|}\sum_{x,y\in V}\left(\mathbf{A}_{x,y} - \frac{k_x k_y}{2|E|}\right) 1_{[c_x=c_y]}, \tag{2.23}$$

where c_x is the community to which the vertex x belongs. The summation term is composed by two factors, all of which are computed only for vertices of the same community due to the indicator function (indicated by $1_{[c_x=c_y]}$). That is, cross-community edges do not contribute to the modularity measure.

Modularity values range from 0 to 1. When the modularity is near 0, it means that the network does not present community structure, suggesting that the edges are disposed at random in the network. Nonzero values indicate deviations from randomness being that as the modularity grows, the community structure gets more and more defined. Values around 0.3 or more usually indicate good divisions.

***Remark* 2.30.** For a weighted network, the terms denoting k_x in Equation 2.23 are exchanged for the strength s_x, as introduced in Definition 2.15, and $|E| = \frac{1}{2}\sum_{x\in V} s_x$.

Definition 2.46. Intracommunity and Intercommunity Edges. Edges formed by vertices belonging to the same community are called intracommunity edges, whilst edges formed by vertices belonging to different communities are called intercommunity edges. The total number of intracommunity and intercommunity edges is represented by z_{in} and z_{out}, respectively, and $|E| = z_{in} + z_{out}$. On the basis of these parameters, we can define the fraction of intracommunity and intercommunity edges, as stated in Equations 2.24 and 2.25, respectively.

$$e_{in} = \frac{z_{in}}{\langle k \rangle}, \tag{2.24}$$

$$e_{out} = \frac{z_{out}}{\langle k \rangle}. \tag{2.25}$$

High values of e_{in} and low values of e_{out} refer to networks with well-defined communities, i.e. there is a high concentration of edges confined within each community and very few edges interconnecting different communities. On the other hand, low values of e_{in} and high values of e_{out} imply in the presence of communities highly mixed with each other.

Definition 2.47. Topological Overlap. The topological overlap measures to what extent two vertices are connected to roughly the same group of other vertices in the network, i.e. the topological overlap evaluates how similar the direct and indirect neighborhoods of two vertices are. To calculate the topological overlap of a pair of vertices, their connections to all of the other vertices in the network are compared. If these two vertices share similar direct and indirect neighborhoods, then they have a high topological overlap. There, is possible to adjust the depth of the neighborhood which is used in the comparison, i.e. we can compare the direct neighborhood of two vertices up to m-th order, with $m > 0$. Therefore, let $\mathcal{N}_m(x)$ denote the set of vertices (excluding x itself) that is reachable from x within a shortest path of length m, i.e. $\mathcal{N}_m(x) = \{y \neq x \mid d_{x,y} \leq m\}$, the m-step topological overlap is given by:

$$t_{x,y}^{[m]} = \begin{cases} \dfrac{|\mathcal{N}_m(x) \cap \mathcal{N}_m(y)| + \mathbf{A}_{x,y}}{\min(|\mathcal{N}_m(x)|,|\mathcal{N}_m(y)|) + 1 - \mathbf{A}_{x,y}}, & if \;\; x \neq y \\ 1, & if \;\; x = y. \end{cases} \tag{2.26}$$

Note that, even in the case that two vertices have the same m-step neighborhoods, the topological overlap only assumes its maximum value when they are directly connected, i.e. when $\mathbf{A}_{x,y} = 1$.

Centrality Measures

Centrality measures quantify how central or important vertices are into a network. The first centrality measure is the vertex degree from which it is natural to assume that vertices with

large degrees are hubs or central to the network, while vertices with small degrees are usually peripheral or terminal ones. In this section, we present other centrality measures which are used throughout this thesis.

Definition 2.48. Closeness or Minisum Criterion. The closeness optimizes a minisum criterion to determine the location of a service facility into the network. The closeness of a vertex $x \in V$ is calculated as follows:

$$c_{Cx} = \frac{1}{\sum_{y \in V} d_{x,y}}. \tag{2.27}$$

Vertices with a small value of closeness are considered as more important as those with a high value.

Definition 2.49. Betweenness. The betweenness measures the extent to which a vertex lies on the shortest paths between every pair of vertices in a network. The betweenness for the vertex x is the number of these shortest paths that pass through it, and is given by:

$$B_x = \sum_{y \neq x \in V} \sum_{z \neq x \in V} \frac{\eta^x_{y,z}}{\eta_{y,z}}, \tag{2.28}$$

where $\eta_{y,z}$ is the total number of shortest paths from y to z, and $\eta^x_{y,z}$ is the number of such shortest paths that pass through the vertex x. Vertices with high betweenness may have considerable influence within a network by virtue of their control ability over information passing between others. The vertices with the highest betweenness are also the ones whose removal from the network will most disrupt communications between other vertices because they lie on the path of several messages.

2.1.5 Categories of Complex Networks

As previously established, complex systems from different real-world domains can be modelled and analysed as complex networks. Therefore, complex networks can be classified according the types of information about these systems. In this section, we introduce briefly five of the most popular categories of complex networks and some of the most important networks included within each type (NEWMAN, 2010; YAVEROGLU, 2013).

Biological Networks

Some complex biological systems may be represented and analyzed as networks. Therefore, nodes can represent a wide-array of biological units, from individual organisms to individual neurons in the brain, as well as edges can represent the different kinds of interactions among these biological units.

The most representative biological networks are: protein-protein interaction networks, metabolic interaction networks, and disease interaction networks, each one with its own peculiarities. For instance, *Protein-Protein Interaction (PPI) networks* represent the binding information

among all proteins of an organism, where nodes represent the proteins and edges represent physical interactions (bindings) between two proteins (SZKLARCZYK *et al.*, 2011; KOTERA *et al.*, 2012; SINGH-BLOM *et al.*, 2013; MENCHE *et al.*, 2015).

Metabolic networks explain the collection of all biochemical reactions that occur in a cell. A metabolic network is a bipartite network of metabolites and reactions, where each metabolite is connected with the reactions that it is involved in. The biochemical reactions are represented by directional edges since they represent chemical conversion of the metabolites from one form to another. However, most biochemical reactions are bidirectional. There is important to note that, the metabolic network of all species can be viewed as a very large single network that contains all possible reactions in all species, i.e. bipartite networks in the form of metabolite-enzyme, metabolite-protein, metabolite-gene, and others (JEONG *et al.*, 2000; TANAKA, 2005).

Diseases have been grouped and studied in terms of the similarities of their symptoms and the organs they affect. Therefore, *disease interaction networks* can be divided into: i) *disease association networks*, in which vertices correspond to diseases and two vertices are connected when the two corresponding diseases are linked with at least one gene in common (GOH *et al.*, 2007), ii) *disease-disease networks*, in which two vertices representing different diseases are linked if they occur in the same person at the same time (HIDALGO *et al.*, 2009), and iii) *disease-drug networks*, which connect the genomic expression profiles of human diseases and drugs (HU; AGARWAL, 2009).

Despite the fact that many efforts of research and applications are focused on the understanding of previously cited networks, there exists other biological networks that are receiving increasing attention, such as: protein structure networks (MURZIN *et al.*, 1995), genetic interaction networks (DIXON *et al.*, 2009), transcriptional regulatory networks (SHEN-ORR *et al.*, 2002), and signal transduction networks (SCHACHERER *et al.*, 2001).

Technological Networks

Technological networks are physical networks that form the backbone of modern technological societies. The most representative technological networks are: the Internet, telephone networks, power grids, transportation networks, and distribution networks.

One of the best examples of technological networks is the *Internet*, in which nodes corresponding to electronics such as computers, laptops, mobile phones, satellites with different IP addresses, and edges corresponding to direct physical communication channels among them (FALOUTSOS; FALOUTSOS; FALOUTSOS, 1999; CHEN *et al.*, 2002; YOOK; JEONG; BARABáSI, 2002; PASTOR-SATORRAS; VESPIGNANI, 2004).

The *telephone network* is made of landlines and wireless links that transmit telephone calls. Due to the advance of the technology, telephone network nodes evolved from circuit switched to digital packet switched, rather similar to that of the Internet. Therefore, the telephone

network and the Internet are not disjoint networks, and in the last years we are viewing as this two networks are merging into a single network (AIELLO; CHUNG; LU, 2002).

A *power grid network* is the network of high-voltage transmission lines that provide long-distance transport of electric power within and between countries. Low-voltage local power delivery lines are normally excluded. Therefore, the vertices of a power grid network correspond to generating stations and switching substations, and the edges correspond to the high-voltage lines. Failures on power grids may have cascading effects, i.e. the failure of one node may recursively provoke the failure of connected nodes (WATTS; STROGATZ, 1998; AMARAL *et al.*, 2000).

Transportation networks include airline routes, road and rail networks, whilst *distribution networks* include oil and gas pipelines, water and sewerage lines, and routes used by the post office and package delivery companies (AMARAL *et al.*, 2000; KALAPALA *et al.*, 2006). One class of distribution networks that has been relatively well studied is river networks, where the edges are rivers or streams and the vertices are their intersections (DODDS, 1969; MARITAN *et al.*, 1996).

One interesting feature about all the technological networks is related to the fact that their structures are clearly governed to some extent by space and geography, i.e. vertices of technological networks are connected to others based on a function of what is technologically desirable and what is geographically feasible (YOOK; JEONG; BARABáSI, 2002).

Information Networks

Information networks, also called knowledge networks, are those which their vertices store some specific type of information. Two of the most famous information networks are the citation networks and the World Wide Web (WWW). A *Citation network* is made of academic papers, i.e. nodes are articles and a directed edge from article A to article B indicates that A cites B. Therefore, citation networks are acyclic because papers can only cite previously published papers, making closed loops impossible or at least extremely rare (WHITE; WELLMAN; NAZER, 2004; PRICE, 1965; YAN; DING, 2012; GOLDBERG; ANTHONY; EVANS, 2015).

The *World Wide Web* (WWW) is an information network of nodes representing web pages containing information, and edges formed by hiperlinks which link one page to another. The WWW should not be confused with the Internet, which is a technological network as previously presented. Unlike a citation network, the WWW is cyclic due to that there is no natural ordering of web sites and no constrains that prevent closed loops (KLEINBERG *et al.*, 1999; HUBERMAN, 2001; MIRZAL, 2010; BRISABOA; LADRA; NAVARRO, 2014).

A few other examples of information networks have been studied to a lesser extent, such as: peer-to-peer networks (IAMNITCHI; RIPEANU; FOSTER, 2002; ADAMIC; LUKOSE; HUBERMAN, 2005; TANTA-NGAI; MILIOS; KESELJ, 2009), semantic word networks (SIG-

MAN; CECCHI, 2002; STEYVERS; TENENBAUM, 2005; DEYNE; G., 2008), and preference networks (KAUTZ; SELMAN; SHAH, 1997; TRUYEN; PHUNG; VENKATESH, 2007).

Economic Networks

Economic networks represent different types of complex micro-scale and macro-scale economic information. Among the most studied economic networks we have: interbank relation networks, investment networks, supply-chain networks, and world-trade networks. *Interbank relation networks* are networks where nodes represent banks and the edges represent the credit-debt relations among them (BOSS *et al.*, 2004; BARGIGLI *et al.*, 2015). *Investment networks*, also called inter-company networks, are networks in which nodes represent companies and edges link with companies that co-invest on the same portfolio (BYGRAVE, 1988; NEVILLE; JENSEN, 2000; BERNSTEIN; CLEARWATER; PROVOST, 2003; BATTISTON; RODRIGUES; ZEYTINOGLU, 2007; ZHANG *et al.*, 2015; ZHANG *et al.*, 2016).

Supply-chain networks are formed by nodes corresponding to organisations/companies and edges representing the flow and movement of materials and information among them (CHOI; DOOLEY; RUNGTUSANATHAM, 2001; CHOI; HONG, 2002; SURANA *et al.*, 2005). The most popular among the economic networks is the *world trade network*, in which nodes correspond to countries and edges to the trade links among them (FAGIOLO; REYES; SCHIAVO, 2008; FAGIOLO; REYES; SCHIAVO, 2009; LI; JIN; CHEN, 2003; CINGOLANI; PICCARDI; TAJOLI, 2015; ZHANG *et al.*, 2016).

Social Networks

A *social network* is a set of entities or groups of sets of entities with some pattern of contacts or interactions between them, i.e. such patterns are into a social context. Depending of the type of social pattern, different social networks have been identified: friendship networks (WANG; WELLMAN, 2010; TRAUD; MUCHA; PORTER, 2012; XIE *et al.*, 2012; LIANG; WANG; ZHU, 2016; LEE; LIM, 2016), collaboration networks (DRABEK, 1981; CAMARINHA-MATOS; AFSARMANESH, 2007; DALL'ASTA; MARSILI; PIN, 2012), e-mail networks (TASHIRO *et al.*, 2010; KOLLI; NARAYANASWAMY, 2013), co-authorship networks (BARABÁSI *et al.*, 2002; SUN *et al.*, 2011; YAN; DING, 2012), co-purchasing networks (LESKOVEC; ADAMIC; HUBERMAN, 2007; ARTHUR *et al.*, 2009; DINH *et al.*, 2014), and others.

Among the different types of social networks, friendship networks, in which nodes represent people and edges connect people with friendship relations, are the popular ones due to the recent boom in *online social networking* (OSN) sites, such as: Facebook[1], Google+[2], Instagram[3], and others. The recent developments in OSNs raised a new set of interesting network

[1] <https://www.facebook.com/>
[2] <https://plus.google.com/>
[3] <https://www.instagram.com/>

analysis issues focused on understand, mainly, the main characteristics of social networks as well as the principles governing the evolution of these networks.

2.1.6 Recent Trends

Due to the inability of the standard complex network theory to describe some specific scenarios observed in complex real-world scenarios, network theory has been extended within the last years to include concepts such as temporal and multilayer networks.

Temporal Networks

Temporal network is characterized by the fact that the connections among vertices are not continuously active, i.e. some edges can be part of the graph by a period of time disappearing later, and eventually reappearing after other period of time. As an example, in networks of communication via e-mail, text messages, or phone calls, edges represent sequences of instantaneous contacts. Sometimes, edges are active for non-negligible periods of time, for instance in a network of hospital patients, where relationships are established between individuals while they are in the same ward.

The temporal structure of edge activations can affect the dynamics of the elements interacting through the network. Therefore, greater efforts should be made to better understand the structure and applications of this network. For further details on temporal networks one can refer to Holme and Saramäki (2012), Holme and Saramäki (2013), Holme (2015), Zhang *et al.* (2015) and Williams and Musolesi (2016).

Multilayer Networks

Multilayer network is a specific type of heterogeneous graph (see Definition 2.8). This type of network emerged by observing that connections between nodes from a single graph are seldom of a single type, i.e. in multilayer networks several different layers of connections are taken into account.

Multilayer networks explicitly incorporates multiple channels of connectivity and therefore constitute the natural framework to describe systems interconnected through different types of connections. A layer represents a specific type of edge, e.g. relationship, activity, etc. Therefore, the same node in the network may have different types of interactions, i.e. different set of neighbors in each layer. For example, in social networks, we can consider different types of relationships for the same people into the network, i.e. friendship, kinship, partnership, job contact, and others.

Despite multilayer network research constitutes a new frontier in many areas of science and it will rapidly attract more attention of the research community, one of the major efforts into this research field is related to provide proper and suitable frameworks to lead with the topology

of this type of networks. For further details about multilayer networks one can refer to Kivela *et al.* (2014), Boccaletti *et al.* (2014), Domenico *et al.* (2015), Zanin (2015) and Diakonova *et al.* (2016).

2.2 Link Prediction

Link prediction is a long-standing challenge in modern information science. It is the fundamental problem faced in *link mining*, a recently emerged research field with special emphasis in links, because they are considered the most important in data analysis. In simple terms, link prediction attempts to estimate the likelihood of the existence of a link between two nodes based on the observed links and on vertices attributes (GETOOR; DIEHL, 2005; LIBEN-NOWELL; KLEINBERG, 2007; LÜ; ZHOU, 2011; WANG *et al.*, 2014).

Link prediction algorithms are used to extract missing information, identify spurious interactions, evaluate network evolving mechanisms, among others. Therefore, link prediction has a range of applications in different domains (SRINIVAS; MITRA, 2016), e.g. bioinformatics, to discover genetic or protein-protein interactions (KOTERA *et al.*, 2012; SINGH-BLOM *et al.*, 2013; MENCHE *et al.*, 2015); security, to identify groups of terrorist or criminals (HASAN *et al.*, 2006; ANIL *et al.*, 2015); information retrieval and extraction, to predict words, topics, or documents in very large collections of documents (ARNOLD; COHEN, 2009; ITAKURA *et al.*, 2011; LI *et al.*, 2016); and recommender systems, to suggest general items and friendships, popular users, and media content (CHILUKA; ANDRADE; POUWELSE, 2011; WEI *et al.*, 2013; LI *et al.*, 2014; BARBIERI; BONCHI; MANCO, 2014; AHMED; ELKORANY, 2015).

This section provides a comprehensive review for link prediction in social networks encompassing standard as well as the latest link prediction techniques, challenges, and applications. Section 2.2.1 starts with the definition of link mining in contrast with traditional data mining, and defines link prediction as a link mining task. Section 2.2.2 provides the link prediction statement including its formal definition, general solution process, and evaluation measures. Section 2.2.3 analyzes several standard and new link prediction techniques. Finally, Section 2.2.4 concludes with an overview of typical link prediction applications.

2.2.1 Link Prediction as a Link Mining Task

Traditional data mining algorithms commonly attempt to find patterns in a attribute-value dataset, which is characterized by a collection of independent instances of a single relation. This is consistent with the classical problem related to the identification of a model given an *independent, identically distributed sample* (IID) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; HAND; SMYTH; MANNILA, 2001; ZAKI; JR., 2014). This process might be applied to a network in order to infer the model for its nodes attributes, while ignoring the links between

them. However, the naïve application of traditional statistical inference procedures may lead to inappropriate conclusions (JENSEN, 1999).

It is a fact that the elements of a network intrinsically store information, which means that the attributes of vertices are correlated and links are more likely to be observed between vertices with common properties; therefore, it is possible to exploit this feature to improve the predictive accuracy of learning models (VALVERDE-REBAZA *et al.*, 2014). Furthermore, the network structure itself may be a meaningful element in such models. In this context, link mining (LM) has emerged as a research area whose purpose is to develop predictive or descriptive models that can be applied to networks taking maximum advantage of their elements and structure (GETOOR; DIEHL, 2005; SRINIVAS; REDDY; GOVARDHAN, 2010; ALAVIJEH, 2015; ALI, 2016).

It is important to note that LM focuses on techniques that explicitly consider links when building predictive or descriptive models. Links (also called edges, relationships, or ties) are perhaps the most important elements of networks; they are the core elements for the definition of network topology, measurements, and categories. Links often exhibit patterns that can indicate properties such as importance, rank, or category of the nodes. In some cases, not all links are visible; therefore, it might be interesting (or necessary) to predict the existence of links between nodes. When links are considered, more complex patterns also arise; that might lead to other challenges related to discovering substructures, such as communities, groups, or common subgraphs (GETOOR; DIEHL, 2005).

LM encompasses a wide range of tasks focused on exploring the link structure to discover and/or understand patterns and behaviors of nodes and links, as well as patterns and behaviors of entire networks. Based on the taxonomy presented by Getoor and Diehl (2005), Chart 2 shows the eight main LM tasks grouped in their respective high-level tasks.

Chart 2 – The taxonomy of link mining tasks.

High-level Task	**Task**
Object-related	Link-Based Object Ranking
	Link-Based Object Classification
	Object Identification
	Object Clustering
Link-related	Link Prediction
Graph-related	Subgraph Discovery
	Graph Classification
	Generative Models for Graphs

Source: Getoor and Diehl (2005).

Object-related tasks aim to understand the different patterns inherent to nodes in order to build predictive or descriptive models capable of identifying the specific characteristics of

each node or group of nodes based on their relationships. This high-level task includes: i) *link-based object ranking* (MEDO, 2013; SUN; HAN, 2014; ROA-VALVERDE; SICILIA, 2014; MARIANI; MEDO; ZHANG, 2015); ii) *link-based object classification* (LU; GETOOR, 2003; MACSKASSY; PROVOST, 2007; ROSSI; ZHOU; AHMED, 2016; BERTON *et al.*, 2017; FALEIROS; ROSSI; LOPES, 2017); iii) *object identification*, also called entity resolution or graph alignment (LIU *et al.*, 2013; ZHANG *et al.*, 2014; NIE *et al.*, 2016; CAO; YU, 2016); and *object clustering*, also called group detection or community detection (NEWMAN, 2004; PAN *et al.*, 2010; XIE; KELLEY; SZYMANSKI, 2013; VALEJO; VALVERDE-REBAZA; LOPES, 2014; BEDI; SHARMA, 2016).

Graph-related tasks aim to find interesting patterns of subgraphs, or understand behaviors that define the networks as a whole. This high-level task includes: i) *subgraph discovery* (INOKUCHI; WASHIO; MOTODA, 2000; MOTODA, 2007; REHMAN *et al.*, 2014; THOMAS; NAIR, 2016); ii) *graph classification* (JIN; YOUNG; WANG, 2010; LI *et al.*, 2012; WU *et al.*, 2016); and iii) *generative models for graphs* (GETOOR *et al.*, 2003; LESKOVEC *et al.*, 2010; DAVIS *et al.*, 2014; ALAM *et al.*, 2016; HADIAN *et al.*, 2016).

Link-related tasks aim to understand all kinds of behaviors and phenomena specific to the links. *Link prediction* is the main task performed by this high-level task. This section will provide an extensive discussion on link prediction task.

2.2.2 Problem Statement

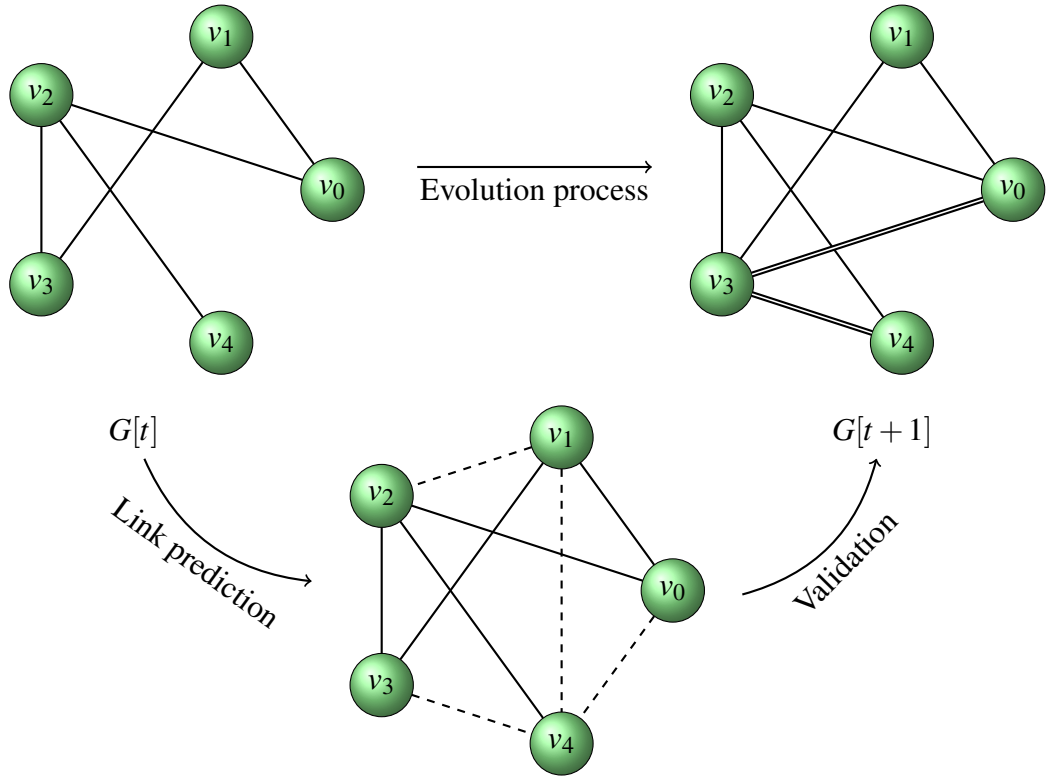
Consider a network $G = (V, E)$ at a particular time t, where V and E are the sets of nodes and links, respectively. The purpose of the link prediction problem is to predict, among all possible pairs of nodes that have not established any connections, those that will have a future association at time $t+1$. The link prediction problem also attempts to estimate the likelihood of missing or unobserved links in the current network (LIBEN-NOWELL; KLEINBERG, 2007; LÜ; ZHOU, 2011; WANG *et al.*, 2014).

Figure 14 explains the link prediction problem. At time t, the undirected network G has a specific structure in which solid links indicate existing interactions among nodes. Following a natural evolution process, the network G at time $t+1$ has a new structure, showing new links (represented by double lines). Therefore, a link prediction algorithm is applied over network G at time t to estimate which links will appear (links represented by dashed lines). A validation process can be performed by comparing the predicted links to the actual new links in network G at time $t+1$.

In this thesis, link prediction will be performed according to the following assumptions (unless otherwise described):

1. The network is undirected and unweighted.

Figure 14 – Description of link prediction problem.



Source: Elaborated by the author.

2. Multiple links and self-loops are not allowed.

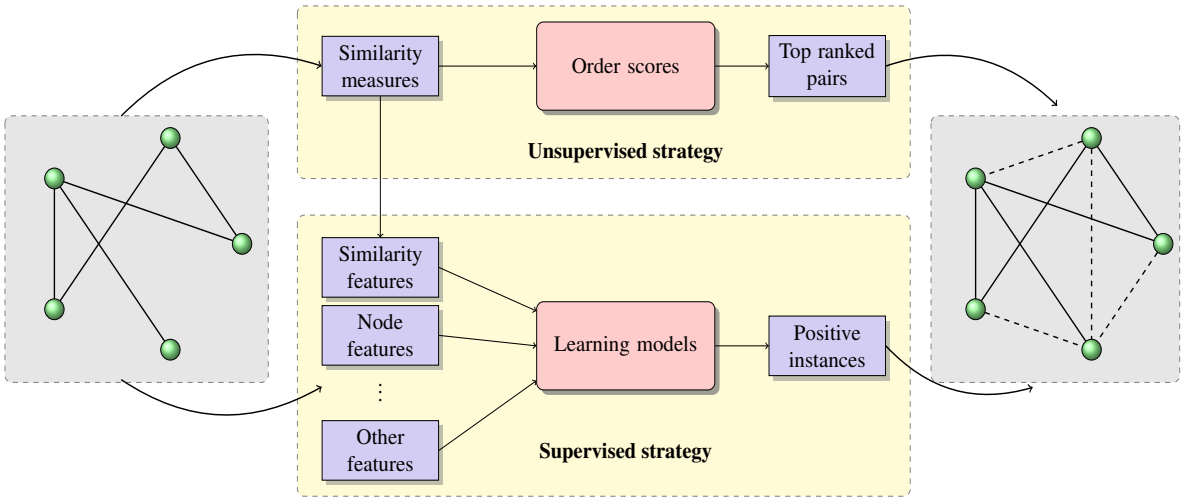
3. The number of nodes at different time of the network evolution process is static.

The likelihood of formation or dissolution of links between all node pairs must be determined to address the link prediction problem. Such likelihoods can usually be calculated following a supervised or unsupervised strategy. Figure 15 shows a network G at time t as input of a link prediction process, which can apply a supervised or unsupervised strategy to obtain a set of predicted links. This section will describe how these strategies work.

Unsupervised Link Prediction

Link prediction is naturally an unsupervised problem. Consider as a *potential link* any pair of disconnected nodes $x, y \in V$ such that $(x, y) \notin E$. Given the universal set U (see Definition 2.35), a *missing link* is any potential link in the set of nonexistent links $U - E$. Then, the fundamental link prediction task, here, is to find out the missing links in the set of nonexistent links, assessing for each link in this set a score. The higher the score, the more likely the link will be. Thus, a *predicted link* is any potential link that has received a score, higher than zero, by any link prediction algorithm. The scores represent the connection probability of pairs of disconnected nodes. Then a ranked list in decreasing order of scores is obtained and links at the

Figure 15 – The generic link prediction process.



Source: Elaborated by the author.

top of list are most likely to appear (LIBEN-NOWELL; KLEINBERG, 2007; LÜ; ZHOU, 2011; WANG *et al.*, 2014; VALVERDE-REBAZA *et al.*, 2016).

Most of these link prediction algorithms are based on specific measures that capture the existing similarity or proximity between nodes. Measures based on similarity are the most used in the literature mainly due to its low computational cost and easy calculation, which make them candidates approaches for real-world applications (LÜ; ZHOU, 2011; SRINIVAS; MITRA, 2016).

Evaluation Measures

To quantify the performance of any link prediction algorithm in unsupervised context it is necessary to investigate the adequacy of some standard evaluation measures (LICHTENWALTER; LUSSIER; CHAWLA, 2010; ALLALI; MAGNIEN; LATAPY, 2013; NAUDÉ; GREYLING; VOGTS, 2015).

Consider that every unsupervised link prediction algorithm generates a set of predicted links, L_p, which represents the prediction space of the algorithm. A prediction space is characterized by contains a small amount of pairs of nodes that really will connect and a huge amount of pairs of users that will never establish a friendship. An extremely skewed distribution of classes can have a negative influence on the performance of link prediction methods (SCELLATO; NOULAS; MASCOLO, 2011).

Efforts are focused to not only reduce the number of wrong predicted links but also, obviously, to increase the number of accurate ones. However, some link prediction algorithms can generate a smaller amount of correctly predicted links than other methods, but these links may have a higher chance of actually appear in the future. That is, assessing the quality of a

link prediction algorithm is not an easy task (LICHTENWALTER; LUSSIER; CHAWLA, 2010; YANG; LICHTENWALTER; CHAWLA, 2015).

Assuming we know the set of future new connections that truly will appear between pair of nodes, E^P, where $E^P \subset U - E$, i.e. E^P is part of the set of nonexistent links. We define the set of *true positives*, TP, as all correctly predicted links, the set of *false positives*, FP, as all wrongly predicted links, and the set of *false negatives*, FN, as all truly new links that were not predicted. Clearly, $L_p = TP \cup FP$ and $FN = E^P - TP$.

To better understanding the performance of link prediction algorithms, we divide the different evaluation measures in two groups: i) measures to analyze the prediction space and, ii) measures to analyze the predictive power.

Evaluation Measures to Analyze the Prediction Space

Evaluation measures to analyze the prediction space aim to quantify the relation between wrong and correct predictions into the overall prediction space size of unsupervised link prediction algorithms. Some evaluation measures to analyze the prediction space are:

Imbalance Ratio. This is the fraction between the total number of predicted links and the number of correct predictions within it (SCELLATO; NOULAS; MASCOLO, 2011; LIBEN-NOWELL; KLEINBERG, 2007). The imbalance ratio (IR) is computed as:

$$IR = \frac{|L_p|}{|TP|}. \tag{2.29}$$

This measure expresses how many predictions should be computed, on average, before find a correct prediction. The lower the imbalance ratio higher the number of right predictions into a reduced prediction space.

Precision. This is the proportion of correctly predicted links into the prediction space (FATOURECHI *et al.*, 2008; PHAM; SHAHABI; LIU, 2013; VALVERDE-REBAZA *et al.*, 2016). This measure is computed as:

$$P = \frac{|TP|}{|TP| + |FP|}. \tag{2.30}$$

Recall. This is defined as the proportion of correctly predicted links into the total number of truly new links (FATOURECHI *et al.*, 2008; PHAM; SHAHABI; LIU, 2013; VALVERDE-REBAZA *et al.*, 2016), as stated in Eq. 2.31.

$$R = \frac{|TP|}{|TP| + |FN|}. \tag{2.31}$$

F-measure. This is defined as the harmonic mean of precision and recall (RIJSBERGEN, 1979; FATOURECHI *et al.*, 2008). This measure is computed as:

$$F_1 = 2 \times \frac{P \times R}{P + R}. \tag{2.32}$$

F-measure represents a trade-off between precision and recall and, in general, improving one degrades the other. Therefore, the goal of a link prediction method is to maximize the f-measure (ALLALI; MAGNIEN; LATAPY, 2013).

Evaluation measures to space analysis provide a better idea about the performance of link prediction algorithms when they are applied over large-scale networks. This fact is because in large-scale networks, the size of L_p is bigger, facilitating the analysis of class distributions of predicted links.

Evaluation Measures to Analyze the Predictive Power

Evaluation measures to analyze the predictive power evaluate the scores of links in the set of predicted links L_p, aiming quantify how important are the correctly predicted links. Some evaluation measures to analyze the predictive power are:

AUC. The *area under the receiver operating characteristic curve* (AUC) (HANLEY; MCNEIL, 1982) can be interpreted as the probability that for a randomly chosen correctly predicted link was given a higher score than for a randomly chosen wrongly predicted link (LÜ; ZHOU, 2011). Therefore, for n independent comparisons among predicted links, if n_1 times for the correctly predicted links were given higher scores than for wrongly predicted links whilst n_2 times they were given equal scores, the AUC is defined as:

$$AUC = \frac{n_1 + 0.5 \times n_2}{n}. \tag{2.33}$$

If the scores are generated from an independent and identical distribution, the AUC should be about 0.5. Therefore, the extent to which AUC exceeds 0.5 indicates how better the link prediction method performs than pure chance.

Precisi@L. Similar to the traditional precision measure, but focused on a specific sample from prediction results (HERLOCKER *et al.*, 2004). Therefore, considering the L_r correctly predicted links from the L top-ranked predicted links (LÜ; ZHOU, 2011; WANG *et al.*, 2011), the precis@L is defined as:

$$precisi@L = \frac{Lr}{L}. \tag{2.34}$$

Precis@L, also called as *top-L predictive rate*, is a good metric for the link prediction task because here sensitivity and specificity are linearly dependent. However, it depends on how appropriate is the selection of the L values due to the use of arbitrary L values could provide too sensitive results (YANG; LICHTENWALTER; CHAWLA, 2015).

Evaluation measures to prediction power analysis are more frequently used in the literature. Since these evaluation measures have a low dependence of class distribution of predicted links by focusing only on the prediction scores of a specific amount of predicted links.

Supervised Link Prediction

The supervised strategy deal with the link prediction problem as a classical supervised learning task. Therefore, network information such as the nodes attributes, nodes behavior patterns, or even similarity scores, can be used to build a set of features vectors for both linked and not linked pairs of nodes. On this set of features vectors is possible apply some typical machine learning model such as classifier or probabilistic model. Therefore, the machine learning model will determines the positive (existent) or negative (nonexistent) class label of new links instances (HASAN *et al.*, 2006; LICHTENWALTER; LUSSIER; CHAWLA, 2010; BENCHETTARA; KANAWATI; ROUVEIROL, 2010; VALVERDE-REBAZA; LOPES, 2014).

Compared to unsupervised process, the supervised one can improve the prediction accuracy significatively since, in this case, it is possible to explore multiple features of node pairs at once (LICHTENWALTER; LUSSIER; CHAWLA, 2010; LIU *et al.*, 2016).

Evaluation Measures

By using the supervised link prediction strategy, it is possible to use the classical evaluation measures, such as *accuracy*, *precision*, *recall*, *F-measure*, *AUC*, and others, to compare the performance of classifiers and other learning models (HASAN *et al.*, 2006; LICHTNWALTER; CHAWLA, 2012; VALVERDE-REBAZA; LOPES, 2014).

2.2.3 Link Prediction Methods

There are different methods to link prediction, each one using different network information sources as well as procedures. Therefore, we divide all the existing techniques into two major groups: i) similarity-based methods, and ii) learning-based methods and approximation methods. In this section, we will present a review of these link prediction methods.

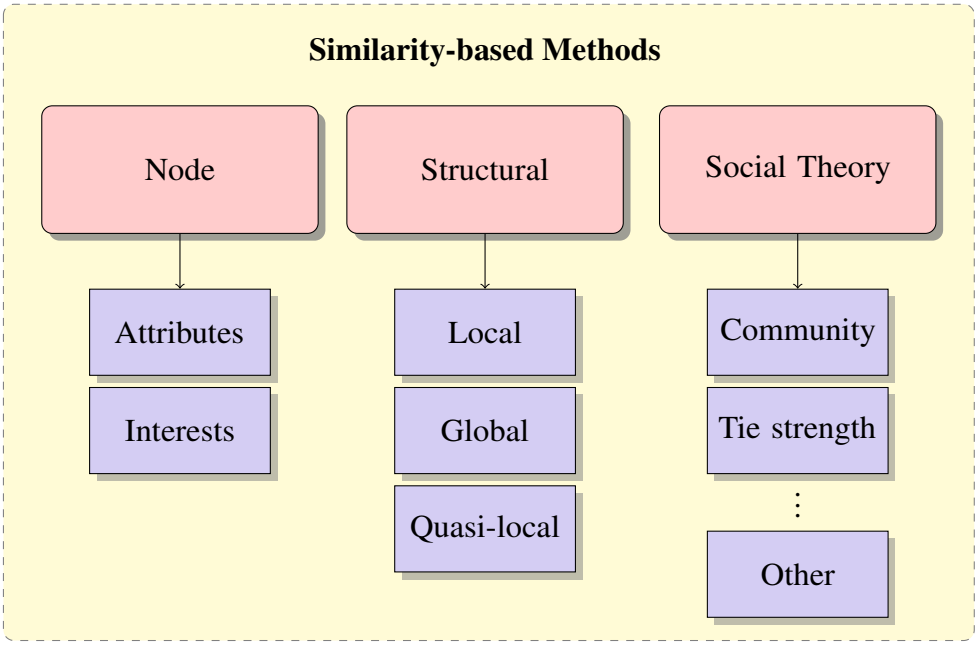
It is important to note that, in this review we do not take into account link prediction methods considering the temporal evolution of networks, e.g. link prediction methods based on time-series (HUANG; LIN, 2009; SOARES; PRUDêNCIO, 2012; GÜNEŞ; GÜNDÜZ-ÖĞÜDÜCÜ; ÇATALTEPE, 2016; MORADABADI; MEYBODI, 2017) and link prediction methods in dynamic networks (SARKAR; CHAKRABARTI; JORDAN, 2012; ZHU; STEEG; GALSTYAN, 2014; RAHMAN; HASAN, 2016).

2.2.3.1 Similarity-based Methods

The simplest framework of link prediction methods is based on similarity, where each pair of nodes x and y is assigned a score $s_{x,y}$ which is directly defined as the similarity or proximity between such pair of nodes. All non-observed links are ranked according to their scores, and the links connecting more similar nodes will probably exist (LÜ; ZHOU, 2011; WANG *et al.*, 2014).

The definition of similarity is not a trivial task, since it has a heuristic component. The same similarity function works differently for distinct networks even from the same domain. Therefore, a large number of similarity-based methods with different definitions of similarity have been proposed (MARTÍNEZ; BERZAL; CUBERO, 2016). For the sake of clarity, we group them into three categories: node similarity, structural similarity, and based on social theory. In Figure 16 we show the categorization of similarity-based link prediction methods into the three previously mentioned groups. Furthermore, for each one of the three groups, we also show the respective information sources or techniques used to perform the similarity calculation.

Figure 16 – Categorization of similarity-based methods for link prediction.



Source: Elaborated by the author.

Methods based on Node Similarity

In a real-world network, a node usually has some informations such as the user profile in OSNs, mail name in email networks, publication record in academic social networks, etc. These information can be directly used to similarity calculation between two nodes. Therefore, node similarity can be defined by using the essential information of nodes being that two nodes are considered to be similar if they have many common features (LIN, 1998; LÜ; ZHOU, 2011).

Generally, essential informations of nodes are constituted by their attributes and actions, which reflect particular interests and behaviors. Therefore, node-based methods are useful in link prediction if we obtain users's attributes and actions from real-world networks, such as social networks (WANG *et al.*, 2014).

A. Methods based on Node Similarity using User Attributes

Since in most cases the node attributes are formed by textual information, the text-based and string-based similarity metrics are usually used here (NAVARRO, 2001). For instance, Bhattacharyya, Garg and Wu (2011) defined a multiple categorization tree model to study the keywords of user profiles, then a distance is defined between keywords to determine the similarity between a pair of users. Akcora, Carminati and Ferrari (2013) proposed a method to infer the missing values of user profiles to improve the computing of similarity and the link prediction task.

Other authors use node attributes not necessarily in textual format. For instance, Gong *et al.* (2014) use users' profile attributes such as occupation, employment, education, places lived, among others, whilst Han *et al.* (2015b) explored other users' profile attributes such as workplace, high school and hometown.

B. Methods based on Node Similarity using User Interests

User interests also are briefly used by node-based similarity methods. Interests are represented by the actions that users take such as writing a comment, create and editing a blog, asking a question on a forum, among other. Therefore, Anderson *et al.* (2012) consider users's interests overlap to measure the similarity applying the cosine between the interests vectors of a pair of users.

Other authors focused their efforts on building robust feature vectors. For instance, Ma (2014) applied interest similarities using a finer granularity to avoid different factors that affect the interest similarities such as subgraph topology, connected components, number of co-friends, among others. Han *et al.* (2015a) investigated how users' interest similarity relates to various social features, such as geographical distance, and accordingly infer whether the interests of two users are alike or unalike where one of the users' interests are unknown.

Methods based on Structural Similarity

Sometimes essential information of nodes are hidden, so it is necessary to use the structure of network to capture the similarity between a pair of vertices. Methods based solely on network structure are called *structural similarity* or *topological similarity*. Since the widespread work of Liben-Nowell and Kleinberg (2007), several link prediction methods based on structural similarity have been proposed, constituting the most explored and used techniques in real-world applications.

Structural similarity methods can be organized into different groups, such as local versus global, parameter-free versus parameter-dependent, etc. (LÜ; ZHOU, 2011; MARTÍNEZ; BERZAL; CUBERO, 2016). In this thesis, we will give a systematic explanation of popular

structural-based link prediction methods and will divide them into three categories: local, global, and quasi-local methods. In following descriptions, we employ definitions previously presented in Sections 2.1.2 and 2.1.4.

A. Methods based on Local Structural Similarity

Link prediction methods based on local similarity use node neighborhood-related structural information to calculate the similarity between pairs of nodes in the network. These methods are faster than nonlocal methods and highly parallelizable, becoming an excellent option to handle efficiently the link prediction problem in real-world networks such as OSNs (MARTÍNEZ; BERZAL; CUBERO, 2016).

One important consideration of local similarity methods is the fact that using only local information restricts the calculation of similarity to nodes separated by a specific and restricted distance, for instance 2-hops (neighbors of neighbors), 3-hops (neighbors of neighbors of neighbors), etc. However, the performances of local similarity methods using 3-hops or more are worse than using 2-hops (LICHTENWALTER; LUSSIER; CHAWLA, 2010). This fact is a big drawback since many links are formed between nodes more than 2-hops away in many real-world networks (LIBEN-NOWELL; KLEINBERG, 2007).

Since local similarity methods are limited to 2-hops, their time complexities range from $O(|V|^2)$ to $O(2|V|^2)$. Furthermore, these methods have shown a very competitive prediction accuracy compared to more complex methods, characteristic which has led to many of them to be considered as state-of-the-art methods (LÜ; ZHOU, 2011; WANG *et al.*, 2014; MARTÍNEZ; BERZAL; CUBERO, 2016). Bellow, we summarize five of them.

Common Neighbors (CN). Common neighbors is the simplest local method (LORRAIN; WHITE, 1971). Its basic assumption is that two nodes are more likely to be connected if they share more common neighbors. It makes sense to assume that, if two individuals share many acquaintances, they are more likely to become friends than two individuals without common contacts. Furthermore, different studies have confirmed this hypothesis by observing a correlation between the number of shared neighbors of two nodes and their probability of being linked (NEWMAN, 2001; LIBEN-NOWELL; KLEINBERG, 2007).

This method refers to the size of the set of all common neighbors, $\Lambda_{x,y}$, of a pair of disconnected nodes x and y according to Equation 2.35.

$$s_{x,y}^{CN} = \left|\Lambda_{x,y}\right| = \left|\Gamma(x) \cap \Gamma(y)\right|. \tag{2.35}$$

Despite its simplicity, CN performs surprisingly well on many real-world networks and beats very complex approaches. Furthermore, CN is the basis for other link prediction methods.

Jaccard Coefficient (Jac). Proposed over a hundred years ago by Jaccard (1901), still is widely used in information retrieval systems, this method indicates whether two nodes of a network have a significant number of common neighbors regarding their total neighbors set size. The Jaccard coefficient is calculated according to Equation 2.36.

$$s_{x,y}^{Jac} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}. \tag{2.36}$$

The Jaccard coefficient can be easily seen as a variation of the common neighbors method where there is a penalization due nonshared neighbors.

Adamic-Adar (AA). Initially proposed by Adamic and Adar (2003) for computing similarity between two entities based on their shared features at first, currently it has been widely used in social networks. This method refines the simple counting of common neighbors by assigning more weight to the less-connected neighbors, as defined in Eq 2.37.

$$s_{x,y}^{AA} = \sum_{z \in \Lambda_{x,y}} \frac{1}{\log |\Gamma(z)|}. \tag{2.37}$$

Adamic-Adar method is a variation of the common neighbors method where each shared neighbor is logarithmically penalized by its degree. It makes sense to assume that, the amount of resources or time that a node can spend on each of its neighbors decreases as its degree increases, also decreasing its influence on them.

Resource Allocation (RA). Motivated by the resource allocation dynamics on complex networks, Zhou, Lü and Zhang (2009) proposed this method by modeling the transmission of units of resources between a pair of disconnected nodes x and y through their common neighbors, which play the role of transmitters. Each one of common neighbors gets a unit of resource from x and equally distributes it to all its neighbors. The similarity between x and y can be defined as the amount of resources received by y from x, according to Eq. 2.38.

$$s_{x,y}^{RA} = \sum_{z \in \Lambda_{x,y}} \frac{1}{|\Gamma(z)|}. \tag{2.38}$$

Resource allocation method punishes the high-degree common neighbors more heavily than Adamic-Adar method. The resource allocation method has been shown to be the local method that achieves best results when compared with other local similarity methods in a large number of networks (PAN *et al.*, 2010; VIRINCHI; MITRA, 2013; MARTÍNEZ; BERZAL; CUBERO, 2016).

Preferential Attachment (PA). This method is a direct result of the well known work of Barabási and Albert (1999) referred to the scale-free network formation model. Many real network node degrees follow a power law distribution, resulting in scale-free networks. In this network, newcomers have preference to connect to more “popular” nodes, leading to the

concept of "the rich get richer". Therefore, this similarity method is proportional to the number of neighbors of each node, as defined in Equation 2.39.

$$s_{x,y}^{PA} = |\Gamma(x)| \times |\Gamma(y)|. \tag{2.39}$$

Preferential attachment method can be also applied in nonlocal contexts, since it does not rely on common neighbors. Due to that fact, the time complexity of preferential attachment method is $O(2|V|)$. Furthermore, its prediction accuracy is usually poor even worst than pure chance (LIBEN-NOWELL; KLEINBERG, 2007; LÜ; ZHOU, 2011; MARTÍNEZ; BERZAL; CUBERO, 2016).

Other traditional local similarity methods also are well known in the literature. For instance, the *Sørensen* index (Sor) proposed by Sorensen (1948), which besides considering the size of the common neighbors also points out that lower degrees of nodes would have higher link likelihood. The *Salton* index (Sal) proposed by Salton and McGill (1983), which is similar to Jac but less sensitive to outliers. Ravasz *et al.* (2002) proposed the *Hub Promoted Index* (HPI) and *Hub Depressed Index* (HDI), which compute the link likelihood by comparing the number of common neighbors to the lower or higher degree of analyzed nodes, respectively. Leicht, Holme and Newman (2006) prosed the *Local Leicht-Holme-Newman* (LLHN) index, which is defined as the ratio of actual paths of length two between two nodes and a value proportional to the expected number of paths of length two between them.

Recently, a large number of novel local similarity methods have been proposed. Some of these methods improve the performance of state-of-the-art methods by incorporating probabilistic tools. For instance, Liu *et al.* (2011) proposed the *Local Naïve Bayes* (LNB) method, which assumes that each common neighbor has a different role or degree of influence that can be estimated using a naïve Bayes model. The authors also proposed some variants of LNB, which are based on CN (LNB-CN), AA (LNB-AA), and RA (LNB-RA) methods. Given $\Omega = \frac{|E|}{|U|-|E|}$, the definitions of methods based on Naïve Bayes model proposed by Liu *et al.* (2011) are as follows:

Local Naïve Bayes (LNB). The local naïve Bayes considers the contribution of each element of set of all common neighbors to compute the likelihood of existence of a pair of disconnected nodes x and y. Given $N_z = \frac{\Delta_z}{\Lambda_z}$, where Δ_z and Λ_z are respectively the number of connected and disconnected pairs of nodes whose common neighbors include $z \in \Lambda_{x,y}$, the local naïve Bayes method is computed as stated in Equation 2.40.

$$s_{x,y}^{LNB} = \prod_{z \in \Lambda_{x,y}} \Omega^{-1} N_z. \tag{2.40}$$

Local Naïve Bayes of Common Neighbors (LNB-CN). The local naïve Bayes of Common Neighbors instances the traditional CN method on the local naïve Bayes model. The local

naïve Bayes of Common Neighbors is defined as stated in Equation 2.41.

$$s_{x,y}^{LNB-CN} = |\Lambda_{x,y}| \log(\Omega^{-1}) + \sum_{z \in \Lambda_{x,y}} \log(N_z). \tag{2.41}$$

Local Naïve Bayes of Adamic-Adar (LNB-AA). The local naïve Bayes of Adamic-Adar instances the traditional AA method on the local naïve Bayes model. The local naïve Bayes of Adamic-Adar is defined as stated in Equation 2.42.

$$s_{x,y}^{LNB-AA} = \sum_{z \in \Lambda_{x,y}} \frac{1}{\log(k_z)} (\log(N_z) + \log(\Omega^{-1})). \tag{2.42}$$

Local Naïve Bayes of Resource Allocation (LNB-RA). The local naïve Bayes of Resource Allocation instances the traditional RA method on the local naïve Bayes model. The local naïve Bayes of Resource Allocation is defined as stated in Equation 2.43.

$$s_{x,y}^{LNB-RA} = \sum_{z \in \Lambda_{x,y}} \frac{1}{k_z} (\log(N_z) + \log(\Omega^{-1})). \tag{2.43}$$

Other local similarity methods have incorporated other strategies to improve the link prediction accuracy. Chua, Sung and Wong (2006) proposed the *Functional Similarity Weight* (FSW), which is derived from Sor but apply a penalization when any of the analyzed nodes has a small degree. Liu, Li and Wong (2008) proposed *Local Interacting Score* (LIT), which is an iterative variation of FSW by optimizing the penalization function used. Dong *et al.* (2011) proposed the *Individual Attraction Index* (IAI), which is based on RA but taking into account how connected the common neighbors are. Zhu *et al.* (2012) proposed the *Parameter Dependent* (PD) method, which is based on Sal and LLHN methods, and by using a free parameter is capable of predict both popular and unpopular links. Tan, Xia and Zhu (2014) proposed the *Mutual Information* (MI) method, which tackles the problem from the perspective of information theory by computing the conditional self-information of the existence of a link given the set of common neighbors. Zhang *et al.* (2014) proposed the *Resource Allocation Based on Common Neighbor Interactions* (RA-CNI) method, which is based on RA but considers the return of resources in the opposite direction into the allocation process. Zhu and Xia (2015) introduced the *Neighbor Set Information* (NSI) method, which from a perspective of information theory, uses both the common neighbors and the links across the two neighbor sets of analyzed nodes to improve the prediction accuracy. Liu *et al.* (2017) proposed the *Extended Resource Allocation* (ERA) method, which is based on RA but considers a parameter to adjust the amount of resources transferred, i.e. the allocation process is performed by sending more than unit of resources between a pair of disconnected nodes.

As observed, many local similarity methods are based on the use of the set of common neighbors. It is due to the fact that common neighbors can indirectly reflect node's behavior and directly affect connection choice (AKCORA; CARMINATI; FERRARI, 2013; WANG

et al., 2014). Moreover, most of the local similarity methods outperform more complicated link prediction methods in practical applications. This fact has many empirical evidences and theoretical justifications (SARKAR; CHAKRABARTI; MOORE, 2011). Therefore, one should choose the link prediction method according to the characteristics of network to be analyzed, due to many experiment evaluation results have shown that there is no an absolutely dominating link prediction method for different datasets (LIBEN-NOWELL; KLEINBERG, 2007; WANG *et al.*, 2014; YANG; LICHTENWALTER; CHAWLA, 2015).

B. Methods based on Global Structural Similarity

Global similarity-based methods use the whole network topological information to score each disconnected pair of nodes. These methods are not limited to measure similarity between 2-hops and provide much more accurate prediction than the local similarity methods. However, global similarity methods suffer two big disadvantages. The first, is the fact that their time complexity is around $O(|V|^3)$, which is infeasible for large-scale networks. The second disadvantage is related to the fact that, sometimes, the global topological information is not available, making the parallelization process of these methods very complex (LÜ; ZHOU, 2011; MARTÍNEZ; BERZAL; CUBERO, 2016).

Different global similarity methods based on network paths have been proposed. Liben-Nowell and Kleinberg (2007) proposed the *Negated Shortest Path* (NSP), a basic global similarity measure defined as the negative value of shortest path distance between a pair of disconnected nodes. NSP accuracy is poor even when compared to local similarity methods. Other global similarity methods are based on multiple paths, obtaining significantly better results. Katz (1953) proposed the *Katz index* (KI), which sums the influence of all possible paths between a pair of disconnected nodes, incrementally penalizing paths by their length. Leicht, Holme and Newman (2006) proposed the *Global Leicht-Holme-Newman* (GLHN) index, which is the global version of LLHN and based on the same fundamentals of KI. The GLHN assigns a similarity proportional to the number of paths between a pair of disconnected nodes.

Other global similarity methods are based on the use of random walks, a random process introduced by Pearson (1905) and which have been applied to describe stochastic processes in many fields such as economics, physics, biology, among others. Given a graph and a starting node, a random walk process consists in randomly select a neighbor of this node and move to it, after that, repeat this process for each reached node. This Markov chain of randomly selected nodes is known as a random walk on the network. Therefore, Liu and Lü (2010) proposed the basic *Random Walks* (RW) similarity method, which considers that the connection probability between a pair of disconnected nodes x and y is defined as the probability of a random walk reaching y starting from x. Jeh and Widom (2002) proposed *SimRank* (SR), a method that computes how soon two random walkers starting from disconnected nodes x and y are expected to meet at the same vertex. Tong, Faloutsos and Pan (2006) proposed the *Random Walks with Restart*

(RWR), which modifies the traditional random walk process by move to a selected node with probability α or return to the starting node with probability $(1-\alpha)$. A slight variation of RWR is the *Flow Propagation* (FP) method, proposed by Vanunu and Sharan (2008), which apply RWR but replacing the normalized adjacency matrix with the normalized Laplacian matrix. Burda *et al.* (2009) introduced the *Maximal Entropy Random Walk* (MERW), which improve the prediction accuracy by maximizing the entropy rate of the random walk process. Liu and Lü (2010) also proposed the *Average Commute Time* (ACT) method, which is defined as the average number of steps that a random walker starting from node x takes to reach a node y for the first time and go back to x.

Authors also have exploited alternative graph representations to compute the similarity between pairs of disconnected nodes. Blondel *et al.* (2004) proposed the *Blondel Index* (BI), which is capable of measuring the similarity between a pair of nodes belonging to different graphs and also to the same graph. The BI computes the similarity iteratively over a Frobenius matrix. Chebotarev and Shamis (2006) proposed the *Random Forest Kernel* (RFK) index, which using a spanning three of network, computes the similarity between a pair of disconnected nodes x and y by the ratio of the number of spanning rooted forests such that nodes x and y belong to the same tree rooted at x to all spanning rooted forests of the network. Fouss *et al.* (2007) proposed the *Pseudoinverse of the Laplacian Matrix* (PLM), also called *Cosine based on* L^+ method, which computes the cosine similarity between a pair of disconnected nodes over the Moore-Penrose pseudoinverse of the Laplacian matrix.

C. Methods based on Quasi-local Structural Similarity

Quasi-local methods have recently emerged as a promising tradeoff to the balance between local and global measures. Quasi-local methods consider more topological information than local methods while abandon superfluous information that makes no contribution or very little contribution to the prediction accuracy, as global methods do. Neither they take into account the similarity between any arbitrary pair of nodes nor they are limited to neighbors of neighbors (LÜ; ZHOU, 2011; MARTÍNEZ; BERZAL; CUBERO, 2016).

Some quasi-local methods have access to the whole network, but their algorithmic time complexity is still below the time complexity of global methods. For instance, Lü, Jin and Zhou (2009) proposed the *Local Path Index* (LPI), which is strongly based on the Katz index but it only considers a finite number of path lengths. Liu and Lü (2010) introduced the *Local Random Walks* (LRW), which exploits the concept of random walks but limiting this process to a fixed number of iterations. Liu and Lü (2010) also introduced the *Superposed Random Walks* (SRW) method, which computes the LRW but superposing each walker contribution by continuously releasing the walker at the starting node. Lichtenwalter, Lussier and Chawla (2010) introduced the *PropFlow* (PF), which computes the probability that a restricted random walk starting from x ends at y in l steps or fewer. Papadimitriou, Symeonidis and Manolopoulos (2012) proposed

FriendLink (FL) which is based on the path count between nodes of interest, like LPI, but using different normalization and penalization mechanisms. Zhang *et al.* (2014) proposed the *Third-Order Resource Allocation Based on Common Neighbor Interactions* (ORA-CNI), which is an extension of RA-CNI but taking into consideration distance between paths.

Similarity Methods based on Social Theory

In recent years, a large number of techniques have been inspired in social theories, such as strong and weak ties, homophily, triadic closure, structural holes, community, and others, to face the link prediction problem (WANG *et al.*, 2014). Here it is important to note that, depending on the scope of network structure used by social theory methods, they also can be considered as local, global or quasi-local.

Social ties are defined as information-carrying connections between people, and can be classified as strong, weak or absent (GRANOVETTER, 2005; WUCHTY, 2009). Different authors have used social ties to improve the link prediction accuracy. Lü and Zhou (2010) used state-of-the-art local similarity methods over weighted and unweighted networks and found that the weak ties play a significant role in the link prediction task by contributing directly to enhance the prediction accuracy for some networks, i.e. the weak links in some networks are not as weak as their weights can suggest. Liu *et al.* (2013) consider that each common neighbor play a different role to the node connection likelihood according to their centralities. Therefore, the authors proposed the *Degree-Centrality-based Common Neighbors* (DC-CN), *Closeness-Centrality-based Common Neighbors* (CC-CN) and *Betweenness-Centrality-based Common Neighbors* (BC-CN) methods. These methods compute the similarity of a pair of disconnected nodes based on weak ties and by quantifying the degree, closeness and betweenness centralities, respectively. On the other hand, Socievole, Rango and Marano (2013) construct a contact graph based on the stronger ties of users of a social network, then use the relationships existent in this graph together to those originally existent into the original social network to improve the predictive power of state-of-the-art local similarity methods. Backstrom and Kleinberg (2014) proposed *dispersion* (Dsp), a measure to estimate tie strength. Using Dsp, the authors are capable of predicting specific relationships, specially those characterized by representing strong social ties as spouses or romantic partners. Xu *et al.* (2017) introduced the *Weighted Path Entropy* (WPE) method, which differently of previously mentioned, considers that a weak tie is not a small weight link but a path with small weight. Therefore, WPE quantifies the contribution of a path with both path entropy and path weight improving the prediction accuracy.

Homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people, and such rate can be measured by cultural, behavioral, genetic, or material information flowing through networks (MCPHERSON; SMITH-LOVIN; COOK, 2001; CHANG *et al.*, 2014; FARALLI; STILO; VELARDI, 2015). Aiello *et al.* (2012) observed a strong correlation between the social connectivity and the intensity of explicit user activities like

tagging and participation in groups. Therefore, the authors built a user similarity network for each homophily factor identified, after that is possible to apply different link prediction methods over such network and obtain accurate prediction results. Yuan *et al.* (2014) exploited sentiment homophily for link prediction. The authors consider sentiments of social network users toward topics of mutual interest as homophily factors to be used to compute traditional link prediction methods. Ciotti *et al.* (2016) identified homophily factors related to academic papers, authors, research sub-fields, and scientific journals, in a citation network, to use them to perform the link prediction task aiming to discover missing citations between pairs of highly related articles.

Triadic closure is a property among three nodes x, y and z such that if the node x links to y, and y links to z, then one should arguably expect an increased likelihood that x will link to z (GRANOVETTER, 2005; KOSSINETS; WATTS, 2006). On the other hand, *structural holes* are nodes that act as spanners among communities or groups of nodes without direct connections (BURT, 1992; GRANOVETTER, 2005). We can observe that, while triadic closure try to predict new links under the argument that a network is formed by strongly interconnected nodes, the structural holes argument says that the link formation is due to the fact that some nodes are important to the connectivity of local regions (BURT, 2001). Despite this apparent contradiction, both the triadic closure (ROMERO; KLEINBERG, 2010; SINTOS; TSAPARAS, 2014) and structural holes (AHUJA, 2000; KLEINBERG *et al.*, 2008) have been widely studied for being considered as important factors to the formation of new relationships among nodes in networks.

Community is one of the most relevant features of networks representing real systems by organizing nodes in groups, modules or clusters, with many links joining nodes of the same community and comparatively few links joining nodes of different communities (CLAUSET; NEWMAN; MOORE, 2004; FORTUNATO, 2010). Studies have found that accuracy of link prediction methods based on local similarity drastically improves when community structure of networks grows (FENG; ZHAO; XU, 2012; LIU *et al.*, 2013). Therefore, Zheleva *et al.* (2008) proposed new local similarity methods based on the participation of users in different social groups. Soundarajan and Hopcroft (2012) proposed adaptations of state-of-the-art local similarity methods which awarding extra points to pairs of nodes that share many communities and penalizing pairs of nodes that do not share communities. Differently of previous works, Valverde-Rebaza and Lopes (2012a) consider that by playing different roles in a network, the common neighbors also may contribute differently if they are within or out the same communities of the pair of disconnected nodes. Taking into account this consideration and using a Bayesian probabilistic framework, the authors proposed the *Within and Inter Community* (WIC) measure, which by better exploiting the community structure of networks performs better than state-of-the-art methods. Additionally, by considering only the common neighbors in the same communities of analyzed nodes, the authors also proposed a variety of adaptations of traditional local similarity methods, called *W form* methods, which perform better than their respective counterpart method.

Because the community structure encloses itself different structural and social properties as social ties, triadic closure, structural holes, and others, much of researchers focused on the use of social theory on link prediction have put their attention on better use the community information to improve the link prediction accuracy. Therefore, based on the primary works of Zheleva *et al.* (2008), Soundarajan and Hopcroft (2012), and Valverde-Rebaza and Lopes (2012a), a considerable amount of works using community information to enhance the link prediction have been proposed (HOSEINI; HASHEMI; HAMZEH, 2012; CANNISTRACI; ALANIS-LOBATO; RAVASI, 2013; KEMAL; TSUYOSHI *et al.*, 2014; MALLEK *et al.*, 2015; DAMINELLI *et al.*, 2015; WU *et al.*, 2016; DING *et al.*, 2016; MA *et al.*, 2016; KUANG; LIU; YU, 2016; CAIYAN; CHEN; LI, 2016; BISWAS; BISWAS, 2017). Considering the fact that social grouping is a natural behavior of users in OSNs and that social groups can be associated directly with community structure (YANG; LESKOVEC, 2015), we will use the concepts of social group and community interchangeably throughout this thesis.

2.2.3.2 Learning-based Methods and Approximation Methods

Based on features provided by similarity methods previously described as well as by node attributes and network structure in general, a variety of link prediction methods based on learning and approximation have been proposed in recent years (LÜ; ZHOU, 2011; WANG *et al.*, 2014; MARTÍNEZ; BERZAL; CUBERO, 2016). *Learning-based methods* include a variety of probabilistic and statistical models used to calculate the probability of existence of unobserved links, such as hierarchical structure model, stochastic block model, cycle formation model, local co-occurrence model, and others. Moreover, among the learning-based methods we consider the large amount of classification models used to label the links as existent or not.

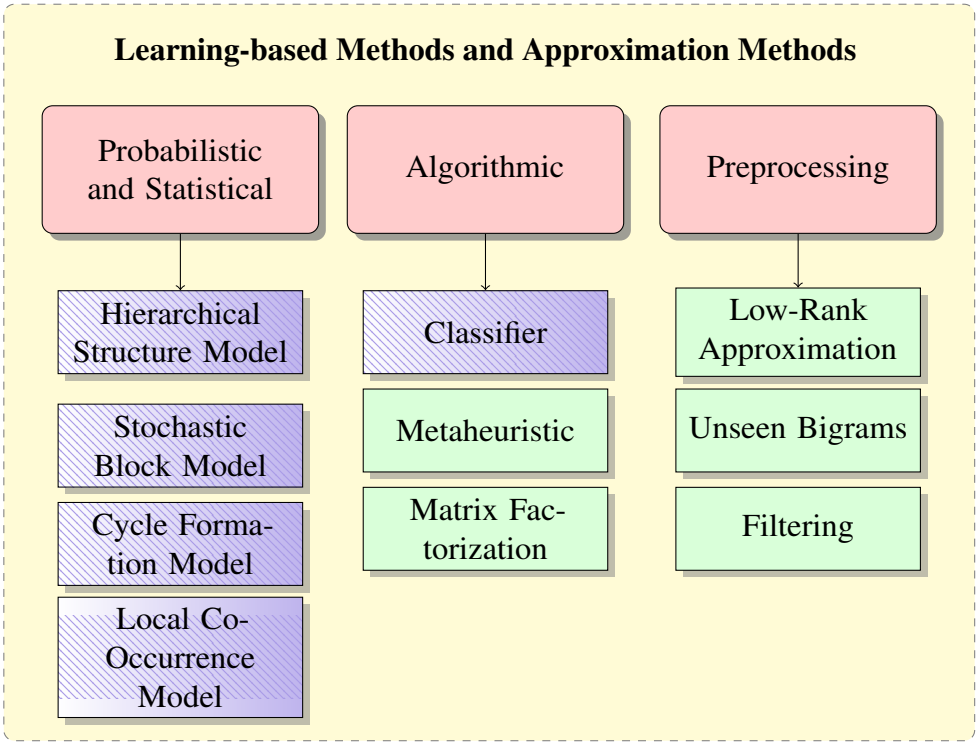
On the other hand, the *approximation methods* are considered as "meta-approaches" since they can perform the link prediction process by searching an optimal solution, as the metaheuristic methods do, or by properly modeling the network, as the matrix factorization-based methods do. Among the approximation methods we also consider those called as preprocessing methods, which cleaning-up the existent noise in the network to after apply any well-known link prediction method, such as low-rank approximation, unseen bigrams, and filtering.

Based on the work of Martínez, Berzal and Cubero (2016), in Figure 17 we show the categorization of learning-based methods and approximation methods for link prediction. This categorization basically group the existing literature into probabilistic and statistical, algorithmic, and preprocessing methods. Furthermore, for each category we show the more representative approaches.

Probabilistic and Statistical Methods

Different techniques based on statistical analysis and probability theory usually assume that the network has a known structure from which is possible to build a model fitting such

Figure 17 – Categorization of learning-based methods and approximation methods for link prediction.



Source: Elaborated by the author.

Note – Probabilistic and statistical methods as well as algorithmic methods grouped in lilac color boxes (filled with lines) are considered as learning-based methods.

Note – Algorithmic and preprocessing methods grouped in green color boxes (without filling lines) are considered as approximation methods.

structure and estimate parameters capable of modeling the network. These parameters are used to calculate the probability of formation of unobserved links. The probabilities computed can be ranked to choose the most potential links to exist as the same way that similarity-based methods work.

A variety of probabilistic and statistical methods for link prediction have been proposed. Based on the work of Martínez, Berzal and Cubero (2016), we divide them as: i) hierarchical structure model, ii) stochastic block model, iii) cycle formation model, and iv) local co-occurrence model.

A. Methods based on Hierarchical Structure Model

Empirical evidence indicates that many real networks are hierarchically organized, i.e. nodes can be divided into groups, further subdivided into groups of groups, and so forth over multiple scales (SALES-PARDO *et al.*, 2007). In hierarchical networks, nodes with higher degree are expected to have a lower clustering coefficient than lower degree nodes. Therefore,

the hierarchical structure is formed due to the fact that hub nodes weakly connect isolated communities of highly clustered nodes (RAVASZ *et al.*, 2002).

The primary work of Clauset, Moore and Newman (2008) proposes that the hierarchy of a network can be represented by a dendrogram with $|V|$ leaves and $|V|-1$ internal nodes. Each leaf represents a node from the network and each internal node represents a relationship among its descendant nodes in the dendrogram. Each internal node has an associated probability, which is equal to the probability of a link between nodes of both branches descending from it. Since a same network may has multiple representations based on dendrograms depending on how internal nodes are set, the link prediction task is performed by: i) sampling a large number of dendrograms with probability proportional to their likelihood; ii) for each pair of disconnected nodes x and y, calculate the average connecting probability $\langle p_{x,y} \rangle$ by averaging the corresponding probability over all sampled dendrograms; and, iii) sorting these node pairs in descending order of $\langle p_{x,y} \rangle$ and the highest-ranked ones are those to be considered as predicted.

An important remark to be considered is that the proposal by Clauset, Moore and Newman (2008) may give poor predictions for networks without clear hierarchical structures. To avoid this problem, Yang *et al.* (2015) proposed to exploit the hierarchical nature of brain networks. Therefore, the authors take a network from any domain and build a new one using the brain network model. Over the network data modeled as a brain network, a large number of dendrograms are build and, after that, the link prediction process follows similarly to the one proposed by Clauset, Moore and Newman (2008). By applying the link prediction process over a hierarchical structure build over data modeled as a brain network, Yang *et al.* (2015) have obtained better AUC performance than the proposal by Clauset, Moore and Newman (2008), which apply the link prediction process over a hierarchical structure build directly from the network data.

The hierarchical structure model provides a smart way to perform the link prediction task, and, maybe more significantly, it uncovers the hidden hierarchical organization of networks. However, a big disadvantage is that the link prediction process is very slow. The process to sample dendrograms usually takes between $O(|V|^2)$ and exponential time. This fact makes the link prediction methods based on hierarchical structure model being infeasible for practical applications.

B. Methods based on Stochastic Block Model

Stochastic block model is one of the most general network models, where nodes are partitioned into groups and the probability that two nodes are connected depends solely on the groups to which they belong to. The stochastic block model can capture the community structure, role-to-role connections, and maybe other factors for the establishing of connections (WHITE; BOORMAN; BREIGER, 1976; DOREIAN; BATAGELJ; FERLIGOJ, 2005).

Based on the stochastic block model approach, Guimerà and Sales-Pardo (2009) established that the probability of link formation between two nodes directly depends on the block they belong to. Therefore, after calculate all the blocks to which all the nodes belong to, and based on the likelihood of the network given the partition to which the blocks of a pair of disconnected nodes belong to, the authors applied the Bayes theorem to calculate the maximum likelihood of existence of a link between such pair of nodes. Whang, Rai and Dhillon (2013) applied the same link prediction process but considering overlapping stochastic block models.

It is important to consider that the number of blocks to be computed grows fast as the number of nodes in the network increases, leading to the computational cost to be exponential. Despite some strategies can be applied, such as the use of *Metropolis algorithm* to sample partitions (METROPOLIS *et al.*, 1953), the whole process is still very time consuming and infeasible for large-scale networks. Recently, Liu *et al.* (2013) proposed a fast blocking probabilistic model based on a greedy strategy, which can reduce the computation complexity and improve the link prediction accuracy.

C. Methods based on the Cycle Formation Model

The cycle formation model is based on the assumption that networks have the tendency to close cycles in their formation process. Based on that, Huang (2006) proposed a link prediction method which computes the likelihood of connection of a pair of disconnected nodes by counting the number of cycles of length k that would be formed if the evaluated link existed.

Since relations between users on OSNs often reflect a mixture of positive (friendly) and negative (antagonistic) interactions, some authors believe that the cycle formation model provides insight into some of the fundamental principles that drive the formation of such signed links. In this direction, Leskovec, Huttenlocher and Kleinberg (2010) showed that higher order cycles in a signed network generate a measure of imbalance which, as suggested by the general *theory of social balance*, can help to improve the prediction of positive relationships more than negative ones. Chiang *et al.* (2011) reinforce this discovery by showing that higher order cycles have relatively better information for nodes with a low amount of common neighbors, fact that benefits the accuracy of sign prediction and lower the false positive rate.

Despite the good performance of this methods, they are somehow limited to the sparsity problem of the network. Furthermore, methods based on the cycle formation model need to compute one or more times cycles of length $k \geq 3$, which can be very time consuming and infeasible for large-scale networks.

D. Methods based on the Local Co-Occurrence Model

The previously presented methods are prohibitive for large-scale networks due to their high computational complexity. To avoid this fact, Wang, Satuluri and Parthasarathy (2007)

proposed the local co-occurrence model, which is a scalable probabilistic method based on the use of local topological features of the network to perform the link prediction task. To compute the likelihood of the existence of a link between a pair of disconnected nodes, Wang, Satuluri and Parthasarathy (2007) use two elements: the central neighborhood set and the collection of nonderivable itemsets.

The set of relevant nodes, also called the central neighborhood set, can be obtained by compute all simple paths (without cycles) of length $1,\ldots,k$. The t nodes in the most frequent paths are selected as part of the central neighborhood set. On the other hand, the collection of nonderivable itemsets is efficiently computed for each one of these pairs by a depth-first search. Nonderivable itemsets are those itemsets whose occurrence statistics cannot be inferred from other itemset patterns, providing nonredundant constraints that can be used to learn probabilistic models without losing information.

Finally, a Markov random field (MRF) undirected graph model is iteratively built to learn the central neighborhood set and satisfying constraints associated to the collection of nonderivable itemsets. The built model allows one to compute the probability of existence of a link between the analyzed pair of nodes.

Algorithmic-based Methods

The previously presented methods are based on computing a score for each unobserved link by defining a similarity or a probability function. However, link prediction can also benefit from other algorithmic approaches, including classification and optimization techniques. These approaches have been less explored in the literature of link prediction but present interesting properties.

A. Classifier-Based Methods

The link prediction problem can be approached as a classical classification problem with two classes: existence and non-existence of a link. This is a very powerful technique since it can use any topological property and measure, or even any other link prediction method as a feature. Therefore, a set of features vectors is built for both the existent and non-existent links. Over the set of features vectors, several classifier-based approaches can be applied.

Some authors have compared different traditional classifiers including decision trees, support vector machines, k-nearest neighbors, multi-layer perceptrons, naïve Bayes, and others, as well as different ensembles of these classifiers (HASAN *et al.*, 2006; BENCHETTARA; KANAWATI; ROUVEIROL, 2010; LICHTENWALTER; LUSSIER; CHAWLA, 2010; FIRE *et al.*, 2011; FIRE *et al.*, 2014; GIMENES *et al.*, 2014). Other authors have obtained good results using random forest classifiers (CUKIERSKI; HAMNER; YANG, 2011), or using classifiers

more sophisticated, such as polynomial kernel support vector machines (LI; NIU; TIAN, 2014) and feedforward neural networks (NANDURI; RANGWALA, 2015).

The main challenge that link prediction methods based on classifiers has to deal with is the well-known class imbalance problem. Since almost all real-world networks are sparse, the number of non-existent links is extremely higher than the number of existent links. This extremely skewed distribution of classes impairs on the link prediction performance (YANG; LICHTENWALTER; CHAWLA, 2015).

B. Metaheuristic-Based Methods

Most of the approaches facing the link prediction problem are heuristic, in the sense that they try to outperform a random baseline predictor by making some assumptions about the link formation in an analyzed network. However, other approaches try to explore metaheuristic, which is a higher-level procedure or heuristic designed to find, generate, or select a heuristic that may provide a sufficiently good solution to an optimization problem. Metaheuristics do not guarantee that a globally optimal solution can be found, but they may make few assumptions about the optimization problem and solve it (BLUM; ROLI, 2003).

Recently, Bliss *et al.* (2014) proposed a method based on an evolutionary algorithm. The authors assume that different link formation heuristics can coexist and cooperate in the same network. Therefore, the proposal of Bliss *et al.* (2014) uses an evolution strategy to optimize the influence of different base link predictors including local and global similarity-based methods and node similarity features in a Twitter reciprocal reply network. Chamani, Pourebrahimi and Shirazi (2014) propose the use of a naïve Bayes classifier to generate an initial set of solutions, over which various metaheuristic algorithms such as particle swarm optimization, genetic algorithm and imperialist competitive algorithm, are used to search a final set with the best solutions with respect to the initial ones. Also, Sherkat, Rahgozar and Asadpour (2015) proposed a method based on the ant colony approach. The authors assume that a single node is an ant and a community is an ant colony. As the number of ants in a colony increases and relations between a specific source ant and this colony increase, the probability of creation of a relation between the source ant and a specific target ant in that colony would also increase. Based on this assumption, the authors optimize the process of finding triangles by taking a time complexity of $O(|V|)$. After that, over the subgraph formed by the found triangle nodes and all their neighbors, the link prediction process is performed via a subgraph matching process.

C. Matrix Factorization-Based Methods

Matrix factorization models have been widely used in recommender systems since they can extract latent features or use additional features to perform prediction. Due to the fact that recommender systems are closely related to the link prediction problem, this type of models

can be used to learns latent features from the topological structure of a network and make the prediction of possible existence of a link between a pair of disconnected nodes (KOREN; BELL; VOLINSKY, 2009; LAK; CAGLAYAN; BENER, 2014).

One of the first works using matrix factorization to face the link prediction problem was proposed by Menon and Elkan (2011). These authors introduced a model to learn both the latent features from the topological structure of an analyzed network and explicit features for nodes or edges. The model proposed by Menon and Elkan (2011) is optimized with stochastic gradient descent and address the class imbalance problem by directly optimizing for a ranking loss. To improve the link prediction accuracy, Wu and Chen (2016) explored the use of bagging technique as combination approaches for matrix factorization. Yokoi, Kajino and Kashima (2016) proposed apply the matrix factorization process over the incidence matrix of the network instead of adjacency matrix. According to the authors, the incidence matrix factorization models a partially-observed graph more accurately than traditional matrix factorization, achieving better link prediction accuracy.

Preprocessing-based Methods

Link prediction methods based on preprocessing are characterized by reducing the noise present in an analyzed network previously to the specific prediction process. The noise reduction process consists of identifying the called "weak" or "false" links, and remove them from the network. Over the reduced network, the link prediction process is performed following some previously described method.

A. Methods based on Low-Rank Approximation

This method simplifies the structure of the network to reduce its noise using the adjacency matrix representation of the graph by solving the low-rank approximation problem. This optimization problem tries to minimize a cost function that measures the fit between the original matrix and an approximation matrix of reduced rank. This problem can be algorithmically solved in an efficient way using the *singular value decomposition* (SVD) method (SARWAR *et al.*, 2000).

One of the first works based on low-rank to perform the link prediction task was proposed by Kunegis and Lommatzsch (2009). These authors generalized accurate link prediction methods to graph kernels and dimensionality reduction methods to provide a way to estimate their parameters efficiently. After that, these parameters are learned by reducing the problem to a one-dimensional least-squares regression problem whose runtime only depends on the chosen reduced rank, and is independent of the original graph size. Finally, by applying the exponential graph kernel, the authors arrived at the matrix hyperbolic sine, which provides the rating prediction results.

A single low-rank approximation may not be sufficient to represent the behavior of the entire network. Therefore, authors have been used different strategies to improve the link prediction accuracy. Shin, Si and Dhillon (2012) proposed a multi-scale link prediction method, which is based on the technique of cluster low rank approximation for massive graphs, can get global information of network structure, handles large-scale networks quickly, and obtains accurate predictions. Dong, Li and Xie (2014) introduced a probabilistic latent variable model for link prediction, which combines both the concepts of block structure and low rank approximations for matrices. First, the authors use any modularity clustering algorithm to generate blocks. Then, a low rank matrices approximations algorithm named convex nonnegative matrix factorization is used to get the link prediction results within the blocks. A very similar work have been proposed by Yang, Dong and Xie (2014). Pech *et al.* (2017) introduced the robust principal component analysis method into link prediction and designed a novel global information based prediction algorithm based upon low rank and sparsity property of the adjacency matrix. Based on that, the authors construct a network that is close to the original network and accordingly identify missing links by discovering the matrix with minimum nuclear norm.

B. Methods based on Unseen Bigrams

A bigram is a sequence of two adjacent elements in a string composed of tokens or words. The frequency distribution of bigrams has been extensively studied in many applications such as linguistics, speech recognition, or cryptography. Therefore, unseen bigrams are valid bigrams not observed in a given string set (ESSEN; STEINBISS, 1992; LEE, 1999). It has been observed that the same tokens in different bigrams with similar appearance distributions are likely to be interchangeable and to form unseen bigrams, i.e. if we observe the bigrams “a method”, “the method”, “a light”, “the light”, and “a car”, then we can infer that “the car” is an unseen bigram. The idea of “substitution” presented by unseen bigrams can be adapted to link prediction in order to reduce noise by replacing a node by its most similar nodes (MARTÍNEZ; BERZAL; CUBERO, 2016).

Based on such assumption, Liben-Nowell and Kleinberg (2007) proposed a link prediction method based on unseen brigrams approach, which by using any link prediction method, computes the score of a pair of disconnected nodes x and y using the score calculated for the pair y and z, being that node z is one of the t most “similar” nodes to x.

C. Methods based on Filtering

Originally called clustering by Liben-Nowell and Kleinberg (2007), but in order to avoid any ambiguity with link prediction methods based on clustering, Martínez, Berzal and Cubero (2016) recalled it as filtering. Link prediction methods based on filtering consider the remotion of “weak’ or “tenuous” links aiming to obtain a cleaned-up subgraph over which an accurate link prediction process can be performed.

Liben-Nowell and Kleinberg (2007) consider as "weak' or "tenuous" links those observed between nodes with a small number or no shared neighbors. Therefore, using any link prediction method are calculated the scores for all the links in the analyzed network. Then, the authors delete the γ fraction of these links for which the previously calculated score is lowest. Finally, the scores are recomputed for all pairs of disconnected nodes on this subgraph.

2.2.4 Applications

The link prediction problem has been the focus of attention of different research communities, mainly due to its broad applicability (SRINIVAS; MITRA, 2016). For biological networks, for instance, experimental discoveries of genetic or protein-protein interactions is costly; therefore, a highly accurate prediction model might reduce research costs and speed the pace of discoveries (REDNER, 2008; LÜ; ZHOU, 2011).

Link prediction has also been used to analyze different real-world networks. Link prediction can be useful to predict words, topics, or documents in information networks formed by very large collections of documents (ARNOLD; COHEN, 2009; ITAKURA *et al.*, 2011; LI *et al.*, 2016); to identify anomalous emails in email networks (HUANG; ZENG, 2006), anomalous relationships in friendship social networks (PEROZZI *et al.*, 2016), or underground relationships between criminals in terrorist social networks (HASAN *et al.*, 2006; CLAUSET; MOORE; NEWMAN, 2008; ANIL *et al.*, 2015); and to predict co-participation of individuals in organizational events (O'MADADHAIN; HUTCHINS; SMYTH, 2005).

However, the most well-known and broadly used applications of link prediction in the domain of OSNs are the recommendation systems, to personalize recommendations of items (ZHOU *et al.*, 2007; LI; CHEN, 2009; ZENG *et al.*, 2010; CHILUKA; ANDRADE; POUWELSE, 2011; HONG; YANSHEN; XIAOMEI, 2014; LI *et al.*, 2014), academic collaboration (BENCHETTARA; KANAWATI; ROUVEIROL, 2010; PEREZ-CERVANTES *et al.*, 2013), e-commerce (BAHABADI; GOLPAYEGANI; ESMAEILI, 2014), and friendship (BARBIERI; BONCHI; MANCO, 2014; AHMED; ELKORANY, 2015; TSUGAWA; KITO, 2017).

Due to its theoretical significance, several applications of link prediction are used to explain different processes in complex networks, including the natural network evolution process (KOSSINETS; WATTS, 2006; BRINGMANN *et al.*, 2010; CUI *et al.*, 2011; ANIL; SETT; SINGH, 2014), the network construction process from flat data (BERTON; VALVERDE-REBAZA; LOPES, 2015), and the reconstruction network process from network data with missing and spurious links (GUIMERÀ; SALES-PARDO, 2009). Finally, link prediction has also been applied as a support tool to improve the performance of algorithms in other link mining tasks, such as link-based object classification (GALLAGHER *et al.*, 2008; ZHANG; SHANG; LÜ, 2010; PÉREZ-SOLÀ; HERRERA-JOANCOMARTÍ, 2013), object identification (DRURY; VALVERDE-REBAZA; LOPES, 2015), and object clustering (VALEJO *et al.*, 2014b; CHENG; ZHANG, 2016).

2.3 Mining Location-based Social Networks

The increasing availability of positioning technology, e.g. GPS and Wi-fi, is changing the way people interact with the web. Therefore, people have been encouraged to share their location information using different services. Different online social networks (OSNs) have implemented different services based on location to attract more attention of users. For example, adding location-tags to photos published in OSNs focusing on photo-sharing, such as Flickr, Instagram, and Pinterest; comment on an event at the exact place where the event is happening in OSNs focusing on sharing messages or events, such as Twitter and Foursquare; adding the exact location where impressions, feelings, or multimedia contents (e.g., pictures, audio, video, etc.) are being sharing, such as Facebook and Google+ (ZHENG; ZHOU, 2011; ROICK; HEUSER, 2013; SYMEONIDIS; NTEMPOS; MANOLOPOULOS, 2014; CHORLEY; WHITAKER; ALLEN, 2015).

Location information brings social networks back to reality, bridging the gap between the physical world and online social networking services. For example, in addition to informing the locations where they are, people using mobile devices can leave their opinion about these locations to the different OSNs in which are participating. Therefore, people create their own location-related stories in the physical world and browse other people's information as well. OSNs offering services related to locations are called *location-based social networks* (LBSNs).

Location-based social networks expand the traditional structure of social networks from a simple graph with a single type of nodes and links to a heterogeneous structure, with different types of nodes and links. The natural heterogeneity of LBSNs is due to the appearance of the new interdependence existing among users and their locations. Understand this new type of interdependence can offer the support for also understand new user behaviors in LBSNs (EAGLE; PENTLAND; LAZER, 2008; ZHENG; ZHOU, 2011; GROH *et al.*, 2013; CHORLEY; WHITAKER; ALLEN, 2015; FELLEGARA *et al.*, 2016). Therefore, different link mining tasks can be performed to extract, process, and understand the behaviors of users in relation to their visited locations, as well as to discover new applications based on locations and improve the quality of the existing ones (GAO; LIU, 2015).

In this section, we will provide a comprehensive review of LBSNs, covering concepts, basic definitions as well as the main mining tasks performed on LBSNs, applications and challenges. Thus, in Section 2.3.1 we briefly introduce the meaning of LBSNs, discussing about their different categories as well as the close relation between the research in LBSNs and human mobility. Since to work over an LBSN first is necessary understand its underlying network structure, in Section 2.3.2 we formally define the heterogeneous structure of LBSNs, as well as their basic topological properties. In Section 2.3.3, we introduce the basic concepts related to link prediction task in the LBSN domain. Finally, the main challenges to be faced to analyze and understand user behavior in LBSNs are pointed in Section 2.3.4.

2.3.1 Location-based Social Networks

As previously mentioned, an online social network (OSN) is a social structure made up of people connected by one or more specific types of interdependency, such as friendship, common interests, preferences, among others. OSNs reflects the real-life social networks among people through online platforms, providing different services for their users share different types of content, events, activities, and interests over the Internet (MISLOVE *et al.*, 2007; XIANG; NEVILLE; ROGATI, 2010; BHATTACHARYYA; GARG; WU, 2011; ZHENG; ZHOU, 2011).

In the last years, Wi-Fi and GPS-enabled devices are changing the way people interact with the Web and many other real world domains such as sensor tags attached to animals, GPS tracking systems on cars and airplanes and RFID tags on merchandise. With such a device, people is able to acquire present locations, search the information around them, design driving routes to a destination, etc. Furthermore, OSNs are offering services for users to share locations as well as the different impressions and experiences about these locations.

A location-based social network (LBSN) is a specific type of social networking platform in which geographical services are added to traditional social networks. Such additional information enable new social dynamics including not only those derived from the visits of users to same or similar locations but also the knowledge of common interests, activities and behaviors inferred from the set of locations visited by a person and the location-tagged data generated in these visits. Therefore, users of LBSNs can not only track and share their location-related information via either mobile devices or desktop computers, but also leverage collaborative social knowledge learned from user-generated and location-related content, such as GPS trajectories and geo-tagged multimedia content, e.g. pictures, audio and video (ZHENG; ZHOU, 2011; CHO; MYERS; LESKOVEC, 2011; ROICK; HEUSER, 2013; SYMEONIDIS; NTEMPOS; MANOLOPOULOS, 2014).

LBSNs also referred to as *Geographic Social Networks* or *Geo-social Networks* (SCELLATO *et al.*, 2010; LI; HSIEH, 2015), have attracted big attention of scientific community, and consequently have enable many novel applications in different research issues, such as social network analysis (LI *et al.*, 2008; EAGLE; PENTLAND; LAZER, 2008; EAGLE; MONTJOYE; BETTENCOURT, 2009; XIAO *et al.*, 2010; CRANSHAW *et al.*, 2010; QUERCIA *et al.*, 2010; YING *et al.*, 2010; YING *et al.*, 2011; MA *et al.*, 2012; BRAGA *et al.*, 2012; XIAO *et al.*, 2014; GRABOWICZ *et al.*, 2014; CHEN; PANG; XUE, 2014; CHENG; PANG; ZHANG, 2015; BAO *et al.*, 2015; BAGCI; KARAGOZ, 2016; XU-RUI; LI; WEI-LI, 2016; LIAO *et al.*, 2016), spatio-temporal data mining (ZHENG *et al.*, 2009; ZHOU; MENG, 2011; LIU *et al.*, 2011; YUAN *et al.*, 2011; HAI *et al.*, 2012; HAI *et al.*, 2013; YANG; GUO; JENSEN, 2013; CHIANG; HOANG; LIM, 2015; ASGHARI *et al.*, 2015; WANG *et al.*, 2016), spatio-temporal databases (WANG *et al.*, 2008; DOYTSHER; GALON; KANZA, 2010; CHEN *et al.*, 2010; TANG *et al.*, 2011; HASHEM *et al.*, 2013; ZHENG; ZHANG; YU, 2015), ubiquitous computing (ZHENG *et al.*, 2008; ZHENG *et al.*, 2011; ZHENG *et al.*, 2014; ZHANG *et al.*, 2013; CHIKHAOUI *et al.*,

2014; RANVIER *et al.*, 2015; JI; ZHENG; LI, 2016) and information retrieval (FERRARI *et al.*, 2011; TAHRAT; ROCHE; TESSEIRE, 2012; LOGLISCI *et al.*, 2012a; LOGLISCI *et al.*, 2012b; BOUILLOT; PONCELET; ROCHE, 2012; PRIEDHORSKY; CULOTTA; VALLE, 2014; HAN; COOK; BALDWIN, 2014; LIU; HUANG, 2016).

Location-based Social Networks and Human Mobility

The introduction of smartphones in early 2000 signalled a massive transition in the way people were accessing the web. Users became mobile, carrying a computational device capable of accessing the online world from almost anywhere. The first OSNs that explicitly have used locations as their primary feature have appeared around 2008. Foursquare, Gowalla and Brightkite were the first LBSNs, and their service was based on a rather simple notion: *share with your friends information about the place where you are*. Later, other LBSNs as Flickr, Twitter, Instagram, and Facebook, have associated locations with social media content (ZHENG; ZHOU, 2011; NOULAS, 2013; WANG *et al.*, 2013).

LBSNs allow to their users interact among them as well as with locations. This fact makes possible to learn more about users, such as their geographic position, types of activities, times they spend in a location and, from a social network perspective, with whom. Furthermore, since location and geography acquire a key role in LBSNs, we have the opportunity to lay new foundations on our understanding of human mobility (CRANSHAW *et al.*, 2010; CHO; MYERS; LESKOVEC, 2011; ALHARBI; ZHANG, 2014; JURDAK *et al.*, 2015).

Human mobility describes the movement of people over time. The movement is motivated by different factors, such as economic activities, personal preferences, social forces, and others (DOMÍNGUEZ-MUJICA, 2016). The initial studies to frame the understanding of human mobility suggest that geographical distance has a deterring effect on mobility (SJAASTAD, 1962; ZELINSKY, 1971; GREENWOOD, 1975). These studies have analyzed the human mobility process using data collected through census considering a limited number of human participants. This limitation has provided only a very static viewpoint of human movement. Furthermore, census data suffers from limited spatial and temporal granularity in the description of human movement.

Despite census data provides unprecedented insight into large scale human movement patterns, it is not capable of capturing human movement activity occurring daily within and between specific geographic regions, such as neighborhoods, cities, states and countries. On the other hand, since data generated from LBSNs stores knowledge about how people use urban spaces, how they commute to work, and where they live, it is expected that the research in LBSNs has a depth impact for the foundations of novel human mobility and user behavior theories (MA *et al.*, 2012; STATE; WEBER; ZAGHENI, 2013; ZHENG *et al.*, 2014; MESSIAS *et al.*, 2015; GAO; LIU, 2015; CHORLEY; WHITAKER; ALLEN, 2015; JI; ZHENG; LI, 2016).

Location-based Social Network Categories

Zheng and Zhou (2011) have categorized the existing location-based networking services into three categories: *geo-tagged-media-based*, *point-location-driven*, and *trajectory-centric*.

Geo-tagged-media-based. This type of LBSN offers geo-tagging services for users add a location label to media content such as text, photos, and videos generated in the physical world. Therefore, users can browse their content at the exact location where it was created by using a mobile device. This fact allows users interact with the media via comments, content sharing, likes and other type of tools offered by LBSNs. Representative LBSNs included in this category are Flickr, Panoramio[4], Twitter, Instagram, and Facebook.

An important fact to have into consideration is that, although a location dimension has been added to OSNs of this category, the focus of such services is still on the media content. That is, location is used only as a feature to organize and enrich media content, since strong interdependencies among users are based on the media itself.

Point-location-driven. This type of LBSN offers for users the option of sharing explicitly their current locations. Generally, points and badges of locations are awarded for "checking in" at venues. With real-time location information, users can discover new friends around their physical location so as to enable certain social activities in the physical world. Also, this type of LBSNs allows that users interact among them about specific locations via comments, likes and other type of tools. Representative LBSNs included in this category are Foursquare, Google Maps[5], Jiepang[6], Yelp[7], Bightkite, and Gowalla.

In Point-location-driven networks, locations are the main elements determining the interdependency connecting users. On the other hand, user-generated contents such as comments add semantic meanings to a point location.

Trajectory-centric. In this type of LBSN, users pay attention to both point locations and the detailed route connecting these point locations. Generally, besides the basic information, such as distance, duration, and velocity, about a particular trajectory, this type of LBSNs also offers the user's experiences related to their routes represented by tags, tips, and photos. Representative LBSNs included in this category are Bikely[8], Waze[9], and Microsoft GeoLife.

This type of LBSNs answering questions related to "how and what" and "where and when", providing services for support the user's travel experience by browsing or replaying the trajectory on a digital map, and follow the trajectory in the real world with a GPS-device.

[4] <https://www.panoramio.com/>
[5] <https://www.google.com.br/maps>
[6] <http://jiepang.com/>
[7] <https://www.yelp.com/>
[8] <http://www.bikely.com/>
[9] <https://www.waze.com/>

It is worthwhile to metion that geo-tagged-media-based networks are generally considered as OSNs offering location services, whilst point-location-driven and trajectory-centric ones are considered as explicit LBSNs. Furthermore, point-location-driven and trajectory-centric networks lie in two major distinctions. One is that a trajectory offers richer information than a point location, such as how to reach a location, the time length for travelling between two locations, the temporal duration that a user stayed in a location, the traffic conditions of a route, among others (FENG; ZHU, 2016). Hence, it is more likely accurate to understand users behaviors and interests in a trajectory-centric LBSN. However, the second aspect is related to the fact that point-location-driven LBSN users usually share their real-time location while those from trajectory-centric prefer to share information about a trajectory after a trip has finished. This fact can compromise some scenarios based on the real-time location of users, but at the same time it reduces to some extent the privacy issues in an LBSN (SHOKRI *et al.*, 2011; ROSSI *et al.*, 2015), i.e. when people see a user trajectory likely the user is no longer there.

The location data generated by geo-tagged-media-based and point-location-driven LBSNs can be modeled as a trajectory which might be used by trajectory-centric ones. For instance, we can sequentially connect the point locations of geo-tagged photos taken by a user over several days or order by time their check-in records. Any of these procedures generate a low-sampling-rate trajectory which can degenerate in a highly sparse trajectory, i.e. the distance and time interval between two consecutive points could be very big. Therefore, since the uncertainty existing in this type of trajectories is increased, to consider these trajectories into trajectory-centric LBSNs it is necessary special efforts (ZHENG *et al.*, 2012; OSPINA; MORENO; URIBE, 2015; LIU; LI, 2017).

Despite the rich information provided by trajectory-centric LBSNs, in the whole of this thesis we will pay closer attention to data generated from geo-tagged-media-based and point-location-driven LBSNs. The main reason for this choose, is the fact that geo-tagged-media-based and point-location-driven LBSNs offer, in addition to location history data of users, data related to the different types of relationships among users, whilst trajectory-centric LBSNs offer data limited to individual routes of users without consider relationships among them. Hence, geo-tagged-media-based and point-location-driven LBSNs presents an attractive environment for researchers who are interested in the interplay of human mobility and social interaction.

Elements for Location-based Social Network Analysis

The layers of information produced by LBSNs establish a set of elements from which is possible perform the analysis of different properties of this type of complex networks (ZHENG; ZHOU, 2011; WANG *et al.*, 2011; NOULAS, 2013; LONG, 2015). Therefore, the elements for LBSN analysis more relevant to this thesis are:

Check-in. Check-in is the single most central element of LBSNs. When a person is at a location,

he/she can communicate his/her presence there by exploiting the *check-in* feature of the LBSN. Typically the GPS sensor of the mobile phone is exploited to obtain the respective geographic position encoded via latitute and longitude coordinates format. Some LBSNs offer a granularity service which automatically queries a database with millions of recorded venues from around the world and returns to the user a list of nearby places. Finally, the user *checks in* at the venue where he/she really is and choose to share this information publicly or privately in his/her LBSN profile.

Venue Database. The venue database constitutes the core wheel of LBSNs. The venue database not only has knowledge become synchronously available about the exact location users go to, but also semantic information about their types, e.g. the knowledge that -22.007031 and -47.894742 correspond to the latitute and longitude coordinates, respectively, of Institute of Mathematics and Computer Sciences Institute of University of São Paulo, at São Carlos city of São Paulo state, in Brazil. Furthermore, venue databases of some LBSNs also store user generated textual content such as tips, comments or tags as well as multi-media content that includes photos, audio and videos, together with geographic, temporal and social information about user check-ins and meetings.

Geographic Accuracy. LBSNs offer the opportunity to record the geographic position of a user not only with GPS accuracy (10-20 metres approximately), but also associate this position with a specific real-world venue. Thus, when a user declares his/her presence at a location by a check-in, he/she is labeling a specific geographic position. This is significant to the best understanding of human mobility and user behavior since more information about locations, besides offering a granular geographic representation of human movement, may be used as a proxy to infer activities and behaviors of users.

2.3.2 Network Representation

In this section, we will explore the traditional topological structure of an LBSN. Furthermore, we will present a simple and coherent notation to ease represent the different elements of an LBSNs as well as the interaction among them. Finally, using the notation defined and the topological LBSN presented, we will also introduce different LBSN properties that will be used through this thesis.

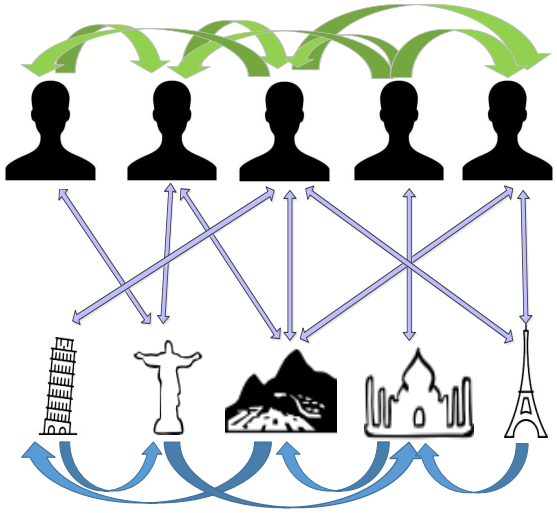
Network Structure

As previously described, users visit some locations in the physical world, leaving their location histories and generating location-tagged media content. If we sequentially connect these locations in terms of time, a trajectory will be created for each user. In this context, user and location are two major actors closely associated with each other in a location-based social network. The interdependencies between users (e.g. friendships), between users and locations

(e.g. check-ins), and between locations (e.g. trajectories) constitute the three fundamental ways in which the actors of an LBSN establish relationships with each other (ZHENG; ZHOU, 2011; CHO; MYERS; LESKOVEC, 2011; ROICK; HEUSER, 2013; SYMEONIDIS; NTEMPOS; MANOLOPOULOS, 2014; FELLEGARA *et al.*, 2016).

In Figure 18, we illustrate the traditional graph structure of an LBSN. From this figure we can observe two types of nodes, users and locations, and three types of links, links between users, links between locations, and links between users and locations. The presence of more than one type of nodes and more than one type of links reveals the heterogeneous nature of LBSNs. Furthermore, the entire graph structure of an LBSN can also be seen as a structure formed by three graphs: a user-user graph, a location-location graph, and a user-location graph.

Figure 18 – Traditional structure of a location-based social network.



Source: Elaborated by the author.

The *user-user graph* is a graph with a traditional structure, i.e. only one type of nodes representing users, and one type of links representing the interactions among users. The user-user graph is also referred to as *social graph*. Given the links in this graph connect one user to other, they are also called to as *social links* or *user links*. Furthermore, users links consist of, at least, two folds. The first is the original connection between two users in an existing social network, i.e. friendship, kinship, partnership, job contact, etc. The second is the new interdependency derived from their locations, i.e. the derived connection between two user by having visited the same location, or similar types of places, in the real world over a certain number of visits. This information, initially inferred from a user's locations, can be transferred to the original connection through a recommendation mechanism, i.e. it is possible to recommend users to an individual based on the inferred location interdependency.

The *location-location graph* is also a graph with traditional structure, in which nodes represent locations and links represent a connection between two locations. Generally, a link

connecting two locations is represented by a directed edge. The directionality of links in a location-location graph is motivated by the fact that links indicate that some users have consecutively traversed these two locations during a trip. It is possible to establish weighted links representing the frequency of traverse between the locations connected by the links. The location-location graph is also called simply as *location graph*.

The *user-location graph* is a heterogeneous graph with two types of nodes, i.e. users and locations, and only one type of link, i.e. the interaction between the users and the different existing locations. Therefore, a link in a user-location graph is also directed since it starting from a user and ending at a location indicating that the user has visited this location. It is possible to establish weighted links representing the frequency of visits to locations.

Notation

Formally, we represent an LBSN as an undirected and unweighted network $G = (V, E, \mathscr{L})$, where V is the set of users, E is the set of edges representing the social links among users, and $\mathscr{L}$ is the set of different locations visited by all the users. Multiple links and self-connections are not allowed.

Since the interaction between users and locations is established by check-ins, we employ the *Location Data Record* (LDR) to represent a check-in (WANG *et al.*, 2011; SCELLATO; NOULAS; MASCOLO, 2011; MENGSHOEL *et al.*, 2013). Let a user $x \in V$ visiting a location $\ell \in \mathscr{L}$, a LDR is defined by a tuple $\theta = (x, t, \ell)$, where t is the check-in time. Consider that the set of all LDRs in G is defined as Φ, and the size of this set, $|\Phi|$, defines the total number of check-ins made by all the users. In this context, the following sets can be defined:

- *Check-ins of a user x*, defined as $\Phi(x) = \{(x,t,\ell) \mid \exists z \in V : x = z \wedge (z,t,\ell) \in \Phi\}$. The size of this set, $|\Phi(x)|$, represents the number of check-ins made by the user x at different locations.
- *Check-ins at a location* ℓ, defined as $\Phi(\ell) = \{(x,t,\ell) \mid \exists \ell' \in \mathscr{L} : \ell = \ell' \wedge (x,t,\ell') \in \Phi\}$. The size of this set, $|\Phi(\ell)|$, represents the number of check-ins made by different users at the location ℓ
- *Locations visited by a user x*, defined as $\Phi_{\mathscr{L}}(x) = \{\ell \mid \forall \ell \in \mathscr{L} : (x,t,\ell) \in \Phi(x)\}$. The cardinality of this set, $|\Phi_{\mathscr{L}}(x)|$, indicates the number of different locations visited by user x.
- *Check-ins of a user x at a location* ℓ, defined as $\Phi(x,\ell) = \{(x,t,\ell) \mid (x,t,\ell) \in \Phi(x) \wedge \ell \in \Phi_{\mathscr{L}}(x)\}$. The size of this set, $|\Phi(x,\ell)|$, represents the number of check-ins made at location ℓ by user x.
- *Visitors of a location* ℓ, defined as $\Phi_V(\ell) = \{x \mid (x,t,\ell) \in \Phi(x) \wedge \ell \in \Phi_{\mathscr{L}}(x)\}$. The size of this set, $|\Phi_V(\ell)|$, indicates the number of different visitors of location ℓ.

It is important to note that all these sets can be influenced by the time. For instance, the set of check-ins of a user x during a time interval $[t_b, t_e)$ is defined as $\Phi(x, t_b, t_e) = \{(x,t,\ell) \mid \exists z \in V : x = z \wedge (x,t,\ell) \in \Phi \wedge t \in [t_b, t_e)\}$, the set of check-ins at a location ℓ during the same interval is defined as $\Phi(\ell, t_b, t_e) = \{(x,t,\ell) \mid \exists \ell' \in \mathscr{L} : \ell = \ell' \wedge (x,t,\ell) \in \Phi \wedge t \in [t_b, t_e)\}$, and so on.

Basic Properties

Different research studies have been carried out on the topological structure of LBSNs considering spatial, social and temporal factors (ALLAMANIS; SCELLATO; MASCOLO, 2012; CHO; MYERS; LESKOVEC, 2011). These studies have identified different LBSNs properties, such as user locality, user displacement, geographic clustering coefficient, radius of gyration, and other (CRANSHAW *et al.*, 2010; SCELLATO *et al.*, 2010; SCELLATO; NOULAS; MASCOLO, 2011; CHENG *et al.*, 2011; GRABOWICZ *et al.*, 2014; ZHANG; PANG, 2015). Here, we focus only on four basic properties since they have been commonly used for most of the link mining task in the LBSN domain.

Geographical Distance. Considering that a location can be represented in absolute (latitude-longitude coordinates), relative (100 meters north of the shopping mall), and symbolic (e.g. home, office, museum, etc.) form (ZHENG; ZHOU, 2011), it is important to define an effective way of determining the closeness or remoteness of a user checked-in at a specific location ℓ to another one checked-in at another location ℓ'. For this purpose, and considering that the location information of ℓ and ℓ' is in absolute form, the geographical distance is defined as:

$$dist(\ell, \ell') = 2R \times \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta lat}{2}\right) + c \times \sin^2\left(\frac{\Delta lon}{2}\right)}\right), \tag{2.44}$$

where R is the general radius of the Earth, Δlat and Δlon are, respectively, the differences among latitudes and longitudes coordinates of locations ℓ and ℓ', and c is the product of the cosine of the latitude of ℓ and the cosine of latitude of ℓ'. Note that all latitudes and longitudes need to be in radians to be processed in trigonometric functions.

As observed, the geographical distance, also called geospatial distance, is simply the Haversine formula to calculate the great-circle distance between two points over the earth's surface (GOODWIN, 1910). Note that Eq. 2.44 does not truly account for altitude regions such as hills, so more accurate versions of this equation can be calculated with additional information on the area in which the geographical distance will be calculated (ROBUSTO, 1957).

Place Entropy. This is a very useful property to quantify the strength of the relationship of a location and its visitors. Place entropy is based on the relation between check-ins of each visitor at a location and all check-ins that have taken place at the location (SCELLATO;

NOULAS; MASCOLO, 2011; CRANSHAW *et al.*, 2010). The place entropy of a location ℓ is defined as:

$$\mathscr{E}(\ell) = -\sum_{x \in \Phi_V(\ell)} q_{x,\ell} \log(q_{x,\ell}), \tag{2.45}$$

where $q_{x,\ell} = \frac{|\Phi(x,\ell)|}{|\Phi(\ell)|}$, called as relevance of check-ins of a user, is the fraction of check-ins made by user x at location ℓ with respect to the total number of check-ins at location ℓ.

Locations with higher place entropy might result in less social links among their visitors than those with lower values. The average of the place entropy of all locations in $\mathscr{L}$ is called as *average of places entropy*, $\langle \mathscr{E} \rangle$.

Location Diversity. The more popular a location the greater its place entropy. Moreover, instead of considering the place entropy directly, we can use the location diversity to represent a location's popularity (CRANSHAW *et al.*, 2010; ZHANG; PANG, 2015). The location or place diversity of location ℓ is defined as:

$$diversity(\ell) = \exp(\mathscr{E}(\ell)). \tag{2.46}$$

Adjusted Geographical Distance. The social strength between a pair of users x and y calculated from the geographical distance between two locations visited by these users can be adjusted if it is multiplied by the maximal location diversity of these two locations (ZHANG; PANG, 2015). Therefore, the adjusted geographical distance between two locations ℓ and ℓ', is defined as:

$$dist_{adj}(\ell, \ell') = dist(\ell, \ell') \times \max(diversity(\ell), diversity(\ell')). \tag{2.47}$$

The adjusted geographical distance certifies that the distance between popular frequent places is increased, while long distances between unpopular places are reduced.

2.3.3 Link Prediction in Location-Based Social Networks

Cho, Myers and Leskovec (2011) have studied the connection between human mobility and friendship by exploring the user movements and behavior in LBSNs. Their work show that 10% to 30% of all user movements can be explained by their social relationships and 50% to 70% human movements are related to periodic behavior. Therefore, it is possible to explore the graph structure of an LBSN and perform over it different mining tasks which can help to understand user behavior (ZHENG; ZHOU, 2011; NOULAS, 2013; LONG, 2015; CHORLEY; WHITAKER; ALLEN, 2015).

Similarly as in link mining on traditional networks, in the LBSN context most of mining tasks also depend of a similarity measure. The similarity estimation between a pair of users in the LBSN context, besides information obtained from their user profiles, topological properties,

and others, should also consider information from their individual's location histories. The user's location history in the real world implies, to some extent, his/her interests and behaviors related to specific locations. Therefore, people who share similar location histories are likely to have common interests and behavior. Different researchers have proposed a variety of similarity measures in the LBSN context (LI *et al.*, 2008; YING *et al.*, 2010; LEE; CHUNG, 2011; YUAN; JIANG; GIDóFALVI, 2013; LV; CHEN; CHEN, 2013; ZOU; XIE; SHA, 2015; MOHAMED; ABDELMOTY, 2016).

Using similarity measures based on location histories is possible perform different mining tasks on LBSNs, such as classification (ROSSI; MUSOLESI, 2014; YU *et al.*, 2015), community detection (BROWN *et al.*, 2012; WANG *et al.*, 2014; LIU *et al.*, 2016), link prediction (MENGSHOEL *et al.*, 2013; BAYRAK; POLAT, 2014; HSIEH; LI; LIN, 2015; KYLASA; KOLLIAS; GRAMA, 2016), among others. As discussed in previous chapters, link prediction is commonly used due to its ability to capture both user and relationship patterns and so identify the different ways in which network actors could to establish new relationships (LÜ; ZHOU, 2011; ZHANG; KONG; YU, 2014; ZHANG; PANG, 2015).

Considering the natural heterogeneity of LBSNs, the link prediction problem should be faced considering the specific existing types of links, e.g. friendship prediction involves predicting user-user links, location prediction focus on predict user-location links, and so on (LI *et al.*, 2008; ZHENG; ZHOU, 2011; BAO *et al.*, 2015).

Friendship Prediction

Friendship prediction is a traditional link prediction application, providing users with promising potential friends based on their relationship patterns, and the social structure of the network (CHO; MYERS; LESKOVEC, 2011; LUO *et al.*, 2013). Friendship prediction in LBSNs is the link prediction task focuses on predicting social links.

Friendship prediction have been widely explored in LBSNs since it is possible to use the large number of existing link prediction methods presented in Section 2.2.3. However, as location information and mobility user patterns significantly improve the effectiveness of friendship prediction (CRANSHAW *et al.*, 2010; MENGSHOEL *et al.*, 2013; PHAM; SHAHABI; LIU, 2013; XIAO *et al.*, 2014), this improvement is even greater when other information sources, such as geographical distance (ZHANG; PANG, 2015), GPS and/or check-ins history (CHO; MYERS; LESKOVEC, 2011; YU *et al.*, 2011; KYLASA; KOLLIAS; GRAMA, 2016), semantic of locations (tags, categories, etc.) (XIAO *et al.*, 2010; BAYRAK; POLAT, 2014; BAYRAK; POLAT, 2016), and others, are employed.

The friendship prediction is commonly applied to friendship recommender systems (BAO *et al.*, 2015). Furthermore, friendship prediction task has opened the doors to new research issues and applications, such as companion prediction (LIAO *et al.*, 2016), local experts prediction (CHENG *et al.*, 2014a; CHENG *et al.*, 2014b), and others. Since in this thesis we are

focusing on friendship prediction, we will show details about some of the most used friendship prediction methods in LBSNs.

Friendship Prediction Methods

As already outlined, it is our aim to study to which extent friendship prediction methods for LBSNs perform better when they are evaluated under different factors. For this purpose, we surveyed and selected 22 different link prediction methods which focus specifically on friendship prediction in LBSNs. Furthermore, we divide these methods into three groups based on the different similarity criteria which are based on: i) frequency, ii) information gain, iii) geographical distance, and iv) social strength.

A. Frequency-based Methods

Friendship prediction methods based on frequency are the most commonly used in the LBSN domain (CRANSHAW *et al.*, 2010; SCELLATO; NOULAS; MASCOLO, 2011; MENG-SHOEL *et al.*, 2013; STEURER; TRATTNER; HELIC, 2013; STEURER; TRATTNER, 2013; BAYRAK; POLAT, 2014; KYLASA; KOLLIAS; GRAMA, 2016). These methods calculate user similarity considering only the count of check-ins or places, or both. The methods we consider in this category are the following:

Collocations (Co). This is one of the most popular methods based on check-ins frequency. For a pair of users x and y, and considering a temporal threshold of time, $\tau \in \mathbb{R}$, Co is defined as:

$$s^{Co}_{x,y,\tau} = |\Phi_{Co}(x,y,\tau)|, \tag{2.48}$$

where $\Phi_{Co}(x,y,\tau) = \{(x,y,t_x,t_y,\ell) \mid (x,t_x,\ell) \in \Phi(x) \ \wedge \ (y,t_y,\ell) \in \Phi(y) \ \wedge \ |t_x - t_y| \leq \tau\}$, is called as the *set of collocations*, i.e. the set of check-ins made by both users x and y at the same place and in the same period of time.

Collocations method, also called as *number of collocations* or *common check-ins count*, expresses the number of times that users x and y visited some location at the same time. For instance, if users x and y meet daily for a week ($\tau = 7$ days) at one particular coffee store, this gives $s^{Co}_{x,y,7} = 7$.

Distinct Collocations (DCo). This is also based on the set of collocations, but by considering the number of different places where users x and y have common collocations in a temporal threshold of $\tau \in \mathbb{R}$, as a homophily indicator. Therefore, DCo can be computed as:

$$s^{DCo}_{x,y,\tau} = |\{\ell \mid (x,y,t_x,t_y,\ell) \in \Phi_{Co}(x,y,\tau)\}|. \tag{2.49}$$

Thus, if users x and y meet themselves daily for a week ($\tau = 7$ days) at one particular coffee store, distinct collocations method gives $s^{DCo}_{x,y,7} = 1$.

Common Location (CL). This is the simplest and most popular method based on place frequency to determine the homophily between a pair of users. Let the *set of common visited places* of a pair of users x and y be defined as $\Phi_{\mathscr{L}}(x,y) = \Phi_{\mathscr{L}}(x) \cap \Phi_{\mathscr{L}}(y)$, the common location method is defined as:

$$s_{x,y}^{CL} = |\Phi_{\mathscr{L}}(x,y)|. \tag{2.50}$$

This measure, also known as *common places* or *distinct common locations*, expresses the number of common locations for users x and y, not necessarily visited at the same time.

Jaccard of Places (JacP). This is defined as the fraction of the number of common locations and the number of locations visited by both users x and y. Therefore, JacP is computed as:

$$s_{x,y}^{JacP} = \frac{|\Phi_{\mathscr{L}}(x,y)|}{|\Phi_{\mathscr{L}}(x) \cup \Phi_{\mathscr{L}}(y)|}. \tag{2.51}$$

Location Observation (LO). This method is defined as the fraction of the number of common visited places and the sum of total number of places visited by each user. Therefore, LO method is computed as:

$$s_{x,y}^{LO} = \frac{|\Phi_{\mathscr{L}}(x,y)|}{|\Phi_{\mathscr{L}}(x)| + |\Phi_{\mathscr{L}}(y)|}. \tag{2.52}$$

Common Location Ratio (CLR). This refers to the ratio of the number of common locations of two users x and y to the sum of their total number of check-ins. CLR is computed as:

$$s_{x,y}^{CLR} = \frac{|\Phi_{\mathscr{L}}(x,y)|}{|\Phi(x)| + |\Phi(y)|}. \tag{2.53}$$

Common Locations Category (CLC). Inspired by the *term frequency-inverse document frequency* (tf-idf) ranking technique from information retrieval, the CLC is defined as the sum of ratios of check-ins count of users x and y at common locations to the all check-ins counts of all users at these locations. CLC is computed as:

$$s_{x,y}^{CLC} = \sum_{\ell \in \Phi_{\mathscr{L}}(x,y)} \frac{|\Phi(x,\ell)| + |\Phi(y,\ell)|}{|\Phi(\ell)|}. \tag{2.54}$$

Preferential Attachment of Places (PAP). This is the product of total different places visited by users x and y. PAP is defined as:

$$s_{x,y}^{PAP} = |\Phi_{\mathscr{L}}(x)| \times |\Phi_{\mathscr{L}}(y)|. \tag{2.55}$$

Preferential Attachment of Check-ins (PAC). This is the product of total different check-ins count of users x and y. PAC is computed as:

$$s_{x,y}^{PAC} = |\Phi(x)| \times |\Phi(y)|. \tag{2.56}$$

Adamic-Adar of Places (AAP). This is based on the traditional Adamic-Adar method but considering the number of check-ins of common visited places of users x and y. AAP is computed as:

$$s_{x,y}^{AAP} = \sum_{\ell \in \Phi_{\mathscr{L}}(x,y)} \frac{1}{\log |\Phi(\ell)|}. \tag{2.57}$$

Min Check-ins (MinC). This is defined as the minimum number of check-ins made by users x and y in all their common locations. For ease of understanding, we define the *set of minimum visits at common locations* of users x and y, $\Phi_{min}(x,y) = \{\min(|\Phi(x,\ell)|, |\Phi(y,\ell)|) \mid \forall\, \ell \in \Phi_{\mathscr{L}}(x,y)\}$, the MinC is computed as:

$$s_{x,y}^{MinC} = \min(\Phi_{min}(x,y)). \tag{2.58}$$

For this case, two users are more likely to have a future friendship if they have a low value of MinC.

B. Methods Based on Information Gain

As previously mentioned, place entropy seems to offer a strong discriminative power to determine of whether a certain place is relevant to the formation of social ties between its visitors. Therefore, some researches try to take advantage of this to calculate the likelihood of two disconnected users can establish a future friendship (CRANSHAW *et al.*, 2010; SCELLATO; NOULAS; MASCOLO, 2011; LUO *et al.*, 2013; BAYRAK; POLAT, 2014). We call all these methods as based on information gain and they are described below.

Min Entropy (MinE). Since entropy is a measure of impurity, MinE method is defined as the minimum place entropy across all the common locations visited by a pair of users x and y. Therefore, considering the set of *common locations entropy* of users x and y, $\Phi_{\mathscr{E}}(x,y) = \{\mathscr{E}(\ell) \mid \forall \ell \in \Phi_{\mathscr{L}}(x,y)\}$, the MinE method is computed as:

$$s_{x,y}^{MinE} = \min(\Phi_{\mathscr{E}}(x,y)). \tag{2.59}$$

Hence, the relevance of a place is higher if it has a low entropy value, i.e. two users are more likely to have a future friendship if they have a low value of MinE.

Adamic Adar of Entropy (AAE). Also apply the traditional Adamic-Adar method but considering the place entropy for common locations of a pair of users x and y. Therefore, AAE method is defined as:

$$s_{x,y}^{AAE} = \sum_{\ell \in \Phi_{\mathscr{L}}(x,y)} = \frac{1}{\log \mathscr{E}(\ell)}. \tag{2.60}$$

Location Category (LC). This calculates the total sum of the ratio of the number of check-ins of all locations visited by users x and y to the number of check-ins of users x and y in these

locations disregarding those with a high place entropy. Therefore, considering an entropy threshold $\tau_{\mathscr{E}} \in \mathbb{R}$, LC method is defined as:

$$s_{x,y}^{LC} = \sum_{\ell \in \Phi_{\mathscr{L}}(x) \ \wedge \ \mathscr{E}(\ell) < \tau_{\mathscr{E}}} \ \sum_{\ell' \in \Phi_{\mathscr{L}}(y) \ \wedge \ \mathscr{E}(\ell') < \tau_{\mathscr{E}}} \frac{|\Phi(\ell)| + |\Phi(\ell')|}{|\Phi(x,\ell)| + |\Phi(y,\ell')|}. \tag{2.61}$$

C. Methods Based on Geographical Distance

Different researches have demonstrated the importance of geographical or geospatial distance in the establishment of social ties (SCELLATO *et al.*, 2010). Therefore, many research have proposed to exploit this fact focusing on friendship prediction (SCELLATO; NOULAS; MASCOLO, 2011; BAYRAK; POLAT, 2014; ZHANG; PANG, 2015; XU-RUI; LI; WEI-LI, 2015; XU-RUI; LI; WEI-LI, 2016). We call these methods as based on geographical distance and they are as follows:

Min Distance (MinD). This is the smallest value among all the distances calculated from all the locations visited by user x and by user y. Therefore, considering the *set of all geographical distances* between the locations visited by users x and y, $\Phi_{dist}(x,y) = \{dist(\ell,\ell') \mid \forall \ell \in \Phi_{\mathscr{L}}(x) \ \wedge \ \forall \ell' \in \Phi_{\mathscr{L}}(y)\}$, the MinD method is defined as:

$$s_{x,y}^{MinD} = \min(\Phi_{dist}(x,y)). \tag{2.62}$$

For this case, two users are more likely to establish a future friendship if they have a low value of MinD.

Check-in Distance (ChD). This is defined as the proportion between the sum of the average of geographical distance between the locations of all check-ins made by users x and y and the total sum of geographical distance between the locations of all their check-ins. Thus, ChD is computed as:

$$s_{x,y}^{ChD} = \frac{\sum_{\ell \in \Phi(x)} \frac{\sum_{\ell' \in \Phi(y)} dist(\ell,\ell')}{|\Phi(y)|}}{\sum_{\ell \in \Phi(x)} \sum_{\ell' \in \Phi(y)} dist(\ell,\ell')}. \tag{2.63}$$

Check-in Location (ChL). This computes the proportion of total number of check-ins in all the places visited by x and y and the number of check-ins made by them only if these place are geographically close. Assuming a geographical distance threshold $\tau_d \in \mathbb{R}$, ChL is defined as:

$$s_{x,y}^{ChL} = \sum_{\ell \in \Phi_{\mathscr{L}}(x)} \ \sum_{\ell' \in \Phi_{\mathscr{L}}(y) \ \wedge \ dist(\ell,\ell') < \tau_d} \frac{|\Phi(\ell)| + |\Phi(\ell')|}{|\Phi(x,\ell)| + |\Phi(y,\ell')|}. \tag{2.64}$$

GeoDist (GeoD). It computes the geographical distance between the home locations of users x and y. Consider as "home location" of user x, ℓ_x^h, to the most checked-in location. Thus, GeoD is calculated as:

$$s_{x,y}^{GeoD} = dist(\ell_x^h, \ell_y^h). \tag{2.65}$$

For this case, two users are more likely to establish a friendship if they have a low value of GeoD.

Weighted Geodist (WGeoD). Defined as the geographical distance between the home locations of users x and y divided by the product of the number of check-ins each user has done in their respective home locations. Therefore, WGeoD is computed as:

$$s_{x,y}^{WGeoD} = \frac{dist(\ell_x^h, \ell_y^h)}{|\Phi(x, \ell_x^h)| \times |\Phi(y, \ell_y^h)|}. \tag{2.66}$$

Hausdorff Distance (HD). Based on the classic Hausdorff distance, this method quantify the distance between users x and y considering only specific visited places. Therefore, considering that sup represents the *supremum* (least upper bound) and inf represents the *infimum* (greatest lower bound) from the set of visited places of a user, HD is defined as:

$$s_{x,y}^{HD} = \max\{\sup_{\ell \in \Phi_{\mathscr{L}}(x)} \inf_{\ell' \in \Phi_{\mathscr{L}}(y)} dist(\ell, \ell'), \sup_{\ell' \in \Phi_{\mathscr{L}}(y)} \inf_{\ell \in \Phi_{\mathscr{L}}(x)} dist(\ell, \ell')\}. \tag{2.67}$$

Despite to HD method try to maximize the geographical distance between two locations, two users will be more likely to establish a relationship if they have a low value of HD.

Adjusted Hausdorff Distance (AHD). Similar to the HD method, but based on the adjusted geographical distance. Therefore, AHD method is defined as:

$$s_{x,y}^{AHD} = \max\{\sup_{\ell \in \Phi_{\mathscr{L}}(x)} \inf_{\ell' \in \Phi_{\mathscr{L}}(y)} dist_{adj}(\ell, \ell'), \sup_{\ell' \in \Phi_{\mathscr{L}}(y)} \inf_{\ell \in \Phi_{\mathscr{L}}(x)} dist_{adj}(\ell, \ell')\}. \tag{2.68}$$

Also similar to HD method, two users will be more likely to establish a relationship if they have a low value of AHD.

D. Social Strength-based Methods

Despite the fact that most of the previously described methods capture different social behavior patterns of users at their visited places, they do not directly use the social strength of ties between a pair of disconnected users and their common friends. In fact, a few number of methods in the literature try to predict the friendship between a pair of users based on the location information of their common friends. We call these methods as based on social strength.

We have identified only one friendship prediction method for LBSNs based on social strength, the *total common friend common check-ins* method, which was proposed by Bayrak and Polat (2014).

Total Common Friend Common Check-ins (TCFCC). This is defined as the sum of the product of the number of collocations of each common friend with each user x and y. Considering a temporal threshold $\tau \in \mathbb{R}$, TCFCC is computed as:

$$s_{x,y}^{TCFCC} = \sum_{z \in \Lambda_{x,y}} |\Phi_{co}(x, z, \tau)| \times |\Phi_{co}(y, z, \tau)|. \tag{2.69}$$

Location Prediction

Location prediction is a conventional task in the LBSN domain due to the nature of these networks. Thus, it focuses on predicting user-location links based mainly on the user's geo-social activity. Location prediction is a very broad topic and the methods in the literature can be divided into stand-alone and sequential location prediction (BAO *et al.*, 2015).

Stand-alone location prediction methods focus on providing a user with individual locations, such as restaurants, museums, touristic venues, cities, etc., according to their preferences and based on user's location history, social strength, and other patterns (PARK; HONG; CHO, 2007; BACKSTROM; SUN; MARLOW, 2010; LIU *et al.*, 2013; CHENG *et al.*, 2013; MCGEE; CAVERLEE; CHENG, 2013; SAIKAEW; JIRANUWATTANAWONG; TAEARAK, 2015; WANG *et al.*, 2016; ZHU *et al.*, 2016). On the other hand, sequential location prediction is a more complex task since it aims to predict a location path which maximize the number of interesting places to visit while minimizing travel time, energy consumption or other restrictions (YOON *et al.*, 2010; CHEN; SHEN; ZHOU, 2011; DOYTSHER; GALON; KANZA, 2011; WEI; ZHENG; PENG, 2012; HSIEH; LI; LIN, 2015; DAI *et al.*, 2015).

Collective Link Prediction

Most studies related to link prediction in LBSNs focus on either friendship or location prediction, and usually assume that the prediction tasks of different types of links are independent (ZHANG; KONG; YU, 2014). However, since user-user and user-location links are closely correlated and mutually influential, it is possible to improve the accuracy of friendship prediction if we have accurate location prediction results and vice versa. Therefore, a variety of methods performing collective friendship and location prediction task in LBSNs have been published (LI *et al.*, 2008; ZHENG *et al.*, 2011; SADILEK; KAUTZ; BIGHAM, 2012; ZHANG; KONG; YU, 2014).

2.3.4 Challenges

Actually, the variety of problems that traditional OSNs have to mining and understanding user behavior also exist in LBSNs. Furthermore, these problems become more challenging due to the following reasons:

Network heterogeneity. As previously mentioned, the network structure of an LBSN consists at least of two types of nodes, e.g. user and location, and three kinds of links, e.g. user-user, location-location, and user-location. Although we can only consider there are at least three tightly associated graphs modeling an LBSNs, e.g. social graph, location graph, and user-locaton graph. However, if it is a trajectory-centric LBSN, trajectories can be regarded as another kind of node in the social network. Similarly, we also can consider geotagged videos and photos as other types of nodes.

Despite the fact that the heterogeneity of an LBSN may be even greater than showed in Section 2.3.2, user and location are the main actors in an LBSN. Under the circumstances, to better understand some user behavior, such as the establishment of new relationships, it is necessary to involve the information from the other graphs forming the LBSN analyzed, such as the linking structure of user-location, location-location, and/or others, besides that of users.

Faster evolution process. LBSNs are constantly evolving at a faster pace than traditional OSNs, in both social structure and properties of nodes and links. This is motivated due to the ease that mobile devices offer to check-ins as well as to establish contact with new friends. Therefore, in LBSNs not only the number of users increase, but also the number of locations as well as the interactions among them.

Given the faster evolution process of LBSNs, it is necessary adequate mechanisms to process all this amount of data as quickly as possible and maintaining a high quality of offered services. For instance, if a user is walking and he/she ask for information about restaurants to dinner, a location prediction algorithm has to provide as quick as possible a set of options of nearby restaurants. It is up to the algorithm to process optimally the entire venue database and provide the best solution.

Location has unique features. Besides the fact of having individual intrinsic properties showed in Section 2.3.2, also the hierarchical and sequential properties of locations are unique. A location can be as small as a restaurant or as big as a city. Locations with different granularities formulate hierarchies between them. For instance, a restaurant belongs to a neighborhood, the neighborhood pertains to a city, which belongs to a country, and so on. Using different granularities, it is possible to obtain different location graphs even given the same location histories. This hierarchical property does not hold in other social networks.

The sequential property of locations is observed when each link between two locations is associated with temporal and directional information. Through the joining these links is possible to construct a particular sequence carrying a particular semantic meaning, e.g. the route from office to home.

These unique features of locations represent a big challenge since it is necessary to consider the natural restrictions they impose for performing different mining tasks. For instance, if any location prediction algorithm does not consider the sequential property, it could recommend to a user go to the disco at lunchtime, which has a high chance of not being a good suggestion.

2.4 Summary

In this chapter, we have widely discussed three main research topics tackled in this thesis. We started by introducing the basic concepts related to graph theory and some of the complex network models. We have also explored a list of network measurements, which are useful to extract structural information from real-world networks. This information may be used as one of the resources to achieve the objectives of this thesis. We have also briefly reviewed some of the classical categories of complex networks to show their peculiarities and highlight the fact that the focus of this thesis is on social networks. Our survey on complex networks concluded with a brief description of some recent research trends in complex networks, such as temporal and multi-layer networks, which can represent different instances of real-world social networks.

Afterwards, we have introduced the link prediction problem, placing it as a specific link mining task and emphasizing on the scope of the link prediction problem and on recent contributions. In the link mining research field, link prediction is the most representative task and the main framework in the context of this thesis. Therefore, we started by formally defining the link prediction problem considering two perspectives: as supervised and unsupervised process. We attempted to detail how each of these processes can work both individually and in combination. We later analyzed large amounts of existing link prediction methods, organizing them into two major groups: the first based on similarity, which can be calculated from node, structural and social information; and the second based on learning and higher-level approaches, which use probabilistic and statistical models as well as algorithmic and preprocessing strategies. Finally, we briefly described some of the most important practical applications of link prediction for a variety of real-world networks as well as their theoretical applications on several complex network processes and other link mining tasks.

It is important to highlight that, although it was not the purpose of this thesis, we have directly contributed to the link prediction research field. In a previous study, we showed how algorithms performing the same data mining task can obtain better results when applied over network data than over flat data (VALVERDE-REBAZA *et al.*, 2014). Furthermore, in collaboration with other researchers, we adapted the state-of-the-art link prediction methods to be used as part of other algorithms, improving results in different tasks, such as network construction (BERTON; VALVERDE-REBAZA; LOPES, 2015), identification of equivalent nodes (DRURY; VALVERDE-REBAZA; LOPES, 2015), and clustering (VALEJO *et al.*, 2014b). Detailing these projects is not in the scope of this thesis.

Considering the advances in link prediction research, we can observe that similarity-based methods may be a promising technique to address the link prediction problem in a fast and accurate way. We highlight methods based on social theory, specifically those based on community structure. The natural presence of communities in networks is important to improve link prediction accuracy without considerable increase in computational cost. Therefore, we believe that investigating user behaviors in communities or social groups allows us to better

understand the mechanisms to establish new relationships as well as to discover other types of hidden user behaviors.

Finally, we have introduced concepts related to location-based social networks (LBSNs), a new type of online social networks. A Location-based social network not considers location only as an attribute of user, but as an important actor of its network structure. Under this consideration, LBSNs are characterized by a heterogeneous topology, which is able to catch both user behavior and mobility. Therefore, we also presented a simple and coherent notation to ease represent the structure of an LBSN as well as its basic properties and challenges.

Due to our goal is understand user behavior by using the link prediction task, we have also presented as this mining task is performed over the heterogeneous structure of an LBSN. Essentially, depending of type of links which we aim to predict, the link prediction task in LBSNs can be instanced as friendship prediction, if we aim to predict friendship or other type of social relationships among users, or location prediction, if we aim to predict the possible locations which users could visit in the future.

In the next chapter, we will discuss in detail the friendship prediction task in OSNs, including LBSNs. Also, we will present a set of proposals which exploit different social patterns not considered by methods of literature. Extensive experimentations in large-scale real-world networks show that our proposals are as competitive as state-of-the-art methods. Furthermore, we will show details of as our proposals offer more benefits, in terms of computational resources optimization, to be used in real-world applications.

CHAPTER

3

MINING USER BEHAVIOR

As discussed in the previous chapter, link prediction is one of the most used mining tasks to understand user behavior. In this chapter we will describe our proposals to face the link prediction problem in both traditional OSNs and LBSNs. Also, we will evaluate and discuss the performance of our proposals regard to current link prediction methods. In Section 3.1, we will discuss the link prediction problem by considering that a user belong to one community found by any community detection algorithm. Furthermore, we will show a possibility to overcome the existing challenges for link prediction methods based on community information on real-world large-scale social networks. Afterwards, in Section 3.2, we will introduce a new link prediction research issue, which consists in take into account information about the participation of users in more than one social group to improve the link prediction accuracy. Then, we will present a set of proposals to overcome the challenges in this new issue. In these two sections, our efforts will be focused on mining, analyzing and understanding user behavior in traditional OSNs. In Section 3.3 we will present our proposals to face the link prediction problem in the context of LBSNs and better understand the behavior of their users. Finally, we conclude in Section 3.4 by addressing an overview of contributions presented in this chapter.

3.1 Friendship Prediction using Community Information

As discussed in the previous chapter, the link prediction task has been successfully explored in the past years by a wide spectrum of approaches, and has been used in a number of real-world applications. Due to its theoretical significance, link prediction is constantly used to explain the process of link creation among nodes, and consequently the network evolution process over time (KOSSINETS; WATTS, 2006; BRINGMANN *et al.*, 2010; CUI *et al.*, 2011; ANIL; SETT; SINGH, 2014). In the context of online social networks (OSNs), the dynamics of

the link creation process differs from that of other types of networks, mainly because users of OSNs constantly establish new friendships.

Many factors may lead to new friendship links among pairs of unknown users, such as preferences, common interests, places visited, and posts shared. All these factors can be captured by different state-of-the-art similarity methods. However, because of the large variety of OSNs offering numerous services, their individual behavior should also be a key factor to better understand the friendship establishment process.

Aware of the need to consider user behavior, different researchers have used several social theories (such as social ties, homophilia, and community) to better understand user behavior in a variety of OSNs and improve link prediction accuracy. Because several social theories can be enclosed and/or explained by the presence of communities, a considerable number of researchers have used this type of information as their main resource to understand the friendship establishment process.

This section shows the importance of using community information to obtain more accurate link prediction results without excessive computational cost. Section 3.1.1 briefly presents the main challenges related to the use of community information for efficient link prediction. Section 3.1.2 explains the strategy adopted to overcome the identified challenges in a comprehensive way. Section 3.1.3 shows experimental results obtained by comparing the efficiency of existing link prediction methods based on community information against state-of-the-art methods in Twitter, a popular large-scale OSN. It is important to highlight that our evaluations are performed using both unsupervised and supervised link prediction strategies to quantify the impact and relevance of using community information in the link prediction task. Section 3.1.4 closes with final remarks.

3.1.1 Challenges in Link Prediction using Community Information

Communities (also called groups, modules, or clusters) are natural structures in real-world networks. A community concentrates a group of nodes which are strongly connected due to their high similarity, whilst different communities are connected by a few links among weakly similar nodes (CLAUSET; NEWMAN; MOORE, 2004; FORTUNATO, 2010; NEWMAN, 2010; BARABÁSI, 2016). Some studies have shown how the presence of communities improve link prediction accuracy (FENG; ZHAO; XU, 2012; LIU *et al.*, 2013); therefore, different authors have proposed a variety of link prediction methods based on the use of community information (ZHELEVA *et al.*, 2008; SOUNDARAJAN; HOPCROFT, 2012; VALVERDE-REBAZA; LOPES, 2012a; CANNISTRACI; ALANIS-LOBATO; RAVASI, 2013; KEMAL; TSUYOSHI *et al.*, 2014; MALLEK *et al.*, 2015; BISWAS; BISWAS, 2017).

Despite their contribution to improve link prediction accuracy, link prediction methods based on community information face a challenging problem: *given a network where the com-*

munity information is unknown, how can we use link prediction methods based on community information without incurring in high computational cost? Most of the link prediction methods based on community information need some community detection algorithm, which can significantly increase the total computational cost of the whole link prediction process.

Consider a function $f(|V|)$ which defines the computational cost of any community detection algorithm. We can consider that the computational cost of a link prediction method based on similarity and using community information is $O(f(|V|)+|V|^2))$, because the computational cost of any community detection algorithm is $O(f(|V|))$ and the computational cost to compute the similarities between all the pairs of disconnected nodes is $O(|V|^2)$. The computational cost of state-of-the-art community detection algorithms is well known (FORTUNATO, 2010): the computational cost of methods based on hierarchical clustering ranges between $O(|V|^2)$ and $O(|V|^2 \log |V|)$; the computational cost of methods based on spectral clustering is $O(|V|^3)$; and the computational cost of methods based on divisive strategies ranges between $O(|V|^2)$ and $O(|V|^3)$. Therefore, the total computational cost of this type of link prediction methods is $O(|V|^2+|V|^3)$.

Probabilistic or statistical methods which naturally incorporate the use of communities in link prediction, such as those based on hierarchical structure or stochastic block models, tend to have exponential computational cost, i.e. $O(2^{|V|})$. Therefore, the use of community information could make the link prediction process practically unfeasible.

Link prediction methods based on community information pose yet another challenge: *what is the best community structure to be explored in order to reach higher link prediction accuracies?* We raise this question because different community detection algorithms (or even the same community detection algorithm with different configurations) may find different communities in the same network. The performance of link prediction may be improved with well-identified communities; however, there is no consensus on the best community detection algorithm.

The challenges mentioned above encourage researchers to find link prediction methods using well-defined communities in order to improve prediction accuracy with affordable computational cost. Overcoming these challenges will allow us to use link prediction methods based on community information in large-scale, real-world networks.

3.1.2 Improving Link Prediction using Community Information

Aware of the challenges of link prediction methods using community information in large-scale networks, here we briefly present the usual strategy adopted by the research community. Given a network G, the strategy consists of two main steps:

- Applying an optimal community detection algorithm which computes well-defined communities with low computational cost.

- Applying a similarity-based link prediction method capable of harnessing the information related to the previously found communities efficiently.

We intend to show the use of community information to improve link prediction accuracy. Therefore, this section briefly describes a community-detection algorithm as well as some similarity-based link prediction methods which can be used as part of the previously described strategy.

Finding Communities: The Label Propagation Algorithm

Several algorithms have been proposed to find community structures in networks (GIRVAN; NEWMAN, 2002; CLAUSET; NEWMAN; MOORE, 2004; WU; HUBERMAN, 2004; NEWMAN, 2004; PONS; LATAPY, 2006; RAGHAVAN; ALBERT; KUMARA, 2007; LEUNG *et al.*, 2009; ALMEIDA; LOPES, 2009; BARBER; CLARK, 2009; PAN *et al.*, 2010; ROSVALL; BERGSTROM, 2010; XIE; SZYMANSKI, 2013; HARENBERG *et al.*, 2014; LIN *et al.*, 2014; ZHOU; MARTIN; PAN, 2016; BEDI; SHARMA, 2016; LIU *et al.*, 2016). Community detection algorithms aim to find groups with an inherent or an externally specified notion of similarity among nodes. Furthermore, the number and size of communities in a network are not previously known; they are determined by the community detection algorithm.

The clustering detection algorithms that stand out are those capable of finding well-defined communities in large-scale networks with near-linear computational cost. One of these algorithms is the fast and simple *label propagation algorithm* (LPA) designed by Raghavan, Albert and Kumara (2007). Initially, LPA assigns a single label to each node, e.g. its node label. At each iteration, a pass over all nodes is performed in a random order: each node takes the label shared by the majority of its neighbors. If there is a tie, one of the labels is selected at random. In this way, labels are propagated across the graph: most labels will disappear, others will dominate. The process converges when each node shares the same label with the majority of its neighbors, or a maximum number of iterations is achieved. Communities are defined as groups of nodes having identical labels at convergence. At the end, each node has more neighbors in its community than in any other community.

This community detection algorithm does not deliver one unique solution. Due to the many relationships encountered along the detection process, it is possible to derive different partitions starting from the same initial condition, with different random seeds. Nonetheless, tests on real networks show that all partition found are similar.

The main advantage of LPA is the fact that the algorithm runs to its completion with a near-linear computational cost. Initializing every node with single labels requires $O(|V|)$ and each iteration for propagation of label takes linear computational cost in the number of edges, i.e. $O(|E|)$. The authors observed that LPA converges significantly after about 5 iterations. Therefore, LPA requires of an overall $O(|V|+|E|)$ computational cost.

Other algorithms may have computational cost similar to LPA, such as the algorithm proposed by Wu and Huberman (2004), which also performs with an $O(|V|+|E|)$ computational cost, and the algorithm proposed by Clauset, Newman and Moore (2004), which requires $O(|V|\log^2|V|)$. However, LPA shows certain advantages. The algorithm proposed by Wu and Huberman (2004) needs to know a priori how many communities are included in the network; this information is not available for real-world networks. Furthermore, if we know that there are M communities in the network, the Wu and Huberman (2004) algorithm can only find communities that are approximately of the same size, that is $\frac{|V|}{M}$. The main advantage of LPA over the Wu and Huberman (2004) algorithm is that LPA does not need any previous information about the number of communities to be found, and does not make restrictions based on community sizes. That means LPA only uses the network structure to guide its progress, requiring no external parameter settings. On the other hand, the algorithm proposed by Clauset, Newman and Moore (2004) tends to join in one single community the partitions with similar structure which were progressively computed, whilst LPA can find multiple partitions with similar structures without combining them. This implies that LPA can find not only one, but multiple significant community structures.

Because LPA is fast and capable of identifying significant and well-defined communities, a number of researchers have added improvements to it in terms of optimization processes (BARBER; CLARK, 2009). Some of these improvements focus on the labeling process by assigning weights or scores to existing nodes according to their importance (LIN *et al.*, 2014), while other strategies use information of known communities as prior labels (ZHOU; MARTIN; PAN, 2016). On the other hand, other improvements focus on the propagation process by using operators to control and stabilize the propagation dynamics (XIE; SZYMANSKI, 2013) or incorporating consensus strategies to optimize the propagation for densely connected networks (LIU *et al.*, 2016). However, we have chosen to use the basic LPA version proposed by Raghavan, Albert and Kumara (2007).

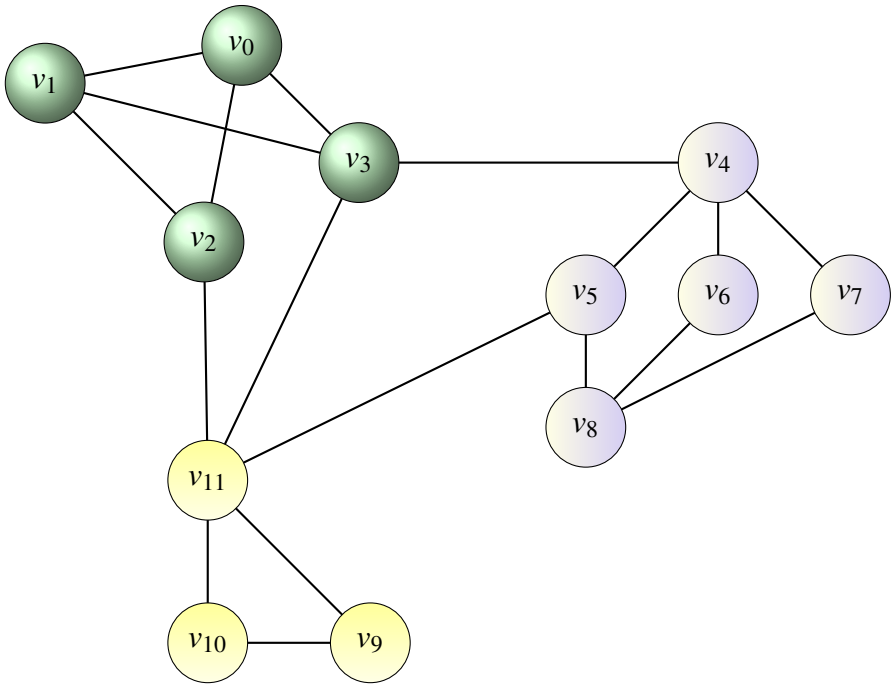
Predicting Links Using Community Information

Experiments have shown that for a network with low community structure, link prediction methods based on structural similarity perform poorly. Nonetheless, the accuracy of these link prediction methods improves drastically as the community structure of the network grows (FENG; ZHAO; XU, 2012; LIU *et al.*, 2013). Based on this consideration and aiming to exploit the advantages of applying link prediction methods based on local similarity, different authors have proposed a variety of these methods incorporating the use of community information (ZHELEVA *et al.*, 2008; SOUNDARAJAN; HOPCROFT, 2012; VALVERDE-REBAZA; LOPES, 2012a; VALVERDE-REBAZA; LOPES, 2013; HOSEINI; HASHEMI; HAMZEH, 2012; CANNISTRACI; ALANIS-LOBATO; RAVASI, 2013; KEMAL; TSUYOSHI *et al.*, 2014; MALLEK *et al.*, 2015; DAMINELLI *et al.*, 2015; WU *et al.*, 2016; DING *et al.*, 2016; MA *et al.*, 2016; KUANG; LIU; YU, 2016; CAIYAN; CHEN; LI, 2016; BISWAS; BISWAS, 2017).

There are several link prediction methods based on local similarity and community information available; the ones that stand out are those capable of directly incorporating the use of community information as an integral part of their calculation and not as a mere reference resource. One of these link prediction methods is the simple and accurate *Within and Inter-Community Common Neighbors* (WIC) proposed by Valverde-Rebaza and Lopes (2012a) in a previous study. The WIC method considers the influence of nodes belonging to the same or different communities to compute the likelihood of connection between pairs of disconnected nodes. Valverde-Rebaza and Lopes (2012a) also proposed a set of methods called *W-forms*, which are variants of state-of-the-art link prediction methods adapted to consider the use of community information.

Consider an undirected network G with M communities found previously by a community detection algorithm like LPA. The communities can have different sizes and are represented by labels $\{C_0, C_1, \ldots, C_{M-1}\}$, where $M > 1$. When a node $x \in V$ belongs to a community with label C, this node is represented as x^C. Here, each node belongs to a single community. For instance, in Figure 19 we show a network G with $|V| = 12$ and $|E| = 18$. Each node is labeled with its index, for easy identification. After applying a community detection algorithm, the nodes $\{v_0, v_1, v_2, v_3\}$ belong to the community with label C_α or green color. Nodes $\{v_4, v_5, v_6, v_7, v_8\}$ belong to the community with label C_β or blue color, and nodes $\{v_9, v_{10}, v_{11}\}$ belong to the community with label C_γ or yellow color. Therefore, since $v_2 \in C_\alpha$, we can represent it as $v_2^{C_\alpha}$, $v_5^{C_\beta}$, and so on.

Figure 19 – Illustration of a network in which three communities are distinguished by different colors. Nodes from the same color belong to the same community.



Source: Elaborated by the author.

Within and Inter-Community Common Neighbors Method

Proposed by Valverde-Rebaza and Lopes (2012a), *Within and Inter-Community common neighbors* (WIC) is a simple and accurate link prediction method which uses efficiently the community information. The WIC is based on Bayesian theorem, i.e., given a network G with M communities detected previously, the posterior probabilities that the same or different community labels are assigned to a pair of disconnected nodes (x,y), given its set of common neighbors $\Lambda_{x,y}$ are, respectively:

$$P(x^{C_\alpha}, y^{C_\alpha} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) P(x^{C_\alpha}, y^{C_\alpha})}{P(\Lambda_{x,y})} \tag{3.1}$$

$$P(x^{C_\alpha}, y^{C_\beta} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) P(x^{C_\alpha}, y^{C_\beta})}{P(\Lambda_{x,y})} \tag{3.2}$$

Consider that $\Lambda_{x,y} = \Lambda^W_{x,y} \cup \Lambda^I_{x,y}$, where $\Lambda^W_{x,y} = \left\{ z^C \in \Lambda_{x,y} \mid x^C, y^C \right\}$ is the set of *within-community common neighbors* (W), i.e. the set of common neighbors belonging to the same community which both x and y belong to. The complement $\Lambda^I_{x,y} = \Lambda_{x,y} - \Lambda^W_{x,y}$, is the *set of inter-community common neighbors* (I), i.e. the set of common neighbors belonging to the same community of x, or the same community of y, or any other community. Clearly, $\Lambda^W_{x,y} \cap \Lambda^I_{x,y} = \varnothing$. For instance, in Chart 3 we show some examples of how the sets $\Lambda_{x,y}$, $\Lambda^W_{x,y}$, and $\Lambda^I_{x,y}$ are related among them for three different pairs of disconnected nodes from the network showed in Figure 19.

Chart 3 – Example of the relation among the set of all common neighbors ($\Lambda_{x,y}$), the set of within-community common neighbors ($\Lambda^W_{x,y}$), and the set of inter-community common neighbors ($\Lambda^I_{x,y}$) for different pairs (x,y) of disconnected nodes from the network showed in Figure 19.

(x,y)	$\Lambda_{x,y}$	$\Lambda^W_{x,y}$	$\Lambda^I_{x,y}$
(v_2, v_3)	$\{v_0, v_1, v_{11}\}$	$\{v_0, v_1\}$	$\{v_{11}\}$
(v_3, v_5)	$\{v_4, v_{11}\}$	$\{\varnothing\}$	$\{v_4, v_{11}\}$
(v_6, v_7)	$\{v_4, v_8\}$	$\{v_4, v_8\}$	$\{\varnothing\}$

Source: Elaborated by the author.

Inspired on the rationale presented by Lopes *et al.* (2009), to estimate the probability of the set of all common neighbors of a pair of nodes $(x^{C_\alpha}, y^{C_\alpha})$ given these nodes belong to the same community with label C_α, consider the number of common neighbors with the same community label divided by the total number of common neighbors, as stated in Equation 3.3.

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) = \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda_{x,y}\right|} \tag{3.3}$$

Similarly, to estimate the probability of the set of all common neighbors of a pair of nodes $(x^{C_\alpha}, y^{C_\beta})$ given these nodes belong to different communities with labels C_α and C_β, consider

the number of common neighbors that may be associated with different labels C_α or C_β or with another community label C_γ divided by the total number of common neighbors, as defined by Equation 3.4.

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) = \frac{\left|\Lambda^I_{x,y}\right|}{\left|\Lambda_{x,y}\right|} \tag{3.4}$$

To compare the likelihood of the link existence between pair of nodes (x,y), the authors followed the scheme used by Liu *et al.* (2011). Thus, the score $s_{x,y}$ is defined as the ratio between Equations. 3.1 and 3.2. Substituting Equations 3.3 and 3.4, we have:

$$s_{x,y} = \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^I_{x,y}\right|} \times \frac{P(x^{C_\alpha}, y^{C_\alpha})}{P(x^{C_\alpha}, y^{C_\beta})} \tag{3.5}$$

The $\frac{P(x^{C_\alpha}, y^{C_\alpha})}{P(x^{C_\alpha}, y^{C_\beta})}$ ratio can be neglected since this faction is 1 if $C_\alpha = C_\beta$. When $C_\alpha \neq C_\beta$, the score will be 0 because $\Lambda^W_{x,y} = \varnothing$. Thus, the final WIC equation is:

$$s^{WIC}_{x,y} = \begin{cases} \left|\Lambda^W_{x,y}\right|, & if \;\; \Lambda^W_{x,y} = \Lambda_{x,y} \\ \\ \frac{\left|\Lambda^W_{x,y}\right|}{\left|\Lambda^I_{x,y}\right|}, & otherwise \end{cases} \tag{3.6}$$

It is important to notice that, Valverde-Rebaza and Lopes (2012b) presented the extension of WIC for applying it in directed networks. The authors defined the WIC method based on the link direction considering the sets of incoming and outgoing within and inter-community common neighbors.

Within Form Methods

As presented in previous chapter, different link prediction methods based on local structural information are based on the set of all common neighbors, except the preferential attachment (PA) method. The simple counting of the number of common neighbors indicates that each common neighbor gives the same contribution to the connection likelihood. However, is well known that different common neighbors may give different contributions to the connection probability.

Based on this consideration, Valverde-Rebaza and Lopes (2012a) consider that within-community common neighbors may contribute more to the connection likelihood than inter-community common neighbors because within-community common neighbors have similar behaviors. Therefore, the set of within-community common neighbors, $\Lambda^W_{x,y}$, is used instead the set of all common neighbors $\Lambda_{x,y}$ in the local structural methods, obtaining a set of new measures referred to as *W-form* methods.

Common Neighbors of W-form (CN-W). Based on CN method, its formal definition is according to Equation 3.7.

$$s_{x,y}^{CN-W} = \left|\Lambda_{x,y}^{W}\right|. \tag{3.7}$$

Notice that CN-W is similar to WIC when the set of within-community common neighbors of a pair of disconnected nodes is the same that the set of all its common neighbors.

Jaccard of W-form (Jac-W). Based on Jac method, its formal definition is according to Equation 3.8.

$$s_{x,y}^{Jac-W} = \frac{\left|\Lambda_{x,y}^{W}\right|}{|\Gamma(x) \cup \Gamma(y)|}. \tag{3.8}$$

Adamic-Adar of W-form (AA-W). Based on AA method, its formal definition is according to Equation 3.9.

$$s_{x,y}^{AA-W} = \sum_{z \in \Lambda_{x,y}^{W}} \frac{1}{\log|\Gamma(z)|}. \tag{3.9}$$

Resource Allocation of W-form (RAW). Based on RA method, its formal definition is according to Equation 3.10.

$$s_{x,y}^{RAW} = \sum_{z \in \Lambda_{x,y}^{W}} \frac{1}{|\Gamma(z)|}. \tag{3.10}$$

Other methods based on the set of common neighbors, such as Sor, Sal, HPI and HDI, also can be adapted to be used in the W-form.

3.1.3 Experimental Evaluation

We consider a scenario where new links of the well-known OSN, Twitter, must be predicted. On this network, the LPA is applied to assign a community label to each node. Next, using unsupervised and supervised strategies, we compare the performance of the WIC and W-form methods to state-of-the-art link prediction methods based on local-similarity information (CN, AA, Jac, RA and PA).

The main objective of this experimental evaluation is to show the relevance of use community information to improve the link prediction accuracy. Furthermore, we want to make it very clear the fact that is possible to use link prediction methods based on community information on large-scale networks from the use of appropriate methods for both community detection and even for link prediction itself.

Twitter Network

Twitter[1] is an online news and social networking service where users post and interact with short messages of up to 140 characters called *tweets*. Registered Twitter users can post

[1] <https://twitter.com/>

tweets, but those who are unregistered can only read them. Twitter differs from other social networks by its directed relationship nature, i.e., a Twitter user is not obligated to reciprocate followers by following them back, i.e. Twitter is naturally a directed network.

The Twitter network used in our experiments has follower information for 41.7 million users and 1.47 billion links. The data has been collected by Kwak *et al.* (2010) from June 6th to June 31st, 2009. Given the directionality of relationships among Twitter users, it is possible to observe that only 22.1% of the collected links are reciprocal.

In our experiments, Twitter users with more than 900 followers have been removed from the Twitter network. On this Twitter sample was employed the LPA using map-reduce formalism with 55 node Hadoop cluster. Two different executions of LPA have been performed, in one the convergence was stipulated when 7th iteration is achieved, and in the second execution when 15th iteration is achieved. Therefore, two subgraphs have been generated with vertices labeled accordingly to the communities obtained at 7th and 15th iterations, *Twitter 7it* and *Twitter 15it*, respectively. Basic properties of these graphs are summarized in Table 1.

Table 1 – Basic properties of the two graphs built from Twitter network after 7th and 15th iterations of LPA.

Properties	***Twitter 7it***	***Twitter 15it***
$\|V\|$	24,617,334	24,617,333
$\|E\|$	363,565,896	363,565,892
M	3,415,051	2,250,964
max community size	1,392,411	10,121,242
ratio of total links per user	14.77	14.77

Source: Elaborated by the author.

In Table 1, we observe that the number of nodes $|V|$ and links $|E|$ is similar for both built networks. For this reason, the ratio of total links per user $\frac{|E|}{|V|} \sim \langle k \rangle$ is the same for both networks. It is important to remark that, for our experiments we consider the directionality existing in Twitter network, therefore the counting of E and the ratio of total links per user as well as the calculation of link prediction methods consider the incoming and outgoing links. In Table 1, we also observe that *Twitter 7it* has a number of communities M higher than *Twitter 15it*. However, despite the number of communities generated, the *Twitter 15it* has a community which size is much larger than the largest community in *Twitter 7it*.

Experimental Setup

For our experiments, we perform two phases: the network pre-processing and the link prediction process. In the network pre-processing, the set E is divided into the training set E^T and the testing set E^P. From the set E, for select the links for E^P, we take randomly one-third of

the links formed by users whose number of followers is two times greater than the ratio of total links per user. The remaining links, except those formed by users whose number of followers is less than one-third of the ratio of total links per user, constitute the training set E^T.

After that, the link prediction process is initiated. This process includes both unsupervised and supervised strategies. In unsupervised strategy, for each pair of nodes from E^T, the connection likelihood is calculated based on the link direction, choosing the highest score between its *in* and *out* scores as final and single score. For instance, if for a pair of disconnected nodes (x,y), any link prediction method is computed considering only the incoming links, we obtain the score $s^{in}_{x,y}$; if only the outgoing links are considered, we obtain the score $s^{out}_{x,y}$; therefore we consider as final score $s_{x,y} = s^{in}_{x,y}$ if $s^{in}_{x,y} > s^{out}_{x,y}$, otherwise $s_{x,y} = s^{out}_{x,y}$.

In supervised strategy, we model the link prediction problem as a classification process. Therefore, we use the Weka[2] implementation of decision tree (J48), naïve Bayes (NB), support vector machine (SMO), and multilayer perceptron with backpropagation (MLP) classifiers to test the supervised strategy. For all the classifiers, we employed their standard configurations. Previously, we compute a total of $6,000,002$ feature vectors from E^T considering that each node generates a feature vector with just 30% of its neighboring nodes randomly. Thus, we have created four different data sets formed by feature vectors combining different link prediction method, i.e. for each pair of disconnected nodes we create a feature vector in which attributes correspond to the different scores computed using different link prediction methods. Each data set have 50% of instances (links) with positive class and 50% of instances with negative class[3].

Since we use different combinations of link prediction methods to generate the feature vectors, the dataset obtained are:

- *VLocal*: dataset whose feature vectors are formed by attributes corresponding to scores computed by link prediction methods based on local similarity, i.e. CN, AA, Jac, RA and PA.
- *VGroup*: dataset whose feature vectors are formed by attributes corresponding to scores computed by link prediction methods based on community information, i.e. WIC and the W-forms of CN, AA, Jac and RA.
- *VTop*: dataset whose feature vectors are formed by attributes corresponding to scores computed by the five best link prediction methods according to unsupervised results obtained in Section 3.1.3.
- *VTotal*: dataset whose feature vectors are formed by attributes corresponding to scores computed by all the link prediction methods evaluated.

[2] <http://www.cs.waikato.ac.nz/ml/weka/>

[3] As previously mentioned, since the main objective of this experimental evaluation is analyze the impact of community information in link prediction accuracy, only for this case, we do not consider the imbalance class problem in our supervised link prediction evaluation.

In Chart 4, we show in detail all the link prediction methods used to constitute the attributes of feature vectors for all the different datasets built.

Chart 4 – List of attributes that constitute all the datasets created to perform the supervised link prediction evaluation.

Dataset	Attributes
VLocal	CN, AA, Jac, RA and PA
VGroup	WIC, CN-W, AA-W, Jac-W and RAW
VTop	WIC, CN-W, AA-W, RAW and RA
VTotal	CN, AA, Jac, RA, PA, WIC, CN-W, AA-W, Jac-W and RAW

Source: Elaborated by the author.

Validating Results and Analysis

To validate our results, we use appropriate evaluation measures for both unsupervised and supervised processes.

Unsupervised Results

For results of unsupervised link prediction process, we employ AUC and Precisi@n to validate the quality of each link prediction method evaluated. Table 2 summarizes the prediction results measured by AUC computed using $n = 1000$ on *Twitter 7it* and *Twitter 15it* networks. Each AUC value is obtained by averaging over 10 implementations with 5 independently divisions of training and testing sets.

Looking at the results of Table 2, one should notice that the AUC performance of each link prediction method is the same for both subgraphs. In the case of WIC and W-form methods, this indicates that although both subgraphs have different number of communities (M), relations, interests or behaviors between nodes remain similar or equivalent, i.e. most of the users of Twitter network sharing similar interests or having similar behaviors are grouped in the same communities. That in turn and as previously commented, it is justified by the fact that results from different executions performed by LPA may produce similar partitions. For the state-of-the-art methods, the same AUC performance to both subgraphs is justified by its similar structure with the similar number of nodes ($|V|$) and links ($|E|$).

Comparing AUC performance for all link prediction methods, WIC outperforms all of them. RAW, RA, CN-W and AA-W are the next best methods, in that order. In addition, all W-form methods outperform, with significant difference, their corresponding basic forms. Also, Jac has the worst performance and do not outperform the assignment by chance.

Table 2 – Unsupervised link prediction results measured by AUC on two subgraphs of Twitter network. The emphasized values correspond to the highest results among the evaluated methods.

Method	***Twitter 7it***	***Twitter 15it***
WIC	**0.62**	**0.62**
CN	0.56	0.56
CN-W	0.59	0.59
AA	0.53	0.53
AA-W	0.58	0.58
Jac	0.45	0.45
Jac-W	0.56	0.56
RA	0.60	0.60
RAW	0.61	0.61
PA	0.51	0.51

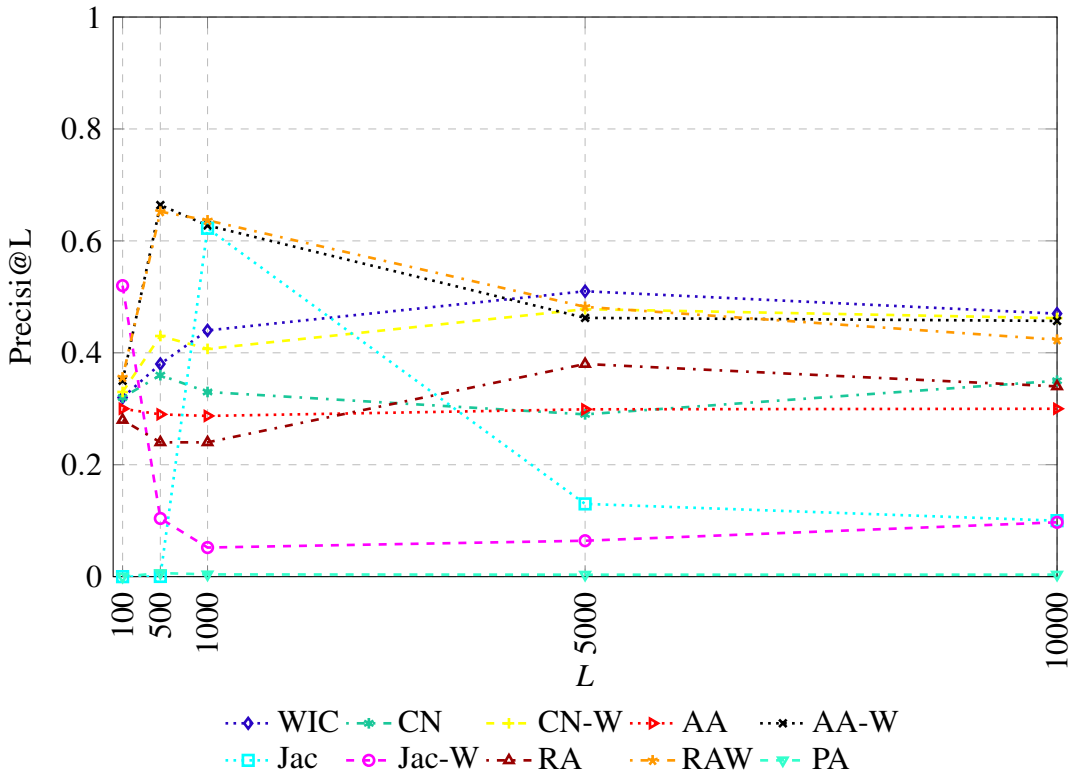
Source: Elaborated by the author.

Figure 20 shows the prediction quality measured by precisi@n on *Twitter 7it* and *Twitter 15it*. Similar to AUC, precisi@n performance of each link prediction measure evaluated is the same in both *Twitter 7it* and *Twitter 15it* subgraphs. Different values of L have been used. In the top-100 links, Jac-W obtains 0.52 and outperforms all other link prediction methods. In the top-500, AA-W and RAW obtain 0.664 and 0.652 precisi@n values, respectively, and outperform all other methods. In the top-1000, RAW, AA-W and Jac obtain 0.636, 0.627 and 0.622, respectively, and outperform all the other methods. In the top-5000, WIC obtains 0.51 precisi@n value and outperform all the other methods. In the top-10000, WIC, CN-W and AA-W obtain 0.47, 0.46 and 0.457 precisi@n values, respectively, outperforming all the other methods.

In general, AA-W and RAW obtained the best precision performance. Also, it was observed that the W-form methods obtained the best precisi@n performance that their respective basic forms, except Jac-W, which performs poorly than Jac. Also, an interesting phenomenon is observed in Jac, which has a peak when $L = 1000$ but this performance decreases sharply when $L = 5000$. PA has the worst precision for all values of L.

Since we intend to determine the relevance of use some link prediction methods in large-scale networks, we also evaluate their performance under execution time. Considering that the link prediction experiments have been performed on a computer with 99 GB of RAM and using Linux as operating system, in Figure 21 we show the average of execution time, in seconds, used by each link prediction method to generate a list of predicted links following the unsupervised strategy. Jac-W and Jac are the most time-consuming methods. Also, all W-form methods need more time than their corresponding basic form due to W-form methods have to process the community information of each node pair analyzed. However, the time spent by

Figure 20 – Precision results on the two graphs from Twitter network. Different values of L are used to select the top-L highest scores for predicting links.



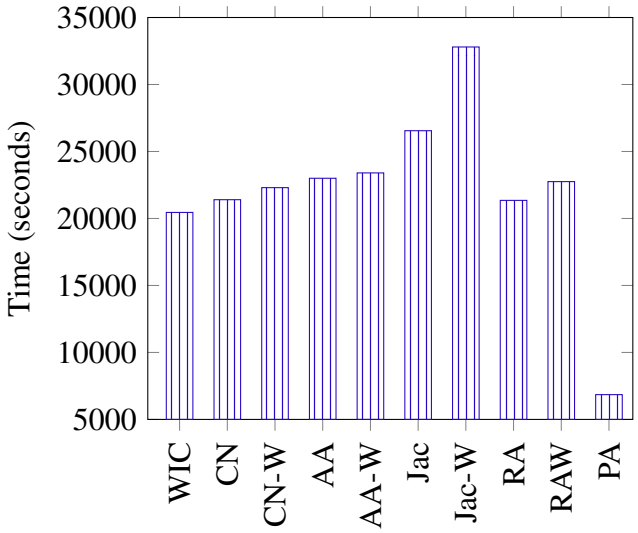
Source: Elaborated by the author.

W-form methods is not excessively higher than their respective basic forms. The fastest measure is PA due to it only makes the product of the neighbors of nodes pairs analyzed. The next fastest measures are WIC and CN, but with a significant difference with respect to PA. Therefore, we can observe that link prediction methods based on community information do not consume too much time than state-of-the-art link prediction methods, and than even can spend less time as does WIC.

Supervised Results

For results under supervised strategy, Accuracy and F-value are employed to validate the quality of the classifiers in *VLocal*, *VGroup*, *VTop* and *VTotal* datasets. Tables 3 and 4, respectively, show accuracy and F-Value average values for four different classifiers after 10-fold cross-validation. For Table 3, values in parenthesis indicate the mean absolute error. For both Tables 3 and 4, values emphasized in black correspond to the highest result among the evaluated data sets for each classifier. Entries highlighted in gray indicate that a classifier get best results in datasets which feature vectors are formed by scores obtained by methods based on community information that *VLocal* dataset, which is the dataset built using feature vectors formed only by scores obtained by methods based on local similarity.

Figure 21 – Average of execution time, in seconds, of all link prediction methods evaluated following the unsupervised strategy on Twitter network.



Source: Elaborated by the author.

Table 3 – Accuracy results (in percent) on four Twitter datasets whose feature vectors are formed by scores obtained by different link prediction methods. Values in parenthesis indicate the mean absolute error.

Dataset	J48	NB	SMO	MLP
VLocal	83.74 (0.24)	71.63 (0.28)	81.35 (0.19)	82.73 (0.25)
VGroup	82.86 (0.25)	71.70 (0.28)	80.00 (0.20)	81.88 (0.26)
VTop	83.08 (0.25)	**72.12** (**0.28**)	80.34 (0.20)	82.01 (0.26)
VTotal	**83.80** (0.24)	72.05 (0.28)	**81.71** (**0.18**)	**82.84** (**0.24**)

Source: Elaborated by the author.

Results of Table 3 show for J48, SMO and MLP classifiers the best accuracy results are obtained in *VTotal* dataset, i.e. the dataset that combines the scores of all methods based on local similarity and all methods based on community information. For NB classifier the best result is in *VTop* dataset, i.e. the dataset that uses the five best link prediction methods according to the results of Table 2. Results of Table 4 show for J48, NB, SMO and MLP classifiers the best F-measure results also are obtained in *VTotal* dataset. Furthermore, for NB classifier the best F-measure result also is obtained in *VTop* data set.

From entries highlighted in gray of Tables 3 and 4, we observe that classifiers perform better in datasets formed by feature vectors that include link prediction methods based on community information. This happens mainly when all methods based on local similarity and all methods based on community information are combined into a feature vector, i.e. in *VTotal* dataset.

Table 4 – F-measure results on four Twitter datasets whose feature vectors are formed by scores obtained by different link prediction methods.

Dataset	J48	NB	SMO	MLP
VLocal	0.837	0.698	0.812	0.827
VGroup	0.829	0.699	0.798	0.819
VTop	0.831	**0.703**	0.801	0.820
VTotal	**0.838**	**0.703**	**0.816**	**0.828**

Source: Elaborated by the author.

Discussion of Results

We use different link prediction methods on two different directed graphs built from a Twitter network. To predict a link between a pair of nodes, the WIC method considers the information on which communities the pair of nodes (and their neighbors) belong to, i.e. whether or not the nodes are in the same community. The W-form methods consider the neighborhood information of nodes belonging only to the same community of the pair of nodes analyzed.

The WIC and W-form methods require the use of community detection algorithms. Here we use the LPA, a fast community detection algorithm, in order to apply the WIC and W-form methods for large-scale networks.

With the unsupervised link prediction strategy, WIC performed better under the AUC criterion when compared to the other methods. All the W-form methods also outperform their respective basic counterpart, and are among the best-performing methods. It is worth noticing that in the link prediction analysis performed by Valverde-Rebaza and Lopes (2012a) on complex networks of different domains, PA showed the worst performance; however, for the Twitter network analyzed in our study, Jac showed the worst performance.

When analyzing precisi@n, RAW, AA-W and WIC outperform other methods. These three methods are characterized by uniform performance in different L values (especially WIC, with sustained growth up to $L = 5000$). Besides, PA performed worst for all L values.

When a supervised link prediction strategy is performed, our results show that combining methods based on local similarity information alone with methods based on community information will improve the performance of classifiers. However, this improvement may not be significant because the selection processes to generate feature vectors from datasets are diverse. Thus, selecting the most appropriate links in a supervised strategy is still a challenge.

Because our analysis used an online, large-scale social network (Twitter), execution time becomes relevant. PA is the fastest method for our unsupervised link prediction process, but it holds the penultimate and ultimate positions in AUC and precision analysis, respectively. WIC is the second fastest method, showing the best performance in AUC analysis; it is also among the

first in precision analysis. W-form methods are slightly slower than their respective basic forms, but outperform them in AUC and precision analyses.

In summary, our experiments suggest that WIC and W-form methods capture information from user behaviors in the communities they belong to, improving link prediction performance on large-scale networks. This happens because nodes in the same communities are likely to have similar interests or behaviors. In the case of Twitter, similar interests or behaviors between users may be a preference for following other users with the same topics of interest, following the same celebrities, disseminating tweets containing certain types of hashtags, etc.

3.1.4 Remarks

In this section, we have introduced the challenges in the context of link prediction problem in real-world, large-scale social networks, specifically when we try to exploit the community information available in these networks. These challenges are summarized in two basic questions: *how can we use link prediction methods based on community information without incurring in high computational cost?* and *what is the best community structure to be explored in order to reach higher link prediction accuracies?* To overcome these challenges, we have presented a strategy combining a fast and efficient community detection algorithm with fast and accurate link prediction methods based on community information. LPA finds well-defined communities with near-linear computational cost. With the communities found by LPA, WIC and W-form methods are capable of accurately predicting new links, which can be used for friendship recommendation in a time frame comparable to state-of-the-art link prediction methods.

Results obtained in the experiments presented in this section allow us to infer that the use of community information in the link prediction context may greatly contribute to accurate friendship predictions as well as to a better understanding of user behavior. Our strategy uses an appropriate community detection algorithm as the key to support the use of link prediction methods based on community information; however, the community detection process represents a computational cost that, although minimum, can be crucial in real-world applications.

3.2 Friendhsip Prediction using Social Group Information

Differently from the Web, which is largely organized around content, online social networks (OSNs) are organized around users. People can join any OSN, publish their profile and any content, create links to other users with whom they associate, and establish social groups to share specific information with other users with similar interests or behaviors (MISLOVE, 2009; KUMAR; NOVAK; TOMKINS, 2006; BRODER *et al.*, 2000). Many OSNs offer services to facilitate establishing social groups (BERNSTEIN *et al.*, 2010; AMERSHI; FOGARTY; WELD, 2012; BARTEL; DEWAN, 2013; WU *et al.*, 2015). One well known service of this type is

groups[4], launched by Facebook, which are human-assisted lists through automation, to facilitate the interaction among specific users, such as people who want to play soccer on weekends, carpool to specific cities, and to follow the tour of their favorite singers, etc.

Participation in social groups is optional, and explicitly declared by each user, i.e. a user belongs to a specific social group only by his or her own choice[5]. Users in a group do not necessarily need to be linked to each other, i.e. not all the participants in a social group are friends. These characteristics allow user groups to represent tightly clustered communities of users in the social network (MISLOVE *et al.*, 2007; MISLOVE, 2009; TRAUD; MUCHA; PORTER, 2012; ESLAMI *et al.*, 2014b). Therefore, social groups can be considered ground-truth communities (YANG; LESKOVEC, 2015; ESLAMI *et al.*, 2014a).

Since social groups exist naturally in OSNs, it is possible to use them to support different link mining tasks, such as link prediction. Using social groups instead of communities as the source of information for link prediction methods based on community information may reduce the challenges of these methods because in general they are well-defined groups of strongly connected users, and algorithms to identify social groups can be unnecessary. Computational cost is, thus, lower because it is limited to link prediction.

Given the benefits provided by social group information, we intend to explore them properly in order to improve link prediction accuracy in large-scale, real-world OSNs. Therefore, this section shows how to face the new challenges related to the use of social group information as a resource to improve link prediction accuracy. Section 3.2.1 briefly presents the main challenges related to the use of social groups to perform the link prediction process in OSNs; Section 3.2.2 shows different basic definitions and properties related to the network structure considering the presence of social groups; and Section 3.2.3 introduces some new topological network properties based on the presence of social groups. Based on these new properties, Section 3.2.4 proposes a set of new link prediction methods. Section 3.2.5 later presents experimental results obtained by comparing the efficiency of state-of-the-art link prediction methods against our proposals in a variety of large-scale OSNs. It is important to consider that the evaluation is performed using both unsupervised and supervised link prediction strategies. Section 3.2.6 concludes with final remarks.

3.2.1 Challenges in Link Prediction using Social Group Information

As previously discussed, an algorithm to find social groups is unnecessary because they are naturally existing entities. Therefore, the use of social group information instead of community information may reduce the computational cost of the link prediction process, since the only cost involved will be that of link prediction itself. In addition, because users themselves choose to participate in certain social groups, it is highly likely that these groups become well-

[4] <https://www.facebook.com/notes/facebook/sharing-with-small-groups/10150158394647131>

[5] For privacy reasons, some OSNs implement a group administrator to filter user membership requests.

defined and strongly cohesive, providing the adequate network structure for optimal performance of link prediction methods based on community/group information.

Although the use of social groups may reduce the computational cost of link prediction based on community/group information, two new challenges still stand: Since OSNs allow users to participate in multiple groups at the same time, it is highly likely that one user will belong to more than one social group. However, most (or in fact all) existing link prediction methods based on community/group information work under the assumption that a user belongs to a single community. Therefore, the first challenge is: *how can one use the information related to the fact that one user participates in one or more social groups?*

The first challenge can be easily overcome by selecting one social group among all those in which the user participates. This selection may be random or based on a given strategy, such as selecting the densest group, or the group in which most of the user's friends participate. Therefore, the information of the selected group can be used directly by an existing link prediction method based on community/group information. However, selecting one single social group may bias the link prediction process; furthermore, using one single social group when there are obviously several of them may be a misuse of a rich source of information.

The first and the second challenges share the same source: since one user can participate in one or more social groups, it is possible to observe the natural overlap between two or more social groups, i.e. one or more common users form at least two or more social groups. Therefore, the second challenge is: *how can we deal with the presence of overlapping social groups in the context of link prediction?*

To the best of our knowledge, there are no studies considering user participation in multiple social groups, or the overlapping among these groups. Therefore, in the next sections we will show how to overcome these challenges in order to perform the link prediction task efficiently using social group information in large-scale, real-world OSNs.

3.2.2 Social Group Properties on Networks

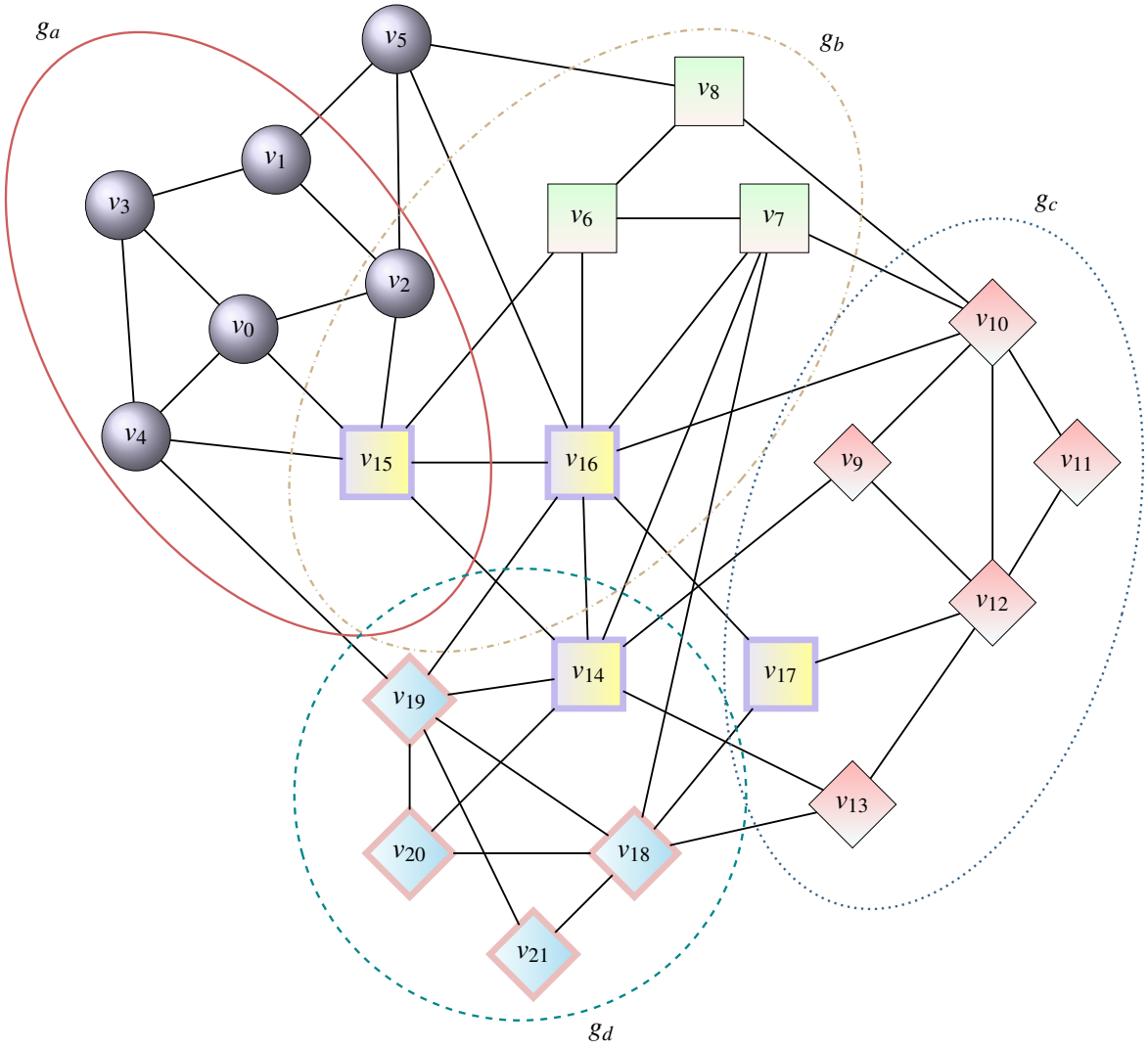
In order to ease introduce the social groups as part of network structure, we propose a simple and coherent notation to define node groups and the presence of overlapping among them.

For a network $G = (V,E)$ with $M > 1$ groups identified by different group labels $\{g_0, g_1, \ldots, g_{M-1}\}$. Each node $x \in V$ belongs to a *set of node groups* $\mathscr{G} = \{g_a, g_b, \ldots, g_p\}$ with size $|\mathscr{G}|$. Thus, $|\mathscr{G}| > 0$ and $|\mathscr{G}_\alpha| \leq M$. Each $g_i \in \mathscr{G}$ is a *group of nodes*, whose elements share interests and behaviors. When a node x belongs to one or more groups in $\mathscr{G}$, this node is represented as $x^{\mathscr{G}}$. A node belongs only to a single set of node groups.

For instance, in Figure 22 we observe an undirected network with $|V| = 22$ and $|E| = 47$. Different groups are formed by the node format and/or color as well as by an explicit grouping using an ellipse. The groups defined by node format/color are: $g_e = \{v_0, v_1, v_2, v_3, v_4, v_5\}$,

$g_f = \{v_6, v_7, v_8\}$, $g_g = \{v_9, v_{10}, v_{11}, v_{12}, v_{13}\}$, $g_h = \{v_{14}, v_{15}, v_{16}, v_{17}\}$, and $g_i = \{v_{18}, v_{19}, v_{20}, v_{21}\}$. On the other hand, the groups defined by an explicit ellipse are: $g_a = \{v_0, v_1, v_2, v_3, v_4, v_{15}\}$, $g_b = \{v_2, v_6, v_7, v_8, v_{15}, v_{16}\}$, $g_c = \{v_9, v_{10}, v_{11}, v_{12}, v_{13}, v_{17}\}$, and $g_d = \{v_{14}, v_{18}, v_{19}, v_{20}, v_{21}\}$. Therefore, we observe the presence of nine groups $\{g_a, g_b, g_c, g_d, g_e, g_f, g_g, g_h, g_i\}$, i.e. $M = 9$. Each node $x \in V$ belongs to one or more groups. For instance, the node $v_5 \in V$ belongs only to g_e group, so its set of node groups is $\mathscr{G}_\alpha = \{g_e\}$, where $|\mathscr{G}_\alpha| = 1$. The node v_8 belongs to g_b and g_f groups, so its set of node groups is $\mathscr{G}_\beta = \{g_b, g_f\}$, where $|\mathscr{G}_\beta| = 2$. The node v_{15} belongs to g_a, g_b and g_h groups, so its set of node groups is $\mathscr{G}_\gamma = \{g_a, g_b, g_h\}$, where $|\mathscr{G}_\gamma| = 3$. As we can observe, each one of the nodes v_5, v_8, and v_{15}, belong only to a single set of node groups, and so we can denote each node of these nodes as $v_5^{\mathscr{G}_\alpha}$, $v_8^{\mathscr{G}_\beta}$, and $v_{15}^{\mathscr{G}_\gamma}$, respectively.

Figure 22 – Illustration of a network in which nine groups are distinguished by different node format/color and by grouping using an ellipse. Therefore, there is possible observe that a node can belong to more than one group as well as the presence of overlapping groups.



Source: Elaborated by the author.

Considering the notation proposed, next we present definitions of common concepts

used in the literature and which are necessary for efficient analysis of network structure given the presence of groups (MISLOVE *et al.*, 2007; MISLOVE, 2009). Therefore, given a network $G = (V, E)$, we have:

Average of the number of groups to which a node belongs to. Considering that a node $x \in V$ belongs only to a single set of node groups $\mathscr{G}$, with size $|\mathscr{G}|$, the average of the number of groups to which a node belongs to is defined as the ratio between the sum of sizes of the sets of node groups of all the nodes and the total number of nodes, as stated in Equation 3.11.

$$\langle m \rangle = \frac{\sum_{x^{\mathscr{G}} \in V} |\mathscr{G}|}{|V|}. \tag{3.11}$$

The average of the number of groups to which a node belongs to indicates the number of groups, in average, to which any node is participating.

Average of groups size. The size of group g is defined by the number of users which belong to the group. Thus, the average of groups size is defined as the ratio between the sum of size of all groups and the total number of groups, as stated in Equation 3.12.

$$\langle P \rangle = \frac{\sum_{i=0}^{M-1} |g_i|}{M}. \tag{3.12}$$

The average of groups size indicates how many nodes, in average, has any group in the network.

Group Clustering Coefficient. The group clustering coefficient of a group g is defined as the average of local clustering coefficient of all the nodes $x \in V$ participating of group g. The group clustering coefficient of a group g is computed by:

$$CC^g = \frac{\sum_{x \in g} CC_x}{|g|} \tag{3.13}$$

where CC_x is the local clustering coefficient presented in Definition 2.44, and $|g|$ is the number of nodes belonging to the group g.

The group clustering coefficient indicates how well connected are the nodes participating of a specific group. The feasible values of group clustering coefficient are $[0, 1]$.

Average Group Clustering Coefficient. The average group clustering coefficient is defined as the average of the group clustering coefficient of all groups in the network. The average group clustering coefficient is computed as stated in Equation 3.14.

$$\langle g \rangle = \frac{\sum_{i=0}^{M-1} CC^{g_i}}{M}. \tag{3.14}$$

3.2.3 Extracting Efficiently Overlapping Group Information

The network properties previously presented extract efficiently different characteristics related to the participation of user in various groups at the same time. However, to the best of our knowledge there is no network property adequate to extract efficiently characteristics related to the presence of overlapping groups.

In order to use the social group information appropriately to better understand the behavior of OSNs users, we formally present new network definitions to handle both the participation of nodes in various groups at the same time as well as the presence of overlapping among such groups. Therefore, given a network $G = (V,E)$, we have:

Neighborhood of Overlapping Groups Membership. The neighborhood of overlapping groups membership of a node $x \in V$ belonging to the set of node groups $\mathcal{G}_\alpha$, i.e. $x^{\mathcal{G}_\alpha}$, is denoted by $\Gamma^{\mathcal{G}}(x)$ and formally is given by Equation 3.15.

$$\Gamma^{\mathcal{G}}(x) = \{y^{\mathcal{G}_\beta} \mid ((x^{\mathcal{G}_\alpha}, y^{\mathcal{G}_\beta}) \in E \ \vee \ (y^{\mathcal{G}_\beta}, x^{\mathcal{G}_\alpha}) \in E) \ \wedge \ \mathcal{G}_\alpha \cap \mathcal{G}_\beta \neq \varnothing\}. \tag{3.15}$$

The neighborhood of overlapping groups membership of the node x is the set formed only by its neighbors participating at least in one of groups in which also x participates.

Degree of Overlapping Groups Membership. The degree of overlapping groups membership of a node $x \in V$ is defined as the size of neighborhood of overlapping groups membership of node x, as stated in Equation 3.16.

$$k_x^{\mathcal{G}} = |\Gamma^{\mathcal{G}}(x)|. \tag{3.16}$$

Average Degree of Overlapping Groups Membership. The average degree of overlapping groups membership is defined as the average of degree of overlapping groups membership of all nodes in the network, as stated in Equation 3.17.

$$\langle k^{\mathcal{G}} \rangle = \frac{\sum_{x \in V} k_x^{\mathcal{G}}}{|V|}. \tag{3.17}$$

Clustering Coefficient of Overlapping Groups Membership. The clustering coefficient of overlapping groups membership of a node $x \in V$ is defined as the local clustering coefficient computed on the subgraph consisting only of nodes belonging to the neighborhood of overlapping groups membership of the node x. Thus, the clustering coefficient of overlapping groups membership of node x can be calculated by:

$$C_x^{\mathcal{G}} = \frac{\Delta_x^{\mathcal{G}}}{\Delta_x^{\mathcal{G}} + \Lambda_x^{\mathcal{G}}} \tag{3.18}$$

where $\Delta_x^{\mathcal{G}}$ and $\Lambda_x^{\mathcal{G}}$ are respectively the number of connected and disconnected pair of nodes whose common neighbors of groups include x. Clearly, $\Delta_x^{\mathcal{G}} + \Lambda_x^{\mathcal{G}} = \frac{k_x^{\mathcal{G}}(k_x^{\mathcal{G}}-1)}{2}$.

Average Clustering Coefficient of Overlapping Groups Membership. The average clustering coefficient of overlapping groups membership is defined as the average of clustering coefficient of overlapping groups membership of all nodes in the network, as stated in Equation 3.19.

$$C^{\mathscr{G}} = \frac{\sum_{x \in V} C_x^{\mathscr{G}}}{|V|}. \tag{3.19}$$

3.2.4 Predicting Links using Overlapping Social Group Information

As presented in previous chapters, different methods for link prediction have been proposed in the literature. Among the existing methods, some of them make use of community/group information to improve the prediction accuracy, which consider the fact that each node in the analyzed network participates only in a single community/group. However, to the best of our knowledge there is no link prediction methods considering both the participation of nodes in multiple communities/groups at the same time and the presence of overlapping among those communities/groups.

In this section, we present a set of proposals to perform the link prediction task. Our proposals are based on a Bayesian probabilistic framework to explore efficiently the overlapping group information in networks. The Bayesian probabilistic framework used (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), considers the different ways that the neighbors of a pair of disconnected nodes x and y interact in the groups in which participate. Therefore, aiming to compute the likelihood of connection of a pair of disconnected nodes, our proposals capture in simple equations the contributions offered by node's neighborhhod, groups in which all these nodes are participating, and the overlapping among these groups.

We classify our proposals in two classes: *based purely on network structure*, and *based on Naïve Bayes model*. For develop all our proposals, we start from the same prior probability. For a network G, we denote by $L_{x,y}$ and $\overline{L}_{x,y}$ the class variables of link existence and nonexistence, respectively, for a pair of nodes $(x,y) \in V$. The prior probabilities of $L_{x,y}$ and $\overline{L}_{x,y}$ are calculated according to Equations 3.20 and 3.21, respectively.

$$P(L_{x,y}) = \frac{|E|}{|U|}, \tag{3.20}$$

$$P(\overline{L}_{x,y}) = \frac{|U| - |E|}{|U|}. \tag{3.21}$$

In Equations 3.20 and 3.21, $|U|$ represents the size of universal set presented in Definition 2.35. From these prior probabilities, we instance the Bayesian probabilistic framework in different ways to obtain our different link prediction methods proposed.

Proposals based purely on Network Topology

Considering only the network definitions to handle the participation of nodes in various groups and the presence of overlapping among these groups, presented in Section 3.2.3, we propose three new link prediction methods. These proposals are called: *common neighbors within and outside of common groups* (WOCG), *common neighbors of groups* (CNG), and *common neighbors with total and partial overlapping of groups* (TPOG).

The proposals consider basically that the similarity between a pair of disconnected users is related to both different roles that each common neighbor plays and the interaction existing among all the common neighbors in all the groups in which they participate, specially in which exist the presence of overlapping.

Common Neighbors Within and Outside of Common Groups (WOCG)

For a network $G = (V, E)$, according to Bayesian theory, the posterior probabilities of link existence and nonexistence between a pair of nodes $(x^{\mathscr{G}_\alpha}, y^{\mathscr{G}_\beta})$, given its set of all common neighbors $\Lambda_{x,y}$, are defined by Equations 3.22 and 3.23, respectively.

$$P(L_{x,y}|\Lambda_{x,y}) = \frac{P(\Lambda_{x,y}|L_{x,y})P(L_{x,y})}{P(\Lambda_{x,y})}, \tag{3.22}$$

$$P(\overline{L}_{x,y}|\Lambda_{x,y}) = \frac{P(\Lambda_{x,y}|\overline{L}_{x,y})P(\overline{L}_{x,y})}{P(\Lambda_{x,y})}. \tag{3.23}$$

Considering that $\mathscr{G}_{\alpha,\beta} = \mathscr{G}_\alpha \cap \mathscr{G}_\beta$, we define the set of all common neighbors such as $\Lambda_{x,y} = \Lambda^{WCG}_{x,y} \cup \Lambda^{OCG}_{x,y}$, where $\Lambda^{WCG}_{x,y} = \{z^{\mathscr{G}_\gamma} \in \Lambda_{x,y} \mid \mathscr{G}_{\alpha,\beta} \cap \mathscr{G}_\gamma \neq \varnothing\}$ is the set of common neighbors within common groups (WCG), i.e., the common neighbors of x and y belonging to at least one group to which both x and y belong to. The complement, $\Lambda^{OCG}_{x,y} = \Lambda_{x,y} - \Lambda^{WCG}_{x,y}$ is the set of common neighbors outside of the common groups (OCG), i.e., the common neighbors of x and y belonging to any group except to one group to which both x and y belong to. Clearly, $\Lambda^{WCG}_{x,y} \cap \Lambda^{OCG}_{x,y} = \varnothing$.

In Chart 5, we show some examples of how the sets $\Lambda_{x,y}$, $\Lambda^{WCG}_{x,y}$, and $\Lambda^{OCG}_{x,y}$, are related among them for four different pairs of disconnected nodes from the network showed in Figure 22. For instance, from Chart 5 we observe that for the pair $v_2^{\mathscr{G}_\alpha}$ and $v_{16}^{\mathscr{G}_\beta}$, where $\mathscr{G}_\alpha = \{g_a, g_b, g_e\}$ and $\mathscr{G}_\beta = \{g_b \cdot g_h\}$, we have that $\mathscr{G}_{\alpha,\beta} = \mathscr{G}_\alpha \cap \mathscr{G}_\beta = \{g_b\}$. Therefore, considering the set of all common neighbors $\Lambda_{v_2,v_{16}} = \{v_5^{\mathscr{G}_\gamma}, v_{15}^{\mathscr{G}_\delta}\}$, where $\mathscr{G}_\gamma = \{g_e\}$ and $\mathscr{G}_\delta = \{g_a, g_b, g_h\}$, we have that $\Lambda^{WCG}_{v_2,v_{16}} = \{v_{15}^{\mathscr{G}_\delta}\}$ since $\mathscr{G}_{\alpha,\beta} \cap \mathscr{G}_\delta \neq \varnothing$, and $\Lambda^{OCG}_{v_2,v_{16}} = \{v_5^{\mathscr{G}_\gamma}\}$ since $\mathscr{G}_{\alpha,\beta} \cap \mathscr{G}_\gamma = \varnothing$.

Hence, to estimate the probability of the common neighbors $\Lambda_{x,y}$ given the connection between $x^{\mathscr{G}_\alpha}$ and $y^{\mathscr{G}_\beta}$, we have to consider the number of common neighbors within common groups by the number of all common neighbors, as stated in Equation 3.24.

$$P(\Lambda_{x,y} \mid L_{x,y}) = \frac{|\Lambda^{WCG}_{x,y}|}{|\Lambda_{x,y}|}. \tag{3.24}$$

Chart 5 – Example of the relation among the set of all common neighbors ($\Lambda_{x,y}$), the set of common neighbors within common groups ($\Lambda^{WCG}_{x,y}$), and the set of common neighbors outside of the common groups ($\Lambda^{OCG}_{x,y}$), for different pairs (x,y) of disconnected nodes from the network showed in Figure 22.

(x,y)	$\Lambda_{x,y}$	$\Lambda^{WCG}_{x,y}$	$\Lambda^{OCG}_{x,y}$
(v_2,v_{16})	$\{v_5,v_{15}\}$	$\{v_{15}\}$	$\{v_5\}$
(v_{15},v_{19})	$\{v_4,v_{14},v_{16}\}$	$\{\varnothing\}$	$\{v_4,v_{14},v_{16}\}$
(v_8,v_{16})	$\{v_5,v_6,v_{10}\}$	$\{v_6\}$	$\{v_5,v_{10}\}$
(v_7,v_8)	$\{v_6\}$	$\{v_6\}$	$\{\varnothing\}$

Source: Elaborated by the author.

Similarly, to estimate the probability of the common neighbors $\Lambda_{x,y}$ given a disconnection between $x^{\mathscr{G}_\alpha}$ and $y^{\mathscr{G}_\beta}$, we have to consider the number of common neighbors outside of the common groups by the number of all common neighbors, as stated in Equation 3.25.

$$P(\Lambda_{x,y} \mid \overline{L}_{x,y}) = \frac{|\Lambda^{OCG}_{x,y}|}{|\Lambda_{x,y}|}. \tag{3.25}$$

In order to compare the likelihood of link existence between $x^{\mathscr{G}_\alpha}$ and $y^{\mathscr{G}_\beta}$, in Equation 3.26, we use the same rationale implemented in WIC (see Section 3.1.2) to define the likelihood score, $s_{x,y}$, of a node pair (x,y) as the ratio between Equation 3.22 and 3.23.

$$s_{x,y} = \frac{P(\Lambda_{x,y} \mid L_{x,y})P(L_{x,y})}{P(\Lambda_{x,y} \mid \overline{L}_{x,y})P(\overline{L}_{x,y})}. \tag{3.26}$$

Substituting Equation 3.24 and 3.25, we have the final score referred to as the *common neighbors within and outside of common groups* (WOCG) method, defined as:

$$s^{WOCG}_{x,y} = \frac{|\Lambda^{WCG}_{x,y}|}{|\Lambda^{OCG}_{x,y}|} \times \Omega, \tag{3.27}$$

where $\Omega = \frac{P(L_{x,y})}{P(\overline{L}_{x,y})} = \frac{|E|}{|U|-|E|}$ is a constant for a network and its computation can be disregarded. To prevent the division by zero, we can use any smoothing method. Thus, using the add-one smoothing, the final WOCG equation is given by:

$$s^{WOCG}_{x,y} = \frac{|\Lambda^{WCG}_{x,y}|+1}{|\Lambda^{OCG}_{x,y}|+1}. \tag{3.28}$$

The WOCG method refers to the fraction between the number of common neighbors of x and y participating of at least one group in which participate both x and y at the same time, and the number of common neighbors participating of any other group, including those in which or x or y participates. Therefore, WOCG can be understood as the relation between the total overlapping groups of node's neighborhod and the simple group participation or partial overlapping of groups.

Common Neighbors of Groups (CNG)

Considering the pair of disconnected nodes $(x^{\mathcal{G}_\alpha}, y^{\mathcal{G}_\beta})$, we define the *set of common neighbors of groups* $\Lambda^{\mathcal{G}}_{x,y} = \{z^{\mathcal{G}_\gamma} \in \Lambda_{x,y} \mid \mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing \ \vee \ \mathcal{G}_\beta \cap \mathcal{G}_\gamma \neq \varnothing\}$. In Chart 6, we show some examples of how the sets $\Lambda_{x,y}$ and $\Lambda^{\mathcal{G}}_{x,y}$, are related among them for four different pairs of disconnected nodes from the network showed in Figure 22. For instance, we observe that for the pair $v_2^{\mathcal{G}_\alpha}$ and $v_{16}^{\mathcal{G}_\beta}$, where $\mathcal{G}_\alpha = \{g_a, g_b, g_e\}$ and $\mathcal{G}_\beta = \{g_b{\cdot}g_h\}$, we have the respective set of all common neighbors $\Lambda_{v_2,v_{16}} = \{v_5^{\mathcal{G}_\gamma}, v_{15}^{\mathcal{G}_\delta}\}$, where $\mathcal{G}_\gamma = \{g_e\}$ and $\mathcal{G}_\delta = \{g_a, g_b, g_h\}$. Therefore, the set of common neighbors of groups of v_2 and v_{16} is $\Lambda^{\mathcal{G}}_{v_2,v_{16}} = \{v_5^{\mathcal{G}_\gamma}, v_{15}^{\mathcal{G}_\delta}\}$, since $\mathcal{G}_\alpha \cap \mathcal{G}_\gamma \neq \varnothing$ and $\mathcal{G}_\alpha \cap \mathcal{G}_\delta \neq \varnothing$. Similarly, for the pair $v_8^{\mathcal{G}_\alpha}$ and $v_{16}^{\mathcal{G}_\beta}$, where $\mathcal{G}_\alpha = \{g_b, g_f\}$ and $\mathcal{G}_\beta = \{g_b{\cdot}g_h\}$, we can observe that the set of all common neighbors is $\Lambda_{v_8,v_{16}} = \{v_5^{\mathcal{G}_\gamma}, v_6^{\mathcal{G}_\delta}, v_{10}^{\mathcal{G}_\zeta}\}$, where $\mathcal{G}_\gamma = \{g_e\}$, $\mathcal{G}_\delta = \{g_b, g_f\}$, and $\mathcal{G}_\zeta = \{g_c, g_g\}$. Therefore, the set of common neighbors of groups of v_8 and v_{16} is $\Lambda^{\mathcal{G}}_{v_8,v_{16}} = \{v_6^{\mathcal{G}_\delta}\}$ since $\mathcal{G}_\alpha \cap \mathcal{G}_\gamma = \varnothing \wedge \mathcal{G}_\beta \cap \mathcal{G}_\gamma = \varnothing$, $\mathcal{G}_\alpha \cap \mathcal{G}_\zeta = \varnothing \wedge \mathcal{G}_\beta \cap \mathcal{G}_\zeta = \varnothing$, and only $\mathcal{G}_\alpha \cap \mathcal{G}_\delta \neq \varnothing$.

Chart 6 – Example of the relation among the set of all common neighbors ($\Lambda_{x,y}$) and the set of common neighbors of groups ($\Lambda^{\mathcal{G}}_{x,y}$), for different pairs (x,y) of disconnected nodes from the network showed in Figure 22.

(x,y)	$\Lambda_{x,y}$	$\Lambda^{\mathcal{G}}_{x,y}$
(v_2, v_{16})	$\{v_5, v_{15}\}$	$\{v_5, v_{15}\}$
(v_5, v_6)	$\{v_8, v_{16}\}$	$\{v_8, v_{16}\}$
(v_8, v_{16})	$\{v_5, v_6, v_{10}\}$	$\{v_6\}$
(v_{10}, v_{14})	$\{v_7, v_9, v_{16}\}$	$\{v_7\}$

Source: Elaborated by the author.

Based on the set of common neighbors of groups, we define a score referred to as *common neighbors of groups* (CNG), as stated in Equation 3.29.

$$s^{CNG}_{x,y} = |\Lambda^{\mathcal{G}}_{x,y}|. \tag{3.29}$$

The CNG method refers to the size of the set of common neighbors of x and y belonging to at least one group to which x or y belongs to. Therefore, the CNG method can be understood as the counting of common neighbors participating of groups in which also participate the analyzed nodes.

Common Neighbors with Total and Partial Overlapping of Groups (TPOG)

We propose a new link prediction method developing the same procedure presented to WOCG but by considering the set of common neighbors of groups instead the set of all common neighbors. Thus, according to Bayesian theory, the posterior probabilities of link existence and

nonexistence between a pair of nodes $(x^{\mathscr{G}_\alpha}, y^{\mathscr{G}_\beta})$, given its set of common neighbors of groups $\Lambda^{\mathscr{G}}_{x,y}$, are defined by Equations 3.30 and 3.31, respectively.

$$P(L_{x,y}|\Lambda^{\mathscr{G}}_{x,y}) = \frac{P(\Lambda^{\mathscr{G}}_{x,y}|L_{x,y})P(L_{x,y})}{P(\Lambda^{\mathscr{G}}_{x,y})}, \tag{3.30}$$

$$P(\overline{L}_{x,y}|\Lambda^{\mathscr{G}}_{x,y}) = \frac{P(\Lambda^{\mathscr{G}}_{x,y}|\overline{L}_{x,y})P(\overline{L}_{x,y})}{P(\Lambda^{\mathscr{G}}_{x,y})}. \tag{3.31}$$

Consider that $\Lambda^{\mathscr{G}}_{x,y} = \Lambda^{TOG}_{x,y} \cup \Lambda^{POG}_{x,y}$, where $\Lambda^{TOG}_{x,y} = \{z^{\mathscr{G}_\gamma} \in \Lambda^{\mathscr{G}}_{x,y} \mid \mathscr{G}_\alpha \cap \mathscr{G}_\gamma \neq \varnothing \wedge \mathscr{G}_\beta \cap \mathscr{G}_\gamma \neq \varnothing\}$ is the set of common neighbors with total overlapping of groups (TOG), i.e., the common neighbors of group of x and y belonging to at least one group of nodes to which x and y belong to. The complement, $\Lambda^{POG}_{x,y} = \Lambda^{\mathscr{G}}_{x,y} - \Lambda^{TOG}_{x,y}$, is the set of common neighbors with partial overlapping of groups (POG), i.e., the common neighbors of groups of x and y belonging exclusively to at least one group of nodes to which x or y belong to. Clearly, $\Lambda^{TOG}_{x,y} \cap \Lambda^{POG}_{x,y} = \varnothing$.

In Chart 7, we show some examples of how the sets $\Lambda_{x,y}$, $\Lambda^{\mathscr{G}}_{x,y}$, $\Lambda^{TOG}_{x,y}$, and $\Lambda^{POG}_{x,y}$, are related among them for four different pairs of disconnected nodes from the network showed in Figure 22. For instance, we observe that for the pair $v_2^{\mathscr{G}_\alpha}$ and $v_{16}^{\mathscr{G}_\beta}$, where $\mathscr{G}_\alpha = \{g_a, g_b, g_e\}$ and $\mathscr{G}_\beta = \{g_b.g_h\}$, we have the set of common neighbors of groups $\Lambda^{\mathscr{G}}_{v_2,v_{16}} = \{v_5^{\mathscr{G}_\gamma}, v_{15}^{\mathscr{G}_\delta}\}$, where $\mathscr{G}_\gamma = \{g_e\}$ and $\mathscr{G}_\delta = \{g_a, g_b, g_h\}$. Therefore, we have that $\Lambda^{TOG}_{v_2,v_{16}} = \{v_{15}^{\mathscr{G}_\delta}\}$ since $\mathscr{G}_\alpha \cap \mathscr{G}_\delta \neq \varnothing \wedge \mathscr{G}_\beta \cap \mathscr{G}_\delta \neq \varnothing$, and $\Lambda^{POG}_{v_2,v_{16}} = \{v_5^{\mathscr{G}_\gamma}\}$ since $\mathscr{G}_\alpha \cap \mathscr{G}_\gamma \neq \varnothing \wedge \mathscr{G}_\beta \cap \mathscr{G}_\gamma = \varnothing$.

Chart 7 – Example of the relation among the set of all common neighbors ($\Lambda_{x,y}$), set of common neighbors of groups ($\Lambda^{\mathscr{G}}_{x,y}$), set of common neighbors with total overlapping of groups ($\Lambda^{TOG}_{x,y}$), and set of common neighbors with partial overlapping of groups ($\Lambda^{POG}_{x,y}$), for different pairs (x,y) of disconnected nodes from the network showed in Figure 22.

(x,y)	$\Lambda_{x,y}$	$\Lambda^{\mathscr{G}}_{x,y}$	$\Lambda^{TOG}_{x,y}$	$\Lambda^{POG}_{x,y}$
(v_2, v_{16})	$\{v_5, v_{15}\}$	$\{v_5, v_{15}\}$	$\{v_{15}\}$	$\{v_5\}$
(v_5, v_6)	$\{v_8, v_{16}\}$	$\{v_8, v_{16}\}$	$\{\varnothing\}$	$\{v_8, v_{16}\}$
(v_8, v_{16})	$\{v_5, v_6, v_{10}\}$	$\{v_6\}$	$\{v_6\}$	$\{\varnothing\}$
(v_{10}, v_{14})	$\{v_7, v_9, v_{16}\}$	$\{v_7\}$	$\{\varnothing\}$	$\{v_7\}$

Source: Elaborated by the author.

Now, we can estimate the probability of the common neighbors of groups $\Lambda^{\mathscr{G}}_{x,y}$ given the probability of link existence and nonexistence between $x^{\mathscr{G}_\alpha}$ and $y^{\mathscr{G}_\beta}$ as stated in Equations 3.32 and 3.33, respectively.

$$P(\Lambda^{\mathscr{G}}_{x,y}|L_{x,y}) = \frac{|\Lambda^{TOG}_{x,y}|}{|\Lambda^{\mathscr{G}}_{x,y}|}, \tag{3.32}$$

$$P(\Lambda^{\mathscr{G}}_{x,y}|\overline{L}_{x,y}) = \frac{|\Lambda^{POG}_{x,y}|}{|\Lambda^{\mathscr{G}}_{x,y}|}. \tag{3.33}$$

In order to compare the likelihood of link existence between $x^{\mathcal{G}_\alpha}$ and $y^{\mathcal{G}_\beta}$, we define the likelihood score of a node pair (x,y) as the ratio between Equation 3.30 and Equation 3.31. Substituting Equation 3.32 and Equation 3.33, we have the final score called as the *common neighbors with total and partial overlapping of groups* (TPOG) method, defined as:

$$s_{x,y}^{TPOG} = \frac{|\Lambda_{x,y}^{TOG}|}{|\Lambda_{x,y}^{POG}|} \times \Omega, \tag{3.34}$$

where $\Omega = \frac{P(L_{x,y})}{P(\overline{L}_{x,y})} = \frac{|E|}{|U|-|E|}$, in the same way that for WOCG, is a constant for a network and its computation can be disregarded. To prevent the division by zero, we can use any smoothing method. Thus, using the add-one smoothing, the final WOCG equation is given by:

$$s_{x,y}^{TPOG} = \frac{|\Lambda_{x,y}^{TOG}| + 1}{|\Lambda_{x,y}^{POG}| + 1}. \tag{3.35}$$

The TPOG method refers to the same concept of WOCG but using the set of common neighbors of groups instead of the set of all common neighbors, i.e. the fraction between the number of common neighbors of x and y participating in at least one group in which participate both x and y at the same time, and the number of common neighbors participating in at least one group in which participates or x or y. Therefore, TPOG can be understood as the relation between the total and partial overlapping groups of node's neighborhood.

Proposals based on Naïve Bayes Model

The naïve Bayes model has been used by Liu *et al.* (2011) to improve the link prediction accuracy. The naïve Bayes model can be used to capture the different roles played by common neighbors and naturally assign to them different weights. The different contributions that each common neighbor offers can be used to predict accurately the link existence among pairs of disconnected nodes.

Considering that the use of naïve Bayes model improve considerably the accuracy state-of-the-art link prediction methods based on local structural similarity, we use this framework to catch both the contribution of common neighbors and their interaction in the different groups in which they are participating. Therefore, based on the naïve Bayes model, we propose a new link prediction method called *group naïve Bayes* (GNB). After that, we adapt the group naïve Bayes method for three of the most accurate state-of-the-art link prediction methods, CN, AA, and RA, obtaining the Common Neighbors, Adamic-Adar, and Resource Allocation versions of group naïve Bayes, respectively.

Group Naïve Bayes (GNB)

For a network $G = (V,E)$, consider the prior probabilities presented in Equations 3.20 and 3.21. Also, consider that each node z owns two conditional probabilities, $P(z \mid L_{x,y})$, which

is the probability that node z is the common neighbor of groups of a connected pair (x,y), and $P(z \mid \overline{L}_{x,y})$ is the probability that node z is the common neighbor of groups of a disconnected pair (x,y). According to Bayesian theory, these two probabilities are:

$$P(z \mid L_{x,y}) = \frac{P(z)P(L_{x,y} \mid z)}{P(L_{x,y})}, \tag{3.36}$$

$$P(z \mid \overline{L}_{x,y}) = \frac{P(z)P(\overline{L}_{x,y} \mid z)}{P(\overline{L}_{x,y})}. \tag{3.37}$$

The posterior probability of connection and disconnection of the pair (x,y) given its set of common neighbors of groups are, respectively:

$$P(L_{x,y} \mid \Lambda^{\mathscr{G}}_{x,y}) = \frac{P(L_{x,y})P(\Lambda^{\mathscr{G}}_{x,y} \mid L_{x,y})}{P(\Lambda^{\mathscr{G}}_{x,y})}, \tag{3.38}$$

$$P(\overline{L}_{x,y} \mid \Lambda^{\mathscr{G}}_{x,y}) = \frac{P(\overline{L}_{x,y})P(\Lambda^{\mathscr{G}}_{x,y} \mid \overline{L}_{x,y})}{P(\Lambda^{\mathscr{G}}_{x,y})}. \tag{3.39}$$

From Equations 3.38 and 3.39, we decompose $P(\Lambda^{\mathscr{G}}_{x,y} \mid L_{x,y})$ and $P(\Lambda^{\mathscr{G}}_{x,y} \mid \overline{L}_{x,y})$ as stated in Equations 3.40 and 3.41, respectively.

$$P(\Lambda^{\mathscr{G}}_{x,y} \mid L_{x,y}) = \prod_{z \in \Lambda^{\mathscr{G}}_{x,y}} P(z \mid L_{x,y}), \tag{3.40}$$

$$P(\Lambda^{\mathscr{G}}_{x,y} \mid \overline{L}_{x,y}) = \prod_{z \in \Lambda^{\mathscr{G}}_{x,y}} P(z \mid \overline{L}_{x,y}). \tag{3.41}$$

Rewriting Equations 3.38 and 3.39 using Equations 3.40 and 3.41, we obtain Equations 3.42 and 3.43, respectively.

$$P(L_{x,y} \mid \Lambda^{\mathscr{G}}_{x,y}) = \frac{P(L_{x,y})}{P(\Lambda^{\mathscr{G}}_{x,y})} \prod_{z \in \Lambda^{\mathscr{G}}_{x,y}} P(z \mid L_{x,y}), \tag{3.42}$$

$$P(\overline{L}_{x,y} \mid \Lambda^{\mathscr{G}}_{x,y}) = \frac{P(\overline{L}_{x,y})}{P(\Lambda^{\mathscr{G}}_{x,y})} \prod_{z \in \Lambda^{\mathscr{G}}_{x,y}} P(z \mid \overline{L}_{x,y}). \tag{3.43}$$

In order to compare the likelihood of the link existence between x and y, we define its likelihood score, $s_{x,y}$, as the ratio between Equations 3.42 and 3.43. Thus, substituting Eqs. 3.36 and 3.37, we have:

$$s_{x,y} = \frac{P(L_{x,y})}{P(\overline{L}_{x,y})} \prod_{z \in \Lambda^{\mathscr{G}}_{x,y}} \frac{P(\overline{L}_{x,y})P(L_{x,y} \mid z)}{P(L_{x,y})P(\overline{L}_{x,y} \mid z)}. \tag{3.44}$$

Indeed $P(L_{x,y} \mid z)$ is equal to the overlapping groups clustering coefficient of node z, as stated in Equation 3.45. Since $P(L_{x,y} \mid z) + P(\overline{L}_{x,y} \mid z) = 1$, using the Equation 3.18, $P(\overline{L}_{x,y} \mid z)$ is calculated as stated in Equation 3.46.

$$P(L_{x,y} \mid z) = C^{\mathscr{G}}_{z}, \tag{3.45}$$

$$P(\bar{L}_{x,y} \mid z) = 1 - C_z^{\mathscr{G}} = \frac{\Lambda_z^{\mathscr{G}}}{\Delta_z^{\mathscr{G}} + \Lambda_z^{\mathscr{G}}}. \tag{3.46}$$

Substituting Equations 3.20, 3.21, 3.45 and 3.46 into Equation 3.44, the likelihood score of a node pair (x,y) is given by:

$$s_{x,y} = \Omega \prod_{z \in \Lambda_{x,y}^{\mathscr{G}}} \Omega^{-1} \frac{\Delta_z^{\mathscr{G}}}{\Lambda_z^{\mathscr{G}}}, \tag{3.47}$$

where $\Omega = \frac{P(L_{x,y})}{P(\bar{L}_{x,y})} = \frac{|E|}{|U|-|E|}$ is a constant for a network and its computation can be disregarded. To prevent the division by zero, we can use any smoothing method. Thus, using the add-one smoothing, we define the *group naïve Bayes* (GNB) method as:

$$s_{x,y}^{GNB} = \prod_{z \in \Lambda_{x,y}^{\mathscr{G}}} \Omega^{-1} N_z^{\mathscr{G}}, \tag{3.48}$$

where $N_z^{\mathscr{G}} = \frac{\Delta_z^{\mathscr{G}}+1}{\Lambda_z^{\mathscr{G}}+1}$. Clearly, larger score means higher probability that two nodes are connected.

Group Naïve Bayes Forms (GNB-Forms)

As previously discussed, the connection likelihood between a pair of nodes can be improved by identifying the different roles that their common neighbors play, for example, identifying their behaviors in the different groups that they belong to. Hence, traditional link prediction methods such as CN, AA, and RA try to capture different roles from the set of all common neighbors. Similar to the adaptations performed by Liu *et al.* (2011) and Valverde-Rebaza and Lopes (2012a) of traditional local similarity methods on another mathematical frameworks, we also adapt these traditional methods to work on GNB.

Adding an exponent $f(k_x^{\mathscr{G}})$ to $\Omega^{-1} N_z^{\mathscr{G}}$ in Equation 3.48, where f is a function of overlapping groups degree. Using Log function on both sides, we obtain the next linear equation:

$$s_{x,y}^{GNB'} = \sum_{z \in \Lambda_{x,y}^{\mathscr{G}}} f(k_z^{\mathscr{G}}) \log(\Omega^{-1} N_z^{\mathscr{G}}). \tag{3.49}$$

Here we consider three forms of function f. The first function form takes the simple form of Common Neighbors method, i.e. $f(k_x^{\mathscr{G}}) = 1$. Therefore, substituting the Common Neighbors function in Equation 3.49, we define the *group naïve Bayes of Common Neighbors* (GNB-CN) as stated in Equation 3.50.

$$s_{x,y}^{GNB-CN} = |\Lambda_{x,y}^{\mathscr{G}}| \log(\Omega^{-1}) + \sum_{z \in \Lambda_{x,y}^{\mathscr{G}}} \log(N_z^{\mathscr{G}}). \tag{3.50}$$

The second function takes the form of Adamic-Adar method, i.e. $f(k_x^{\mathscr{G}}) = \frac{1}{\log(k_x^{\mathscr{G}})}$. Therefore, substituting the Adamic-Adar function in Equation 3.49, we define the *group naïve Bayes*

of Adamic-Adar (GNB-AA) as stated in Equation 3.51.

$$s_{x,y}^{GNB-AA} = \sum_{z \in \Lambda_{x,y}^{\mathcal{G}}} \frac{1}{\log(k_z^{\mathcal{G}})} (\log(N_z^{\mathcal{G}}) + \log(\Omega^{-1})). \tag{3.51}$$

Finally, the third function takes the form of Resource Allocation method, i.e. $f(k_x^{\mathcal{G}}) = \frac{1}{k_x^{\mathcal{G}}}$. Therefore, substituting the Resource Allocation function in Equation 3.49, we define the *group naïve Bayes of Resource Allocation* (GNB-RA) as stated in Equation 3.52.

$$s_{x,y}^{GNB-RA} = \sum_{z \in \Lambda_{x,y}^{\mathcal{G}}} \frac{1}{k_z^{\mathcal{G}}} (\log(N_z^{\mathcal{G}}) + \log(\Omega^{-1})). \tag{3.52}$$

Since GNB-CN is constituted by using a constant function, GNB-CN is technically similar to GNB. However, we will consider both methods as different methods due to they have different representations, which may lead to compute different operations and therefore get different results.

3.2.5 Experimental Evaluation

We consider a scenario where the prediction of new links of different OSNs is mandatory. We compared the performance of all our proposals to the performance of different link prediction methods based on local structural similarity for these networks using unsupervised and supervised strategies.

Network Datasets

The OSNs considered in our experiments were Flickr[6], LiveJournal[7], Orkut[8] and Youtube[9]. We used anonymized datasets of these OSNs, which were previously crawled by Mislove *et al.* (2007) and made publicly available[10]. These networks provide information on links between users as well as natural information on social groups to which each user belongs.

Flickr. Flickr is a photo-sharing website based on a social network. This study uses data from a crawl conducted on January 9, 2007. The Flickr dataset contains over 1.8 million users, 22 million links, and 100 thousand social groups.

LiveJournal. LiveJournal is a popular blogging site whose users form a social network. This study uses data from a crawl conducted between December 9 and December 11, 2006. The LiveJournal dataset contains over 5.2 million users, 77 million links, and 7 million social groups.

6 <https://www.flickr.com/>
7 <http://www.livejournal.com/>
8 <http://www.orkut.com>
9 <https://www.youtube.com/>
10 <http://socialnetworks.mpi-sws.org/datasets.html>

Orkut. Orkut is a social networking site run by Google. Orkut is a "pure" social network, as its sole purpose is social networking. Orkut was closed on September 30, 2014; this study uses data from a crawl conducted between October 3 and November 11, 2006. The Orkut dataset contains over 3 million users, 223.5 million links, and 8.7 million social groups.

Youtube. Youtube is a popular video-sharing site that includes a social network. This study uses data from a crawl conducted on January 15th, 2007. The dataset consists of over 1.1 million users, 4.9 million links, and 30 thousand social groups.

Table 5 shows the topological properties of the four datasets. These networks are considered large-scale due to their high number of nodes ($|V|$) and links ($|E|$). The average graph degree ($\langle k \rangle$) indicates that the average number of neighbors per user is very high in Orkut, perhaps because it is a pure social network and; therefore, friendship is a key factor. Given the presence of directionality of relationships among users of Flickr, LiveJournal, and Youtube, the fraction of symmetric links ($\mathscr{S}$) denotes the degree in which directed links from a source to a destination have an endorsement from the destination to the source. Because of its undirected nature, only Orkut shows symmetry of 100%.

Table 5 – Topological properties of OSNs analyzed.

Properties	**Flickr**	**LiveJournal**	**Orkut**	**Youtube**
$\|V\|$	1,846,198	5,284,457	3,072,441	1,157,827
$\|E\|$	22,613,981	77,402,652	223,534,301	4,945,382
$\langle k \rangle$	12.24	16.97	106.1	4.29
$\mathscr{S}$	62.0%	73.5%	100.0%	79.1%
$\langle \ell \rangle$	5.67	5.88	4.25	5.10
T	27	20	9	21
CC	0.313	0.330	0.171	0.136
r	0.202	0.179	0.072	−0.033
M	103,648	7,489,073	8,730,859	30,087
$\langle m \rangle$	4.62	21.25	106.44	0.25
$\langle P \rangle$	82	15	37	10
$\langle g \rangle$	0.47	0.81	0.52	0.34
$\langle k^{\mathscr{G}} \rangle$	9.65	6.19	50.85	0.42
$C^{\mathscr{G}}$	0.06	0.13	0.18	0.02

Source: Elaborated by the author.

Table 5 shows some general topological properties of the networks. The average path length ($\langle \ell \rangle$) indicates that the shortest path between two users of any of the OSNs analyzed in

this study averages between 4 and 6, i.e. there are a path with less than 6 users between any pair of users. However, diameter (T) shows that the greatest distance between a pair of users ranges between 20 and 27 for all the social networks analyzed but Orkut, whose users show a maximum distance of 9. The global clustering coefficient, also called average clustering coefficient (CC), indicates the connections between users in Flickr and LiveJournal tend to be stronger than in Orkut and Youtube. The assortativity coefficient (r) indicates that with the exception of Youtube, all the analyzed networks are assortative. In fact, despite its intrinsic social network nature, Orkut is less assortative than Flickr and LiveJournal. On the other hand, Youtube is slightly disassortative.

Table 5 also shows that the four networks include a high number of groups (M) and that Youtube users have less participation in social groups. We can better understand this finding by analyzing the average of the number of groups to which users belong to ($\langle m \rangle$). In Orkut, the network with the highest number of groups, any given user participates in an average of 106 social groups; in Youtube, the network with the fewest groups, it is very likely that users will not participate in any social group. The average of group size ($\langle P \rangle$) indicates that, on average, any Orkut or Flickr group has a considerable number of members, as opposed to Youtube or Flickr, whose groups characteristically have few members. The average group clustering coefficient ($\langle g \rangle$) of all the networks analyzed is substantial, i.e. users participating in existing groups are considerably well connected. The average degree of overlapping groups membership ($\langle k^{\mathcal{G}} \rangle$) shows that on average the users of all analyzed networks have many friends participating in the same groups; this is especially true for Orkut, but was not observed in Youtube. That contributes to the fact that the average clustering coefficient of overlapping groups membership ($C^{\mathcal{G}}$) is greater for Orkut and considerably lower for Youtube.

Experimental Setup

Our experiments were divided into two phases: the network pre-processing and the link prediction process. In the network pre-processing, for a network $G = (V, E)$, the set E is divided into the training set E^T and the probe set E^P. To select the links for E^P from the set E, we randomly took two-thirds of the links formed by nodes whose number of neighbors was twice as great as the average degree value. The remaining links constitute the training set E^T, excluding links formed by nodes whose number of neighbors was less than two-thirds of the average degree value.

The link prediction process included both unsupervised and supervised strategies. In the unsupervised strategy, for each pair of disconnected nodes from E^T, the connection likelihood was calculated based on the link direction, choosing the highest score between its *in* and *out* scores as a final and single score as explained in Section 3.1.3.

We evaluated the prediction performance of all our proposals based on overlapping group information, i.e. WOCG, CNG, TPOG, GNB, GNB-CN, GNB-AA, and GNB-RA. We performed

a comparative analysis of our seven proposals against five state-of-the-art link prediction methods based on local structural similarity (CN, AA, Jac, RA, and PA) and the four local similarity methods based on the Naïve Bayes model proposed by Liu *et al.* (2011) (LNB, LNB-CN, LNB-AA, and LNB-RA). The definitions of local similarity methods have been presented in Section 2.2.3.1.

In supervised strategy, we used decision tree (J48), naïve Bayes (NB), multilayer perceptron with backpropagation (MLP) and support vector machine (SMO) classifiers from Weka. For all the classifiers, we employed their standard configurations. Furthermore, for each network, we computed a set of feature vectors formed by randomly selected pair of nodes from E^T. If the pair of nodes taken from the predicted links list is in both E^T and E^P, then the feature vector formed by this pair of nodes takes the positive class (existent link), otherwise takes the negative class (nonexistent link). Table 6 shows the number of instances by class and the total of instances randomly selected for each social network. Note that we considered an imbalanced class distribution.

Table 6 – Number of instances by class for the datasets created from each analyzed network.

Network	**Existent**	**Non-existent**	**Total**
Flickr	7,100	35,500	42,600
LiveJournal	4,500	22,500	27,000
Orkut	16,000	80,000	96,000
Youtube	2,700	13,500	16,200

Source: Elaborated by the author.

For each network, we created ten different datasets. Each data set was formed by features represented by scores computed by the link prediction methods evaluated. We performed different combinations of link prediction methods to create each one of ten datasets:

- *VLocal*: dataset whose feature vectors were formed by attributes corresponding to scores computed by state-of-the-art link prediction methods based on local similarity, i.e. CN, AA, Jac, RA, and PA.

- *VLNB*: dataset whose feature vectors were formed by attributes corresponding to scores computed by local similarity link prediction methods based on Naïve Bayes model, i.e. LNB, LNB-CN, LNB-AA, and LNB-RA.

- *VGroups*: dataset whose feature vectors were formed by attributes corresponding to scores computed by our proposals based on overlapping social group information, but using purely the network topology, i.e. WOCG, CNG, and TPOG.

- *VGNB*: dataset whose feature vectors were formed by attributes corresponding to scores computed by our proposals based on overlapping social group information using the Naïve Bayes model, i.e. GNB, GNB-CN, GNB-AA, and GNB-RA.

- *VLocal-Groups*: dataset whose feature vectors were formed by attributes corresponding to scores computed by the five state-of-the-art link prediction methods based on local similarity and our three proposals based on overlapping social group information using purely the network topology, i.e. CN, AA, Jac, RA, PA, WOCG, CNG, and TPOG.

- *VLocal-GNB*: dataset whose feature vectors were formed by attributes corresponding to scores computed by the five state-of-the-art link prediction methods based on local similarity and our four proposals based on overlapping social group information using the Naïve Bayes model, i.e. CN, AA, Jac, RA, PA, GNB, GNB-CN, GNB-AA, and GNB-RA.

- *VLNB-Groups*: dataset whose feature vectors were formed by attributes corresponding to scores computed by the four local similarity link prediction methods based on Naïve Bayes model and our three proposals based on overlapping social group information using purely the network topology, i.e. LNB, LNB-CN, LNB-AA, LNB-RA, WOCG, CNG, and TPOG.

- *VLNB-GNB*: dataset whose feature vectors were formed by attributes corresponding to scores computed by the four local similarity link prediction methods based on Naïve Bayes model and our four proposals based on overlapping social group information using the Naïve Bayes model, i.e. LNB, LNB-CN, LNB-AA, LNB-RA, GNB, GNB-CN, GNB-AA, and GNB-RA.

- *VGroups-GNB*: dataset whose feature vectors were formed by attributes corresponding to scores computed by all our seven proposals based on overlapping social group information, i.e. WOCG, CNG, TPOG, GNB, GNB-CN, GNB-AA, and GNB-RA.

- *VTotal*: dataset whose feature vectors are formed by attributes corresponding to scores computed by all the sixteen link prediction methods evaluated, i.e. CN, AA, Jac, RA, PA, LNB, LNB-CN, LNB-AA, LNB-RA, WOCG, CNG, TPOG, GNB, GNB-CN, GNB-AA, and GNB-RA.

In Chart 8, we show in detail all the link prediction methods used to constitute the attributes of feature vectors for all the different datasets built.

Validating Results and Analysis

To validate our results, we used appropriate evaluation measures for both unsupervised and supervised processes.

Chart 8 – Features constituting the datasets created for each analyzed network.

Dataset	**Features**
VLocal	CN, AA, Jac, RA, and PA
VLNB	LNB, LNB-CN, LNB-AA, and LNB-RA
VGroups	WOCG, CNG, and TPOG
VGNB	GNB, GNB-CN, GNB-AA, and GNB-RA
VLocal-Groups	*VLocal* and *VGroups*
VLocal-GNB	*VLocal* and *VGNB*
VLNB-Groups	*VLNB* and *VGroups*
VLNB-GNB	*VLNB* and *VGNB*
VGroups-GNB	*VGroups* and *VGNB*
VTotal	*VLocal*, *VGroups*, *VLNB* and *VGNB*

Source: Elaborated by the author.

Unsupervised Results

For results of unsupervised link prediction process, we employed AUC and precisi@L to validate the quality of each link prediction method evaluated. Table 7 summarizes the prediction results measured by AUC, with $n = 5000$. Each AUC value was obtained by averaging over 10 run over 10 independent partitions of training and testing sets. For each network, values highlighted in gray indicate the highest result for each type of link prediction method evaluated, i.e. state-of-the-art methods, local similarity methods based on Naïve Bayes model, our proposals based purely on network topology, and our proposals based on Naïve Bayes model. Values emphasized in bold correspond to the highest AUC achieved for each network analyzed. The last column shows the average performance ranking of each link prediction method. The average ranking is the average of rank positions of each method in all the networks evaluated.

Considering the best AUC for each network, for Flickr, LiveJournal and Orkut, the local similarity methods based on local naïve Bayes model outperform the others. For Youtube network, AA performs better. However, since one method can perform better for a network and worse for another, we analyze the average performance ranking to have a better idea on the general performance of each link prediction method. Therefore, considering the four types of link prediction methods evaluated, among the state-of-the-art methods, AA performs better. Among the local similarity methods based on naïve Bayes model, LNB-RA performs better. Among our proposals based purely on network topology, TPOG performs better. Among our proposals based on naïve Bayes model, GNB-CN and GNB-AA perform better.

Based on results of Table 7, Friedman and Nemenyi post-hoc tests were applied to analyze the difference between all the link prediction methods evaluated (DEMSAR, 2006). The

Table 7 – Unsupervised link prediction results measured by AUC on the four OSNs analyzed. For each network, values emphasized in bold correspond to the highest results among all the evaluated methods. Similarly, values highlighted in Gray indicate the highest result for each subgroup of evaluated methods.

Method	Flickr	Livejournal	Orkut	Youtube	Avg. rank
CN	0.674	0.582	0.572	0.834	10.50
AA	0.656	0.580	0.620	**0.928**	8.25
Jac	0.431	0.624	0.575	0.217	12.50
RA	0.616	0.565	0.566	0.892	11.00
PA	0.566	0.542	0.602	0.917	10.00
LNB	0.860	0.880	0.446	0.872	7.25
LNB-CN	0.859	0.877	**0.706**	0.873	4.50
LNB-AA	0.884	**0.883**	0.342	0.890	5.75
LNB-RA	**0.890**	0.880	0.333	0.896	5.75
WOCG	0.637	0.596	0.649	0.434	10.75
CNG	0.728	0.611	0.621	0.723	9.63
TPOG	0.728	0.665	0.651	0.555	8.63
GNB	0.857	0.853	0.525	0.800	10.0
GNB-CN	0.861	0.855	0.639	0.808	6.25
GNB-AA	0.875	0.862	0.572	0.807	6.75
GNB-RA	0.874	0.856	0.539	0.790	8.50

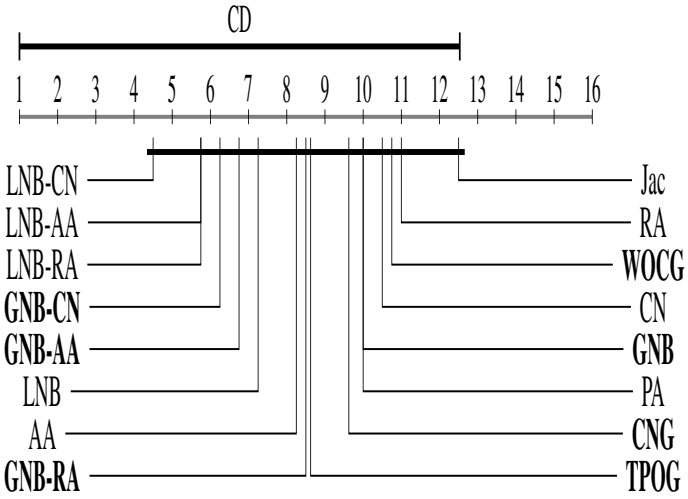
Source: Elaborated by the author.

F-statistics with 15 and 45 degrees of freedom and at 95 percentile is 1.89. According to the Friedman test using the F-statistics, the null-hypothesis that all link prediction methods evaluated behave similarly, should not be rejected. So, there is no significant difference.

Figure 23 presents the Nemenyi test for all the sixteen link prediction methods evaluated. The critical difference (CD) value for comparing the average ranking of two different link prediction methods at 95 percentile is 11.53. On the top of the presented diagram is the CD value and in the axis are the average rank of methods. The lowest (best) ranks are in the left side of the axis. All the analyzed methods have no significant difference, so they are connected by a bold line in the diagram. Our proposals are highlighted in bold in the diagram.

Although there is no significant difference among the link prediction methods, our proposals achieved a competitive accuracy. Therefore, in general terms, the methods based on naïve Bayes model using only local structural information and overlapping group information surpass the other, being LNB-CN, LNB-AA, LNB-RA, and our proposals GNB-CN and GNB-AA, the top five methods in the ranking. Following them, some of our proposals, such as

Figure 23 – Nemenyi post-hoc test diagram obtained from AUC results showed in Table 7. Diagram shows all the link prediction methods evaluated in their respective average rank position. Our proposals are highlighted in bold.



Source: Elaborated by the author.

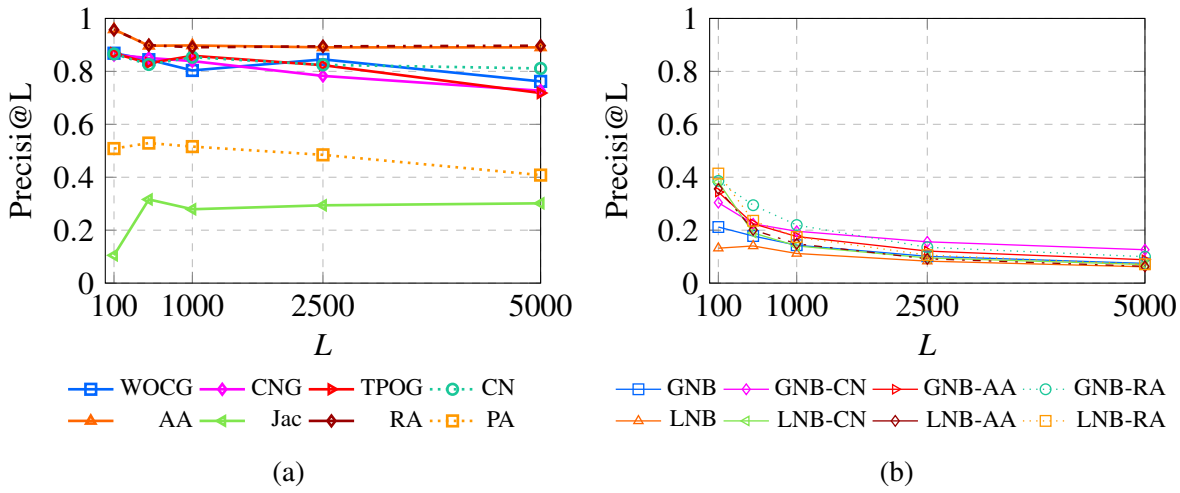
GNB-RA, TPOG, and GNB have similar performance or overcome state-of-the-art methods, such as PA, CN, RA, and Jac. Our proposal with poorest performance is WOCG, which even so overcome two state-of-the-art methods, RA and Jac.

In Figure 23, we can observe that all methods based on Naïve Bayes model, i.e. those based on local similarity as well as our proposals using overlapping social group information, perform similar and occupy the first positions. Similarly, but behind the previously mentioned, we observe the same behavior between state-of-the-art methods based on local similarity and our proposals using overlapping social group information based purely on network topology. Therefore, to facilitate the analysis, in Figures 24, 25, 26 and 27, we show the precisi@L results obtained for Flickr, LiveJournal, Orkut, and Youtube, respectively. Different values of L are used. In these figures we compare our proposals based purely on network topology to state-of-the-art methods, and or proposals based on Naïve Bayes model against local similarity methods also based on Naïve Bayes model.

In Figure 24 we observe the precisi@L results for Flickr network. In Figure 24a we observe the precisi@L performance of state-of-the-art methods based on local similarity and our proposals using social group information based purely on network topology. In Figure 24b we observe the precisi@L performance of local similarity methods based on Naïve Bayes model and our proposals using social group information based also on Naïve Bayes model. In Figure 24a, we observe that all link prediction methods, except PA and Jac, have a similar performance, highlighting AA and RA as the best overall methods in all L values, but reaching their maximum performance when $L = 100$. In Figure 24b, we observe that all link prediction methods also have a similar performance, with maximum precision value equal to 0.4 when $L = 100$. There, we

observe that, in general, state-of-the-art methods based on local similarity and our proposals based purely on network topology outperform any method based on Naïve Bayes model.

Figure 24 – Precisi@L results of Flickr network. Different values of L are used to select the top-L highest scores for predicting links obtained by different link prediction methods evaluated: (a) for state-of-the-art methods based on local similarity and our proposals using social group information based purely on network topology, and (b) for local similarity methods based on Naïve Bayes model and our proposals using social group information based on Naïve Bayes model.

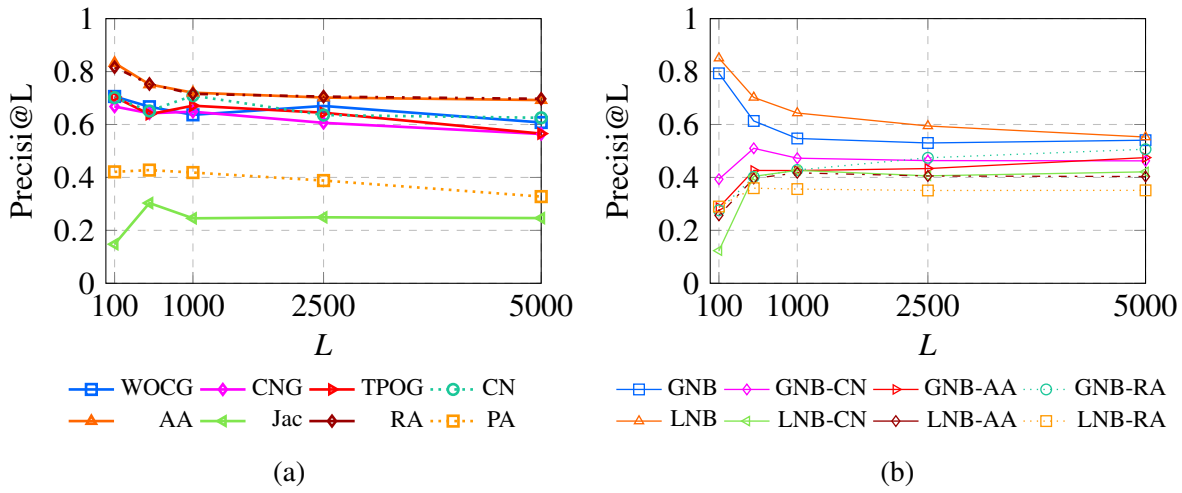


Source: Elaborated by the author.

In Figure 25 we observe the precisi@L results for LiveJournal network. In Figure 25a we observe the precisi@L performance of state-of-the-art methods based on local similarity and our proposals using social group information based purely on network topology. In Figure 25b we observe the precisi@L performance of local similarity methods based on Naïve Bayes model and our proposals using social group information based also on Naïve Bayes model. In Figure 25a, we observe that all link prediction methods, except PA and Jac, have a similar performance, highlighting AA and RA as the best overall methods in all L values, but reaching their maximum performance when $L = 100$. In Figure 25b, we also observe that evaluated methods perform similar, but LNB and GNB have the best overall measures in all L values, with maximum precisi@L values around to 0.8 when $L = 100$. These two methods are followed by GNB-CN, which perform remarkably as third, reaching a maximum precis@n value of 0.52 when $L = 500$. There we observe that, in general, most of state-of-the-art methods based on local similarity and all our proposals based purely on network topology perform consistently above 0.5 of precisi@L for any L value, while, for methods based on Naïve Bayes model, this fact is only reached by LNB and GNB.

In Figure 26 we observe the precisi@L results for Orkut network. In Figure 26a we observe the precisi@L performance of state-of-the-art methods based on local similarity and our proposals using social group information based purely on network topology. In Figure 26b we

Figure 25 – Precisi@L results of LiveJournal network. Different values of L are used to select the top-L highest scores for predicting links obtained by different link prediction methods evaluated: (a) for state-of-the-art methods based on local similarity and our proposals using social group information based purely on network topology, and (b) for local similarity methods based on Naïve Bayes model and our proposals using social group information based on Naïve Bayes model.

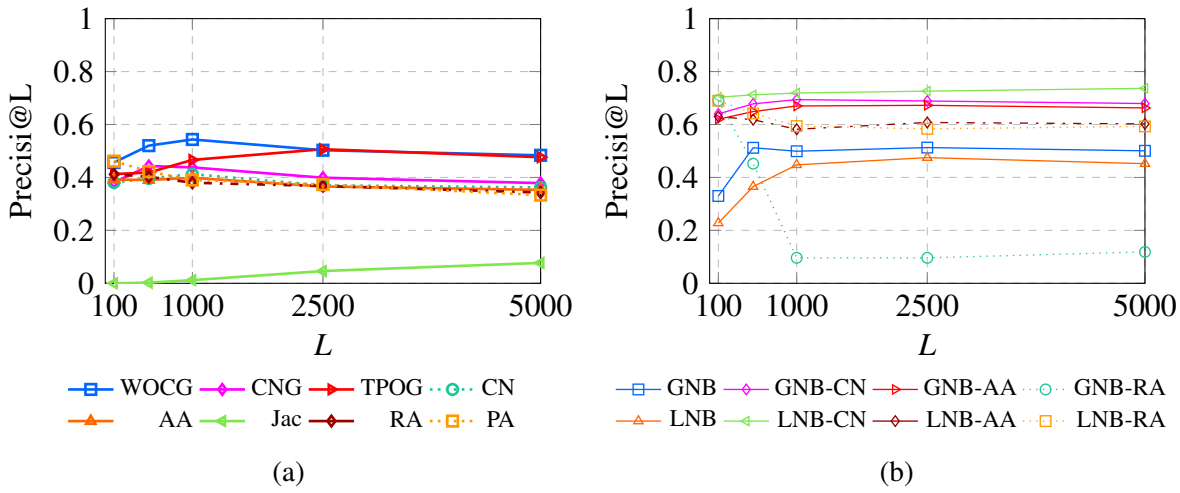


Source: Elaborated by the author.

observe the precisi@L performance of local similarity methods based on Naïve Bayes model and our proposals using social group information based also on Naïve Bayes model. From Figure 26a, we can observe that all link prediction methods but Jac have a similar performance, highlighting WOCG and TPOG as the best overall methods in all L values. WOCG achieves its maximum performance when $L = 1000$, whilst TPOG when $L = 2500$. In Figure 26b, we observe that all link prediction methods but GNB-RA have a similar performance, highlighting LNB-CN, GNB-CN and GNB-AA in all the L values and achieve the maximum precisi@L value of 0.7 when $L = 1000$. There we observe that, in general, most of methods using the Naïve Bayes model, those based on local similarity or our proposals, perform consistently on top of state-of-the-art methods based on local similarity and our proposals using overlapping group information based purely on network topology.

In Figure 27 we observe the precisi@L results for Youtube network. In Figure 27a we observe the precisi@L performance of state-of-the-art methods based on local similarity and our proposals using social group information based purely on network topology. In Figure 27b we observe the precisi@L performance of local similarity methods based on Naïve Bayes model and our proposals using social group information based also on Naïve Bayes model. In Figures 27a and 27b, we observe that all link prediction methods evaluated have similar performances. In Figure 27a, we observe that CNG and TPOG highlight as the best overall methods in all L values. Furthermore, it is important to note the considerable difference in performance that CNG has over TPOG, as well as the fact that Jac and PA still to perform poorly. From Figure 27b, we

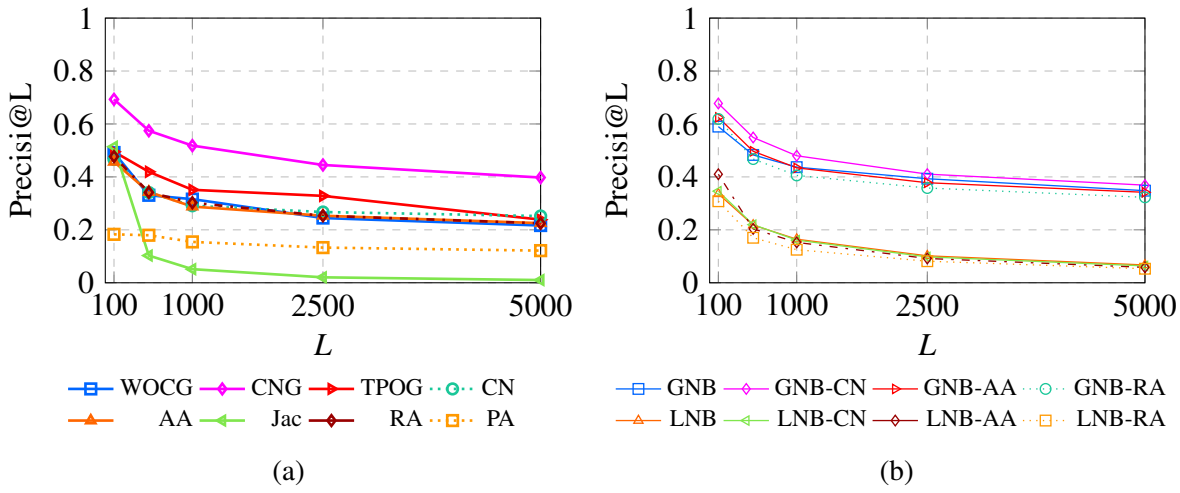
Figure 26 – Precisi@L results of Orkut network. Different values of L are used to select the top-L highest scores for predicting links obtained by different link prediction methods evaluated: (a) for state-of-the-art methods based on local similarity and our proposals using social group information based purely on network topology, and (b) for local similarity methods based on Naïve Bayes model and our proposals using social group information based on Naïve Bayes model.



(a) (b)

Source: Elaborated by the author.

Figure 27 – Precisi@L results of Youtube network. Different values of L are used to select the top-L highest scores for predicting links obtained by different link prediction methods evaluated: (a) for state-of-the-art methods based on local similarity and our proposals using social group information based purely on network topology, and (b) for local similarity methods based on Naïve Bayes model and our proposals using social group information based on Naïve Bayes model.



(a) (b)

Source: Elaborated by the author.

observe that GNB, GNB-CN, GNB-AA, and GNB-RA achieve the best precisi@L performance, with maximum value between 0.6 and 0.7 when $L = 100$. The other methods based on Naïve Bayes model perform poorly. We also observe that, in general, most of methods based on Naïve

Bayes model, specifically our proposals using overlapping social group information, perform better than all the state-of-the-art methods based on local similarity.

Supervised Results

As previously mentioned, for each analyzed network have been built ten datasets, in which their attributes correspond to the scores computed by different link prediction methods as detailed in Table 8. An imbalanced class distribution is present in ach one of these ten datasets, as showed in Table 6. Table 8 shows AUC results obtained for four different classifiers after 10-fold cross-validation over the ten datasets built for each one of OSNs analyzed. Values emphasized in bold correspond to the best results among the evaluated datasets for each classifier. Values highlighted in gray indicate that a classifier get similar or best results in datasets built using our proposals than *VLocal* and *VLNB* datasets.

In Table 8, among the datasets built using only link prediction methods based on local similarity, i.e. *VLocal* and *VLNB*, we observe that most of the classifiers perform better on *VLocal* than *VLNB*. Therefore, scores of state-of-the-art methods used as attributes of feature vectors characterize better the link structure than scores of local similarity methods based on Naïve Bayes model. Considering the datasets built using only our proposals, i.e. *VGroups* and *VGNB*, we observe that most of the classifiers perform better on *VGroups* than *VGNB*. These results suggest that, attributes of feature vectors represented by scores of methods using overlapping social group information based purely on network topology, characterize better the link structure than scores of methods using overlapping social group information based on Naïve Bayes model. The appropriate link characterization offered by *VLocal* and *VGroups* has lead to *VLocal-Groups* and *VTotal* being the datasets in which most of the classifiers perform better.

To better observe the impact of scores of link prediction methods as attributes of feature vectors characterizing the link structure of networks, in Figure 28, we show the Nemenyi post-hoc test diagrams of four network analyzed (DEMSAR, 2006). Each one of these Nemenyi diagrams has been computed from AUC results showed in Table 8. All the diagrams have the same F-statistics and Nemenyi statistics values. The critical value of the F-statistics with 9 and 27 degrees of freedom at 95 percentile is 2.25. Therefore, the null-hypothesis that all datasets have a similar contribution to classification task should be rejected. According to the Nemenyi statistics, the critical difference (CD) for comparing the mean ranking of contribution of two different datasets at 95 percentile is 6.77. The datasets with no significant difference among them are connected by a bold line in each diagram.

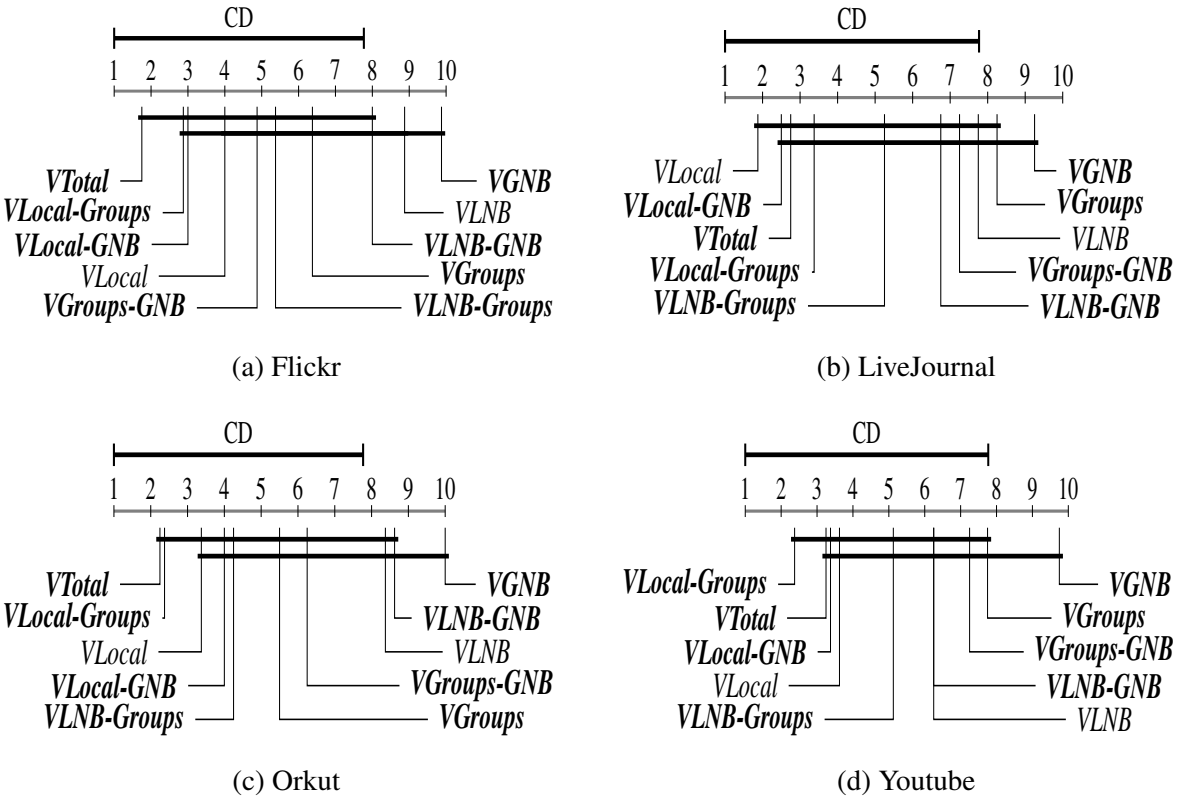
In Figure 28 we observe that *VTotal* and *VLocal-Groups* are always in the two-top in all the networks except in LiveJournal (see the Figure 28b). This fact confirms the previous observation that these two datasets offer a good link characterization. Also, we observe that in most of the cases *VLocal-Groups* and *VLocal-GNB* outperform *VLocal*, *VGroups*, and *VGNB*.

Table 8 – AUC results obtained on ten datasets built over each one of the four OSNs analyzed.

Network	Datasset	J48	NB	SMO	MLP
Flickr	*VLocal*	0.774	0.746	0.583	0.778
	VLNB	0.748	0.664	0.501	0.685
	VGroups	0.761	0.728	0.504	0.734
	VGNB	0.737	0.502	0.501	0.516
	VLocal-Groups	0.789	**0.776**	0.585	0.778
	VLocal-GNB	**0.796**	0.725	0.583	0.780
	VLNB-Groups	0.792	0.723	0.504	0.753
	VLNB-GNB	0.769	0.642	0.502	0.688
	VGroups-GNB	**0.796**	0.698	0.505	0.736
	VTotal	0.793	0.747	**0.586**	**0.782**
LiveJournal	*VLocal*	**0.808**	**0.829**	0.658	0.854
	VLNB	0.732	0.776	0.547	0.800
	VGroups	0.767	0.768	0.607	0.777
	VGNB	0.775	0.503	0.503	0.510
	VLocal-Groups	0.802	0.826	0.654	0.854
	VLocal-GNB	0.807	0.828	0.660	0.852
	VLNB-Groups	0.783	0.806	0.612	0.835
	VLNB-GNB	0.804	0.767	0.550	0.798
	VGroups-GNB	0.768	0.772	0.609	0.781
	VTotal	0.799	0.825	**0.664**	**0.858**
Orkut	*VLocal*	**0.883**	0.862	0.629	0.873
	VLNB	0.823	0.837	0.558	0.859
	VGroups	0.829	0.870	0.626	0.863
	VGNB	0.816	0.500	0.500	0.532
	VLocal-Groups	0.880	0.872	**0.644**	0.871
	VLocal-GNB	0.857	0.862	0.629	**0.876**
	VLNB-Groups	0.872	0.869	0.634	0.861
	VLNB-GNB	0.828	0.830	0.558	0.858
	VGroups-GNB	0.830	0.856	0.626	0.863
	VTotal	0.861	**0.873**	**0.644**	0.873
Youtube	*VLocal*	0.836	0.801	0.551	0.808
	VLNB	0.832	0.687	0.507	0.739
	VGroups	0.734	0.671	0.562	0.726
	VGNB	0.802	0.506	0.501	0.499
	VLocal-Groups	0.822	**0.819**	**0.579**	**0.825**
	VLocal-GNB	**0.851**	0.800	0.551	0.812
	VLNB-Groups	0.822	0.720	0.562	0.755
	VLNB-GNB	0.835	0.683	0.509	0.738
	VGroups-GNB	0.820	0.681	0.562	0.723
	VTotal	0.823	0.768	0.578	0.821

Source: Elaborated by the author.

Figure 28 – Nemenyi post-hoc test diagrams obtained from AUC results showed in Table 8 for ten datasets built from (a) Flickr, (b) LiveJournal, (c) Orkut, and (d) Youtube. Diagrams show all the datasets evaluated in their respective average rank position. Datasets built using our proposals are highlighted in bold.



(a) Flickr (b) LiveJournal

(c) Orkut (d) Youtube

Source: Elaborated by the author.

On the other hand, *VGroups-GNB* and *VLNB-Groups* have a regular performance, whilst *VLNB* and *VGNB* perform poorly in most of the cases.

Discussion of Results

We used different link prediction methods on four different OSNs. We used the social group information naturally available in the analyzed networks to compute the prediction scores of our proposals. Different evaluation measures have been used to analyze the performance of link prediction methods in both unsupervised and supervised strategies.

It is very difficult to identify one single best method using only the unsupervised link prediction strategy. Table 7 shows the AUC results, which offer some information on the performance of local similarity methods based on the Naïve Bayes model when compared to state-of-the-art methods and our proposals. Figure 23 shows the Nemenyi post-hoc test diagram to better illustrate results and confirm the superiority of local similarity methods based on Naïve Bayes model. However, the Nemenyi post-hoc test diagram also shows that our proposals using overlapping social group information based on Naïve Bayes model is very similar to the first

positions. In fact, our proposals GNB-CN and GNB-AA are among the top-5 link prediction methods.

The AUC is the most widely used evaluation measure; it captures the predictive power of a link prediction method by considering its complete list of predicted links. Although this is an important factor to evaluate a link prediction method, considering all the predicted links is not always feasible in real-life applications. In such cases, the precisi@L performance becomes crucial because it offers an outline of a specific portion of the most likely links. Figures 24, 25, 26 and 27 show the precisi@L results for Flickr, LiveJournal, Orkut, and Youtube, respectively. We observe that, in general, the precis@n performance of the local similarity methods based on the Naïve Bayes model ranges between average and poor despite their good performance in AUC. Although some state-of-the-art methods such as AA and RA have good precisi@L performance, we highlight the performance of our proposals, specially of WOCG, CNG, GNB, GNB-CN, and GNB-AA.

We observe that state-of-the-art methods perform well in precisi@L, but poorly in AUC; in contrast, local similarity methods based on the Naïve Bayes model perform well in AUC, but poorly in precisi@L. The performance of our proposals is constant in both AUC and precisi@L, but those based purely on network structure perform better in precisi@L than in AUC, and those based on Naïve Bayes model perform better in AUC than in precisi@L.

It is difficult to analyze the individual performance of the link prediction methods evaluated. Therefore, we used the supervised link prediction strategy to obtain an outline of their collective performance. The AUC classification results in Table 8 show that, in general, the combination of our proposals with any other link prediction method offers a better representation of link structure, leading to a better performance of classifiers to identify potential new relationships. The Nemenyi post-hoc test diagrams in Figure 28 also illustrate that. The diagrams in Figure 28 show that the combination of all the link prediction methods evaluated (*VTotal*) offers the best representation of link structure to perform the link prediction task via classification. However, the combination of our proposals with state-of-the-art methods (*VLocal-Groups* and *VLocal-GNB*) are comparable, and equally good, as in Orkut (Figure 28c). Furthermore, *VLocal-Groups* and *VLocal-GNB* are even superior to *VTotal* in some cases, as in LiveJournal (Figure 28b) and Youtube (Figure 28d).

Some peculiarities have been observed. The combination of state-of-the-art methods (*VLocal*) is very competitive and is, in general, in the top-5 ranking. Thus, even when competing with more sophisticated methods, such as those based on the Naïve Bayes model, or methods using more information, such as our proposals using overlapping social group information, the state-of-the-art methods are a hard nut to crack. Similarly, the combination of local similarity methods based on the Naïve Bayes model with our proposals based purely on network topology (*VLNB-Groups*) has regular performance and, in general, is in the top-5 ranking. The combinations of our proposals using overlapping social group information based purely on network

topology (*VGroups*) or based on Naïve Bayes model (*VGNB*) perform poorly in general.

Our results suggest that, when analyzed individually, our proposals can overcome the state-of-the-art methods, such as CN, Jac, AA, RA, and PA, and may be as competitive as local similarity methods based on the Naïve Bayes model, such as LNB, LNB-CN, LNB-AA, and LNB-RA. When analyzed collectively, our proposals work better when combined with state-of-the-art methods. Therefore, we can assume that the use of overlapping group information may improve link prediction accuracy.

3.2.6 Remarks

In this section, we have introduced the challenges in link prediction in real-world, online social networks, specifically when we try to explore the information related to the participation of users in multiple social groups at the same time. It is known that, like in real life, users of online social network services participate in different groups because other users with similar interests are also members of such groups. This raises one question: *how can we properly use the information related to the fact that one user participates in one or more social groups?* This question poses the challenges related to the decision of selecting one single, representative group and using the information provided by this group in one of the link prediction algorithms based on community/group information previously presented (WIC and W-form methods), or using all the existing groups. Considering all the existing groups is the obvious choice; the second choice would mean wasting an important source of information. However, considering all the existing groups to which a user belongs to implies that two or more groups could share the same users, i.e. there is a natural presence of overlapping groups. This fact leads to a second challenge: *how can we deal with the presence of overlapping social groups in the context of link prediction?*

To overcome these challenges, we have introduced a simple, formal notation to manage information on the participation of a user in different groups. We have proposed new network properties to extract characteristics related to the presence of overlapping groups more efficiently. Among these properties, we highlight the *overlapping groups clustering coefficient*, which measures the degree to which nodes belonging to overlapping groups tend to cluster together. Based on these properties, we have proposed seven new link prediction methods classified in two types: *based purely on network topology* and *based on Naïve Bayes model*. Using the Bayesian theory as a support framework, our proposals based purely on network topology compute link likelihood by considering different sets of nodes formed by the different existing correlations given the presence of overlapping groups. Our proposals based purely on network topology are called *common neighbors within and outside of common groups* (WOCG), *common neighbors of groups* (CNG), and *common neighbors with total and partial overlapping of groups* (TPOG). On the other hand, using the Naïve Bayes classifier theory as a support framework, our proposals based on the Naïve Bayes model try to identify the contribution of both common neighbors of a pair of disconnected users and their interactions in the different groups in which they participate.

Our proposals based on the Naïve Bayes model are called *group naïve Bayes* (GNB), *group naïve Bayes of common neighbors* (GNB-CN), *group naïve Bayes of Adamic-Adar* (GNB-AA), and *group naïve Bayes of Resource Allocation* (GNB-RA).

We have compared the performance of our proposals against other link prediction methods, including state-of-the-art and local similarity methods which also use the Naïve Bayes model. The purpose of our experimentation was to compare the impact of using social group information in link prediction compared to the use of local structure information. We have performed our experiments using four real-world online social networks: Flickr, LiveJournal, Orkut, and Youtube. We have chosen these networks because they offer information on their users' social groups. It is worth noticing the lack of social network datasets containing information on user's social groups. The results obtained in our experimentation show the competitive performance of our proposals in both unsupervised and supervised link prediction strategies, even outperforming state-of-the-art link prediction methods. Therefore, using overlapping social group information in link prediction conveys relevant clues to better understand user's interest and behavior.

To the best of our knowledge, we have conducted the first research considering social group information to improve link prediction accuracy. Therefore, in addition to opening a new research issue in link prediction, we have successfully established the first considerations to overcome the challenges in link prediction task using overlapping group information.

3.3 Friendship Prediction using Location Information

Millions of people use different social networking services to interact with friends and meet new people. With the widespread adoption of various smart mobile devices, many of these people have integrated the use of these services into their daily practices. This fact has been well utilized by the networks LBSNs, which in addition to offer the possibility to establish new friendships, also offer to their users the possibility of share their locations with friends as well as sending messages, tips or other information related to visited places (ZHENG; ZHOU, 2011; CHORLEY; WHITAKER; ALLEN, 2015).

The main example of LBSN is Foursquare, which involves more than 50 million of users, more than 93 million of places, and more than 10 billion of check-ins[11]. Due to these service properties, users can access and share information about friends and places within their social graph. The user-location links in this network are mutually reinforced by its actors, making it possible to take advantage of geographic mobility as an additional information source of information to analyze user behavior (CHO; MYERS; LESKOVEC, 2011; WANG *et al.*, 2011; LUO *et al.*, 2013).

[11] Data reported by Foursquare (<https://foursquare.com/about>) and accessed in July 5, 2017.

As previously discussed in this thesis, one of the most used tools to understand user behavior in social networks is the link prediction task. Due to the heterogeneity of LBSNs and depending of their final objectives, the link prediction task can focus on predict social links, which leads to the traditional and widely known task called as to *friendship prediction*, or predict location links, which leads to the task called *location prediction*. In Section 2.3.3, we have discussed the link prediction task in LBSNs, focusing mainly on friendship prediction.

Despite the fact that friendship prediction to be a well known task, its application in LBSNs opens new challenges since the use of location information represents a new dimension to be taken into account. Aiming to tackle this task, in this section we will show as location information can be effectivelly used for both to improve the accuracy of friendship prediction in the context of LBSNs as well as make this prediction task more suitable for real-world applications. In this way, hence, in Section 3.3.1, we briefly show the main challenges related to friendship prediction in the context of LBSNs. In Section 3.3.2, we present our proposals to cope with the friendship prediction in LBSNs. Afterwards, in Section 3.3.3, we show the experimental results obtained by comparing friendship current prediction methods with our proposals in two well-known real-world LBSNs. We conclude in Section 3.3.4 by with some remarks on the friendship prediction task in LBSNs.

3.3.1 Challenges in Friendship Prediction using Location Information

As widely discussed in this thesis, friendship prediction is a task that consists of predicting social links in a social network. Friendship prediction in the domain of LBSNs has, at the locations level, a new information source to be explored. Therefore, several methods have been proposed to perform friendship prediction using location information in the domain of LBSNs (CRANSHAW *et al.*, 2010; XIAO *et al.*, 2010; CHO; MYERS; LESKOVEC, 2011; YU *et al.*, 2011; MENGSHOEL *et al.*, 2013; PHAM; SHAHABI; LIU, 2013; BAYRAK; POLAT, 2014; XIAO *et al.*, 2014; ZHANG; PANG, 2015; KYLASA; KOLLIAS; GRAMA, 2016; BAYRAK; POLAT, 2016).

To consider the locations as a new actor of LBSN structure constitutes an important fact for a variety of mining tasks, including friendship prediction. However, being a new type of actor, i.e. a new type of node, locations constitute a new dimension to be considered in the calculation of computational cost of friendship prediction methods. So, the main challenge to be faced by friendship prediction methods in LBSNs is referred to the *prediction space size*.

The prediction space of a link prediction method encompasses the “universe” of pairs of users with potential to establish relationships. This universe is formed by a small amount of pairs of users that actually will be connected and a huge amount of pairs of users that will never establish a connection (SCELLATO; NOULAS; MASCOLO, 2011). This extremely skewed distribution of classes of pairs of users in the prediction space impairs on the performance of friendship prediction methods. Therefore, the prediction space challenge is related to the

question: *how can not only reduce the number of wrong predicted links of one link prediction method but also increasing the number of correctly predicted links?*

In the context of LBSNs, instead of perform calculations over all the existing users and locations in a network, the prediction space size challenge can be overcome by exploiting efficiently the information given by the interaction between users and their visited locations. Therefore, any friendship prediction method in an LBSN, for a pair of disconnected users, has to explore efficiently geographic mobility and social neighborhood patterns of these users, who are not friends but who have visited the same places, to predict if they will become friends.

To the best of our knowledge there is large amount of friendship prediction methods in LBSNs using only location information as its main source to perform their predictions. However, only a little number of them using both location and social information sources.

3.3.2 Improving Friendship Prediction in LBSNs

We surveyed the link prediction methods in LBSNs described in Section 2.3.3, and observed that due to the different information sources used to make their predictions, these methods can be divided into three groups: i) based on place, which generally use the frequency or geography distance between pairs of visited places as user similarity criteria, ii) based on check-in, which commonly use frequency of check-ins or information gain calculated for specific places as user similarity criteria, and iii) based on social information, which basically use the social strength between a pair of users and their common friends. In Chart 9 we provide an overview of different information sources used by each friendship prediction method described in Section 2.3.3.

In Chart 9 we observe that currently methods, generally, use more than one information source to improve the accuracy of their predictions. However, we identify that some information sources are not combined or their combination could be improved. Given this gap, we propose eight new friendship prediction methods to better explore the different information sources identified.

Our proposals appear highlighted in bold in Chart 9 and are referred to as: *Check-in Observation* (ChO), *Check-in Allocation* (ChA), *Within and Outside of Common Places* (WOCP), *Common Neighbors of Places* (CNP), *Total and Partial Overlapping of Places* (TPOP), *Friend Allocation Within Common Places* (FAW), *Common Neighbors of Nearby Places* (CNNP), and *Nearby Distance Allocation* (NDA). Our two first proposals correspond to the category of methods based on frequency, whilst the other six correspond to the category of methods based on social strength.

Chart 9 – Summary of current friendship prediction methods for LBSNs and our proposals, as well as the information sources used to make their predictions. Our methods are in bold.

Method	**Place Information**		**Check-in Information**		**Social Information**
	Frequency	**Geographic Distance**	**Frequency**	**Information Gain**	
Co			✓		
DCo	✓		✓		
CL	✓				
JacP	✓				
LO	✓				
CLR	✓		✓		
CLC	✓		✓		
PAP	✓				
PAC			✓		
AAP	✓		✓		
MinC	✓		✓		
MinE	✓			✓	
AAE	✓			✓	
LC	✓		✓	✓	
MinD	✓	✓			
ChD	✓	✓			
ChL	✓	✓			
GeoD		✓	✓		
WGeoD		✓	✓		
HD		✓	✓		
AHD		✓	✓	✓	
TCFCC	✓		✓		✓
ChO	✓		✓		
ChA	✓		✓		
WOCP	✓				✓
CNP	✓				✓
TPOP	✓				✓
FAW	✓		✓		✓
CNNP	✓	✓			✓
NDA	✓	✓	✓	✓	✓

Source: Elaborated by the author.

Proposals Based on Frequency

We have observed that some frequency relations between place and check-in information had not been explored by existing methods. Therefore, we propose the *Check-in Observation* (ChO) and *Check-in Allocation* (ChA) methods.

Check-in Observation (ChO)

ChO considers the ratio of the sum of the number of check-ins made at common locations of users x and y to the total sum of the number of check-ins made at all location visited by both users. Thus, *Check-in Observation* is computed as:

$$s_{x,y}^{ChO} = \frac{\sum_{\ell \in \Phi_{\mathscr{L}}(x,y)} |\Phi(x,\ell)| + |\Phi(y,\ell)|}{\sum_{\ell' \in \Phi_{\mathscr{L}}(x)} |\Phi(x,\ell')| \; + \; \sum_{\ell'' \in \Phi_{\mathscr{L}}(y)} |\Phi(y,\ell'')|}. \tag{3.53}$$

Intuitively, ChO try to catch the frequency of visits at common locations of a pair of disconnected users considering the total number of visits at all their locations.

Check-in Allocation (ChA)

ChA is based on traditional Resource Allocation method, but refining the popularity of all common locations of users x and y through the count of total check-ins of each one of such locations. Thus, *Check-in Allocation* is defined as:

$$s_{x,y}^{ChA} = \sum_{\ell \in \Phi_{\mathscr{L}}(x,y)} \frac{1}{|\Phi(\ell)|}. \tag{3.54}$$

ChA punishes heavily high number of check-ins of popular locations (e.g. public venues) by not applying a logarithmic function on the size of set of all check-ins made at such locations. Similarly to ChO, ChA is considered as a method based on frequency of places and check-ins.

Proposals Based on Social Strength

We have observed that most of existing methods in the literature disregard the social strength to compute the likelihood of existence of a relationship between a pair of users of LBSNs. Therefore, our efforts have been directed to explore different social interactions among users as well as between users and their visited locations to formulate our proposals.

Three of our proposals to perform link prediction in LBSNs using social strength take advantage of mathematical support of WOCP, CNP, and TPOP methods (which are other of our proposals previously presented in Section 3.2.4) by establishing a parallel between users participating in social groups and users visiting locations. These proposals are the *Within and Outside of Common Places* (WOCP), *Common Neighbors of Places* (CNP), and *Total and Partial Overlapping of Places* (TPOP) methods.

The other three of our proposals to perform link prediction in LBSNs using social strength have been formulated inspired by different existing link prediction methods but considering other types of information sources that have not been used previously. These proposals are *Friend Allocation Within Common Places* (FAW), *Common Neighbors of Nearby Places* (CNNP), and *Nearby Distance Allocation* (NDA) methods.

Within and Outside of Common Places (WOCP)

WOCP considers that users x and y are more likely to establish a friendship if they have more common friends visiting the same places than if they have more common friends visiting distinct places. Let the *set of common neighbors within common visited places*, $\Lambda_{x,y}^{WCP} = \{z \in \Lambda_{x,y} \mid \Phi_{\mathscr{L}}(x,y) \cap \Phi_{\mathscr{L}}(z) \neq \varnothing\}$, and the *set of common neighbors outside common visited places*, $\Lambda_{x,y}^{OCP} = \Lambda_{x,y} - \Lambda_{x,y}^{WCP}$, the WOCP method is calculated as:

$$s_{x,y}^{WOCP} = \frac{|\Lambda_{x,y}^{WCP}|}{|\Lambda_{x,y}^{OCP}|}. \tag{3.55}$$

WOCP is inspired in WOCG, which is other of our proposals to perform link prediction using overlapping social group information and defined in Section 3.2.4.

Common Neighbors of Places (CNP)

CNP indicates that a pair of users x and y more likely will have a future friendship if they have many common friends visiting the same places visited by x or y. Let the *set of common neighbors of places* of users x and y, $\Lambda_{x,y}^{\mathscr{L}} = \{z \in \Lambda_{x,y} \mid \Phi_{\mathscr{L}}(x) \cap \Phi_{\mathscr{L}}(z) \neq \varnothing \vee \Phi_{\mathscr{L}}(y) \cap \Phi_{\mathscr{L}}(z) \neq \varnothing\}$, the CNP method is defined as:

$$s_{x,y}^{CNP} = |\Lambda_{x,y}^{\mathscr{L}}|. \tag{3.56}$$

CNP is inspired in CNG, which is other of our proposals to perform link prediction using overlapping social group information and defined in Section 3.2.4.

Total and Partial Overlapping of Places (TPOP)

TPOP considers that a pair of users x and y could establish a friendship if they have more common friends visiting places also visited by both users than common friends who visited places visited by only one of them. Therefore, let the *set of common neighbors with total overlapping of places*, $\Lambda_{x,y}^{TOP} = \{z \in \Lambda_{x,y}^{\mathscr{L}} \mid \Phi_{\mathscr{L}}(x) \cap \Phi_{\mathscr{L}}(z) \neq \varnothing \wedge \Phi_{\mathscr{L}}(y) \cap \Phi_{\mathscr{L}}(z) \neq \varnothing\}$, and the *set of common neighbors with partial overlapping of places*, $\Lambda_{x,y}^{POP} = \Lambda_{x,y}^{\mathscr{L}} - \Lambda_{x,y}^{TOP}$, the TPOP method is defined as:

$$s_{x,y}^{TPOP} = \frac{|\Lambda_{x,y}^{TOP}|}{|\Lambda_{x,y}^{POP}|}. \tag{3.57}$$

TPOP is inspired in TPOG, which is other of our proposals to perform link prediction using overlapping social group information and defined in Section 3.2.4.

Friend Allocation Within Common Places (FAW)

FAW is inspired in the traditional Resource Allocation method, but considering the number of check-ins made by all common friends within common visited places of users x and y. Therefore, FAW is defined as:

$$s_{x,y}^{FAW} = \sum_{z \in \Lambda_{x,y}^{WCP}} \frac{1}{|\Phi(z)|}. \tag{3.58}$$

Despite the use of place and check-in frequency by FAW, we consider it as a method based on social strength, due to this criterion be the main filter used to perform its predictions.

Common Neighbors of Nearby Places (CNNP)

CNNP counts the number of common friends of users x and y whose geographical distance between their home locations and the home location of at least one, x or y, lies within a given radio. Therefore, given a distance threshold τ_d, CNNP is computed as:

$$s_{x,y}^{CNNP} = |\{z \mid \forall z \in \Lambda_{x,y} \ \wedge \ (dist(\ell_x^h, \ell_z^h) \leq \tau_d \ \vee \ dist(\ell_y^h, \ell_z^h) \leq \tau_d)\}|. \tag{3.59}$$

CNNP uses full place information as well as social information to make its predictions, however we consider it as a method based on social strength.

Nearby Distance Allocation (NDA)

NDA refines all the minimum adjusted distances calculated between the home locations of users x and y, and the respective home locations of all their common neighbors of places. Therefore, NDA is defined as:

$$s_{x,y}^{NDA} = \sum_{z \in \Lambda_{x,y}^{\mathscr{L}}} \frac{1}{\min\{dist_{adj}(\ell_x^h, \ell_z^h), dist_{adj}(\ell_y^h, \ell_z^h)\}}. \tag{3.60}$$

NDA is the only method using place, check-in and social information. However, as previously applied for the other proposals, since NDA uses the social information as the main criterion, we consider it as a method based on social strength.

3.3.3 Experimental Evaluation

We consider a scenario where new friendships of two different LBSNs must be predicted. On these networks, we compare the performance of our 8 proposals regarding 22 different methods from the literature. Considering the amount of link prediction methods to be analyzed, we only perform our experiments under unsupervised strategy.

LBSN Datasets

The datasets used in our experiments are real-world LBSNs in which users made check-ins to report visits to specific physical locations. In this section, we describe their main properties and the ways to construct the training and test datasets.

Dataset Selection

Datasets for our experiments have to meet certain requirements: i) they have to represent social and location data, i.e. data defining the existing connections between users as well as the check-ins made by all of them in all their visited locations, and ii) those connections and/or check-ins have to be time stamped. Based on these two criteria, we select two datasets collected from real-world LBSNs, which are commonly used in the link prediction task by the scientific community.

Brightkite. It was once a location-based social networking service provider where users shared their locations by checking-in. The Brightkite service was shut down on 2012, but the dataset was collected over the period of April 2008 to October 2010 (CHO; MYERS; LESKOVEC, 2011). This publicly available dataset[12] consists of 58228 users, 214078 relations, 4491144 check-ins, and 772788 locations.

Gowalla. It is also another location-based social networking service that ceased operation in 2012. The dataset was collected over the period of February 2009 to October 2010 (CHO; MYERS; LESKOVEC, 2011) and also is publicly available[13]. This dataset contains 196591 users, 950327 relations, 6442892 check-ins, and 1280969 different locations.

The various properties of these networks are depicted in Table 9. In this table we observe that the analyzed networks have a small *average degree*, $\langle k \rangle$, which suggests that the users of these networks have between 7 and 9 friends on an average. This fact implies that the *average clustering coefficient*, *CC*, of both networks is too low. However, the low *degree heterogeneity*, H, of Brightkite indicates that its users are less different than the users of Gowalla. Also, the *assortativity coefficient* r shows that only Brightkite is assortative, due to which it has a positive value, indicating the presence of low amount of relationships among the users with similar degree. On the other hand, Gowalla is disassortative, since its assortativity coefficient is negative, indicating the presence of a considerable amount of relationships among users with different degree.

From Table 9 we also observe that the *number of users with at least one check-in*, $|\Phi_V|$, is a little over 85% of total users of both networks. However, despite the fact that Gowalla has

[12] <http://snap.stanford.edu/data/loc-brightkite.html>
[13] <http://snap.stanford.edu/data/loc-gowalla.html>

Table 9 – The main properties of LBSNs analyzed.

Properties	Brightkite	Gowalla
$\|V\|$	58,228	196,591
$\|E\|$	214,078	950,327
$\langle k \rangle$	7.35	9.66
CC	0.17	0.24
H	8.66	31.71
r	0.01	−0.03
$\|\Phi\|$	4,491,144	6,442,892
$\|\Phi_V\|$	50,686	107,092
$\langle \Phi \rangle$	88	60
$\|\mathscr{L}\|$	772,788	1,280,969
$\langle \mathscr{L}_\Phi \rangle$	5	5
$\langle \mathscr{E} \rangle$	0.05	0.25

Source: Elaborated by the author.

more check-ins and locations and more users making check-ins than Brightkite, the *average number of check-ins per user*, $\langle \Phi \rangle$, that users of Brightkite have, is greater than the users of Gowalla. However, the *average of check-ins per place*, $\langle \mathscr{L}_\Phi \rangle$, made by users of Brightkite and Gowalla is similar. Finally, very small value of the *average of places entropy*, $\langle \mathscr{E} \rangle$, of Brightkite suggests that the locations in this LBSN represent a stronger factor to facilitate the establishment of new relationships among its users than for users of Gowalla.

Data Processing

To make the data suitable for the experiments, we perform preprocessing over both the datasets. It consists of two steps: select data samples and split the data into training and testing sets.

A. Select Data Samples

Isolated nodes and locations without visits can generate noise for measuring the performance of different link prediction methods. To eliminate the impact of this noise, for each dataset, we consider only users with at least one friend and with at least one check-in made at any location.

B. Split the Data into Training and Testing Sets

Considering that we aim to predict new friendships among users, we divide each dataset into training and test (or probe) sets taking into account the time stamps information available.

Therefore, links formed by users of Brightkite who made check-ins from April 2008 to January 2010 are used to construct the training set, whilst links formed by users who made check-ins from February 2010 to October 2010 are used for the probe set. Whereas for Gowalla, the training set is constructed with links formed by users made check-ins from February 2009 to April 2010, and the probe set is constructed with links formed by users made check-ins from May 2010 to October 2010. Table 10 shows the training and testing time ranges.

Table 10 – Details of pre-processed LBSN datasets.

Dataset	**Training time range**	**Testing time range**	$\langle\|V\|\rangle$	$\langle\|\mathscr{L}\|\rangle$	$\langle\|E^T\|\rangle$	$\langle\|E^P\|\rangle$
Brightkite	2008/04 - 2010/01	2010/02 - 2010/10	4,606	277,515	49,460	24,800
Gowalla	2009/02 - 2010/04	2010/05 - 2010/10	19,981	607,094	232,194	87,619

Source: Elaborated by the author.

Different researches have used a similar strategy for splitting the data into training and test sets, but they did not have to maintain the consistency between the users in both sets (BAYRAK; POLAT, 2014; LUO *et al.*, 2013). This fact could affect the performance of link prediction methods differently (YANG; LICHTENWALTER; CHAWLA, 2015; LICHTNWALTER; CHAWLA, 2012; ALLALI; MAGNIEN; LATAPY, 2013). To avoid that, we remove all the links formed by users which made check-ins only during the training time interval or only in testing time interval. From the links formed by users with check-ins in both training and testing time range, we choose one-third of the links formed by users at random with a degree higher than the average degree for the test set, while the remaining links will be part of the training set. Therefore, we obtain the training set $G^T(V,E^T,\mathscr{L})$ and test set $G^P(V,E^P,\mathscr{L})$, where both sets have the same users, V, and locations, $\mathscr{L}$.

Table 10 also summarizes the average final number of nodes, $\langle|V|\rangle$, number of different locations, $\langle|\mathscr{L}|\rangle$, number of links for training, $\langle|E^T|\rangle$, and number of links for testing, $\langle|E^P|\rangle$, obtained by averaging 10 independent partitions of each dataset.

Data Limitations

Although the datasets selected contain thousands of users and links, they can be considered as relatively small compared to other online social network datasets analyzed in previous chapters of this thesis. Furthermore, we underline the fact that typical users of LBSNs have different mobility and social behaviors than users of traditional online social networks. This

would lead to some of the observed network properties as the reason to explain certain patterns of users of Gowalla and Brightkite that could not reflect user behaviors of other LBSNs.

Notwithstanding these limitations present in our datasets, we use them in this work since they meet the requirements explained previously in Section 3.3.3 and also because they have been widely analyzed by different researchers, who have identified the social and spatial factors influencing the edge creation process (ALLAMANIS; SCELLATO; MASCOLO, 2012; CHO; MYERS; LESKOVEC, 2011; MENGSHOEL *et al.*, 2013; BAYRAK; POLAT, 2014; GRABOWICZ *et al.*, 2014). Hence, this work sheds light on how to explore the different information sources to improve friendship prediction in Brightkite and Gowalla, but our findings can pave the way to further investigation on other LBSNs.

Here it is important to note that, despite other researches have been used datasets collected from Foursquare (LUO *et al.*, 2013; ZHANG; KONG; YU, 2014; YE; YIN; LEE, 2010), Facebook (BACKSTROM; SUN; MARLOW, 2010; MCGEE; CAVERLEE; CHENG, 2013), Twitter (ZHANG; KONG; YU, 2014; CHENG *et al.*, 2011; GRABOWICZ *et al.*, 2014), Second Life (STEURER; TRATTNER; HELIC, 2013; STEURER; TRATTNER, 2013), and other LBSNs, we do not use them since these datasets are not publicly available, or its edge creation process have not been sufficiently analyzed, or simply the lack in any of the requirements specified in Section 3.3.3.

Experimental Setup

For each one of the 10 independent partitions of each dataset obtained as explained in Section 3.3.3, we consider 10 executions of each link prediction method presented in Section 2.3.3, including our proposals previously described in Section 3.3.2. Due to that our experimentation will consider only the unsupervised strategy, we adopted all the performance measures showed in Section 2.2.2 and applied them on the prediction results to determine which were the most accurate and efficient link prediction methods in LBSN domain. We set the default parameters of link prediction methods analyzed as follows:

- For methods considering a temporal threshold τ, such as Co, DCo and TCFCC, we considered that $\tau = 1$ day.
- For methods considering an entropy threshold $\tau_{\mathscr{E}}$, such as LC, we considered that $\tau_{\mathscr{E}} = \langle \mathscr{E} \rangle$.
- For methods considering a geographical distance threshold τ_d for comparison among place check-ins, such as ChL, we considered that $\tau_d = 300$ meters. When this threshold is used specifically for comparison among different places, we considered that $\tau_d = 1500$ meters.
- For methods based on the calculation of least upper and greatest lower bounds, such as HD and AHD, for a user x and being ℓ the most visited place by him, we considered that the

comparison value for the calculation of supremum is $v_s = \frac{|\Phi(x,\ell)|}{2}$, whilst the comparison value for the calculation of infimum is $v_i = \frac{|\Phi(x,\ell)|}{5}$.

Evaluation Results

For the two LBSNs analyzed, Table 11 summarizes the performance results for each link prediction method through different evaluation measures. Each value in these table was obtained by averaging over 10 run over 10 partitions of training and testing sets as previously detailed in Section 3.3.3. Values emphasized in bold correspond to the best results achieved for each evaluation measure.

From Table 11, results for imbalance ratio, precision, recall, and F-measure were calculated considering the whole list of predicted links obtained by each link prediction method evaluated. On the other hand, results for AUC were calculated considering $n = 5000$.

Due to the number of link prediction methods studied and the different ways they were evaluated, we performed a set of analysis in order to determine which were the best link prediction methods based on location information.

Reducing the Prediction Space Size

Previous studies showed that the prediction space size of methods based only on network topology is in the order of $10^{11} \sim 10^{12}$ links for Brightkite and Gowalla. However, by using methods based on location information, the prediction space can be reduced by about 15 or more times (SCELLATO; NOULAS; MASCOLO, 2011). Based on that and aiming to determine if the reduction of prediction space is related to different location information sources, in Figure 29 we report the average prediction space size of different link prediction methods analyzed in this work.

In Figure 29 we observe for the analyzed networks, methods based on frequency, entropy and geographical distance follow the traditional logic of obtaining a high number of right predictions at the price of a much higher number of wrong predictions (WANG *et al.*, 2011). On the other hand, methods based on social strength lead to a considerable lower number of wrong predictions in cost of a small number of correctly predicted links. Our proposals improve the scheme of the methods based on social strength, but increasing considerably the number of right predictions.

This fact is clearly shown trough the IR results in Table 11, where besides highlighting that Co, DCo and TCFCC methods have a general better IR performance, we observe that some methods based on frequency and most of the methods based on geographical distance have a IR three, or more, times higher than most of methods based on social strength and our proposals. Therefore, Co and DCo are the methods based on frequency with better IR performance, whilst PAP and PAC are the worst ones. Among the methods based on information

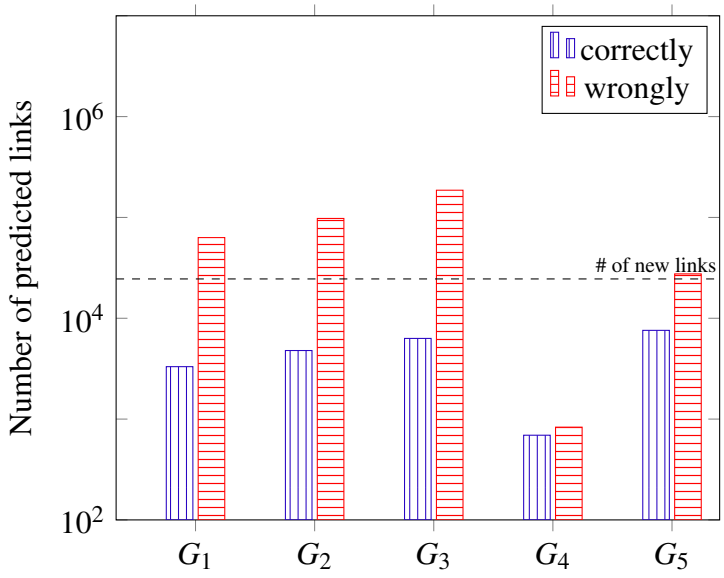
Table 11 – Friendship prediction results for Gowalla and Brightkite. Highlighted values indicate the best results for each evaluation measure considered.

Method		IR	P	R	F_1	AUC		IR	P	R	F_1	AUC
Co		**4.934**	**0.211**	0.042	0.070	0.668		14.972	0.069	0.040	0.051	0.554
DCo		**4.934**	**0.211**	0.042	0.070	0.653		14.972	0.069	0.040	0.051	0.567
CL		13.114	0.079	0.156	0.105	0.676		36.327	0.028	0.115	0.045	0.682
JacP		13.114	0.079	0.156	0.105	0.630		36.327	0.028	0.115	0.045	0.742
LO		13.114	0.079	0.156	0.105	0.630		36.327	0.028	0.115	0.045	0.736
CLR		13.114	0.079	0.156	0.105	0.617		36.327	0.028	0.115	0.045	0.677
CLC		13.079	0.079	0.151	0.104	0.692		31.197	0.033	0.107	0.050	0.747
PAP		35.005	0.030	0.267	0.053	0.659		180.461	0.006	0.189	0.011	0.628
PAC		35.005	0.030	0.267	0.053	0.586		180.461	0.006	0.189	0.011	0.637
AAP		13.190	0.079	0.154	0.104	0.682		36.531	0.028	0.113	0.045	0.728
MinC		15.084	0.069	0.129	0.090	0.642		43.803	0.023	0.089	0.037	0.698
MinE		15.859	0.065	0.118	0.084	0.638		47.076	0.022	0.081	0.034	0.701
AAE		13.190	0.079	0.153	0.104	0.694		36.586	0.028	0.113	0.045	0.736
LC		34.000	0.030	0.261	0.055	0.629		180.945	0.006	0.185	0.011	0.542
MinD	**Brightkite**	35.008	0.030	0.267	0.053	0.739	**Gowalla**	180.458	0.006	0.189	0.011	**0.830**
ChD		43.873	0.023	**0.427**	0.044	0.472		157.348	0.007	**0.397**	0.013	0.426
ChL		16.085	0.064	0.183	0.095	0.610		38.230	0.027	0.126	0.044	0.527
GeoD		35.005	0.030	0.267	0.053	0.710		180.461	0.006	0.189	0.011	0.767
WGeoD		35.005	0.030	0.267	0.053	0.304		180.461	0.006	0.189	0.011	0.211
HD		31.689	0.031	0.260	0.056	0.692		223.714	0.006	0.188	0.011	0.728
AHD		31.689	0.031	0.260	0.056	0.685		223.714	0.006	0.188	0.011	0.681
TCFCC		5.279	0.199	0.051	0.082	0.633		**13.493**	**0.077**	0.032	0.045	0.589
ChO		13.079	0.079	0.151	0.104	0.608		31.197	0.033	0.107	0.05	0.714
ChA		13.173	0.079	0.155	0.104	0.676		36.460	0.028	0.114	0.045	0.736
WOCP		9.678	0.108	0.120	**0.113**	0.515		15.821	0.065	0.073	**0.069**	0.480
CNP		31.180	0.033	0.339	0.060	**0.761**		66.484	0.015	0.282	0.029	0.687
TPOP		13.441	0.077	0.165	0.105	0.673		25.383	0.040	0.099	0.057	0.665
FAW		9.678	0.108	0.120	**0.113**	0.740		15.821	0.065	0.073	**0.069**	0.718
CNNP		9.387	0.110	0.031	0.048	0.552		18.868	0.056	0.039	0.046	0.620
NDA		22.496	0.046	0.221	0.076	0.700		47.54	0.022	0.132	0.037	0.720

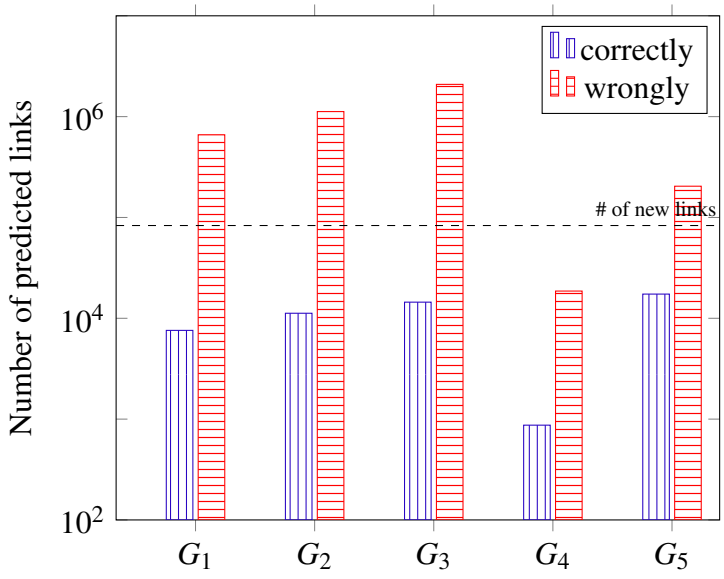
Source: Elaborated by the author.

gain, AAE highlighted as the method with best IR performance, while LC performed more poorly. Among the methods based on geographical distance, the most accurate method with respect to IR is ChL, the others has similar performance, with IR two or three times more than

Figure 29 – Number of correctly and wrongly predicted links for methods based on frequency (G_1), entropy (G_2), geographical distance (G_3), social strength (G_4), and our proposals based on frequency and social strength (G_5) for (a) Brightkite and (b) Gowalla. The dashed horizontal line indicates the number of truly new links (links into the probe set). Results averaged over the 10 analyzed partitions and plotted in log 10 scale.



(a) Brightkite



(b) Gowalla

Source: Elaborated by the author.

ChL, as minimum. TCFCC is the only method based on social strength and shows a good IR performance, positioning among the first for both networks. Among our proposals, we found that FAW and WOCP performed better in IR, followed closely by CNNP. These three methods have social components, which help to significantly reduce the prediction space size. The worst IR of our proposals was obtained by NDA and CNP.

Measuring the Accuracy

Based on Table 11 we observe that the precision (P) of a link prediction method is inversely proportional to its respective IR, i.e. the greater the IR value the lower the precision value. Therefore, Co and DCo have the best precision performance on Brightkite, whilst TCFCC is the best on Gowalla. With regard to the recall (R), ChD outperforms all the prediction methods both in Brightkite and Gowalla.

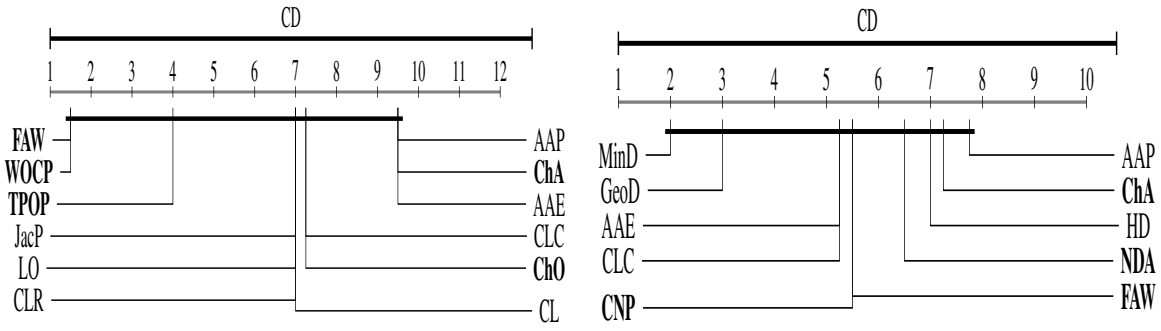
Due to that some methods obtained a considerable number of correctly predicted links whilst others obtained a low number of wrongly predicted links, we have adopted the f-measure (F_1) to observe the performance of prediction methods in terms of relevant predicted links. Two of our proposals, WOCP and FAW, had the better f-measure performance in the two analyzed LBSNs.

To facilitate the analysis of all link prediction methods, from Table 11 we ranked the average results of F_1 obtained by all the link prediction methods in both analyzed networks, after that we took the top-10. Over the top-10 selected methods, we applied the Friedman and Nemenyi post-hoc tests (DEMSAR, 2006). Due to the presence of ties, 12 link prediction methods rank into the top-10 of F_1 performance. Therefore, the F-statistics with 11 and 11 degrees of freedom and at 95 percentile was 2.82. According to the Friedman test using the F-statistics, the null-hypothesis that the top-10 link prediction methods behave similarly when compared by their F_1 performance should not be rejected.

Figure 30a shows the Nemenyi test results for the 12 link prediction methods in the top-10 positions of F_1 ranking. The critical difference (CD) value for comparing the mean-ranking of two different methods at 95 percentile is 11.78, as showed on the top of the diagram. In the axis of the diagram are showed the names of the methods, highlighting our proposals in bold. The lowest (best) ranks are in the left side of the axis. Furthermore, Nemenyi test indicates that the top-10 methods have no statistical significant difference, so they are connected by a bold line in the diagram.

Despite the top-10 methods have no significant statistical difference when analyzed their F_1 performance, from Figure 30a we observe that our proposals based on social strength such as FAW, WOCP and TPOP perform better than others, being that FAW and WOCP tie in the first position, whilst TPOP is in the second position. After these three measures, and a little further away, the rest of the ranking is formed by methods based on frequency: JacP, LO, CLR, and CL tying in the third position, CLC and ChO, which is one of our proposals based on frequency, tie in the fourth position. ChA, which also is other of our proposals based on frequency, ties in fifth position with AAE and AAP. Also, we observe that none method based on geographical distance performed well enough to be in this top-10 ranking.

Figure 30 – Nemenyi post-hoc test diagrams obtained from (a) f-measure and (b) AUC results showed in Table 11. Diagrams show the link prediction methods in the top-10 positions. Our proposals are highlighted in bold.



(a) F-measure (b) AUC

Source: Elaborated by the author.

Analyzing the Predictive Power

Table 11 also shows the prediction results obtained for AUC. From these results we observe that CNP, which is one one of our proposals, and MinD outperformed all the other link prediction methods in Brightkite and Gowalla. Also, we observe that all the link prediction methods perform better than pure chance, except ChD and WGeoD in Brightkite, and WOCP, ChD and WGeoD in Gowalla.

Furthermore, to have a better idea of the real prediction power of evaluated link prediction methods, we followed the same scheme previously used for F_1 analysis. Therefore, we ranked the average results of AUC obtained by all the link prediction methods and took the top-10. Over the top-10 selected methods, we applied the Friedman and Nemenyi post-hoc tests. The critical value of the F-statistics with 9 and 9 degrees of freedom and at 95 percentile was 3.18. Based on F-statistics, the Friedman test suggests that the null-hypothesis that the top-10 link prediction methods behave similarly when compared by their AUC performance should not be rejected.

Figure 30b shows the Nemenyi test results for the top-10 methods ranked by AUC. The diagram indicates that the CD value calculated at 95 percentile is 9.58. This test also shows that the top-10 link prediction methods have no significant difference, so they are connected by a bold line.

From Figure 30b we observe that two methods based on geographical distance, MinD and GeoD, are in the first and second position, respectively, AAE and CLC tie in the third position, whilst FAW and CNP, which are two of our proposals based on social strength, tie in the fourth position. NDA, other of our proposals, is fifth, HD is sixth, ChA, which also is other of our proposals, is seventh, and finally, AAP is in eighth position. It is important to note that, differently from the F_1 ranking, in the AUC ranking we observe the presence of methods based on geographical distance (MinD and GeoD). However, most of the methods in the AUC ranking

are those based on check-in/place frequency and social strength.

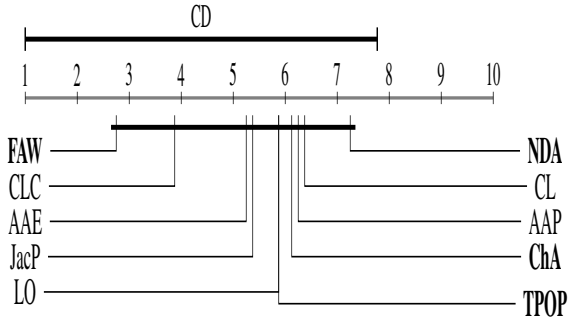
Recommending New Friendships

Since some link prediction methods performed better in the prediction space analysis whilst other ones in the prediction power analysis, there is no consensus on which are, definitively, the more suitable methods for practical implementations, such as the friendship recommendation task. For that purpose, we analyze F_1 and AUC results at the same time.

From Table 11 we rank the average results of F_1 and AUC obtained by all the link prediction methods. From this ranking, we take the top-10 methods and apply over them the Friedman and Nemenyi post-hoc tests. The critical value of the F-statistic with 9 and 27 degrees of freedom and at 95 percentile is 2.25. Based on this F-statistic, the Friedman test suggests that the null-hypothesis that the top-10 methods behave similarly when compared by their F_1 and AUC performances should not be rejected.

Figure 31 shows the Nemenyi test result for the top-10 methods in our final ranking, which contains the methods that in addition to reduce optimally the prediction space size also have a competitive prediction power. The diagram of Figure 31 indicates that the CD value at 95 percentile is 6.77, hence the methods have no significant difference, so they are connected by a bold line.

Figure 31 – Nemenyi post-hoc test diagram obtained over the F_1 and AUC average ranks showed in Table 11. Diagram shows the top-10 link prediction methods with the best performance considering the optimal reduction of prediction space size and high prediction power. Our proposals are highlighted in bold.

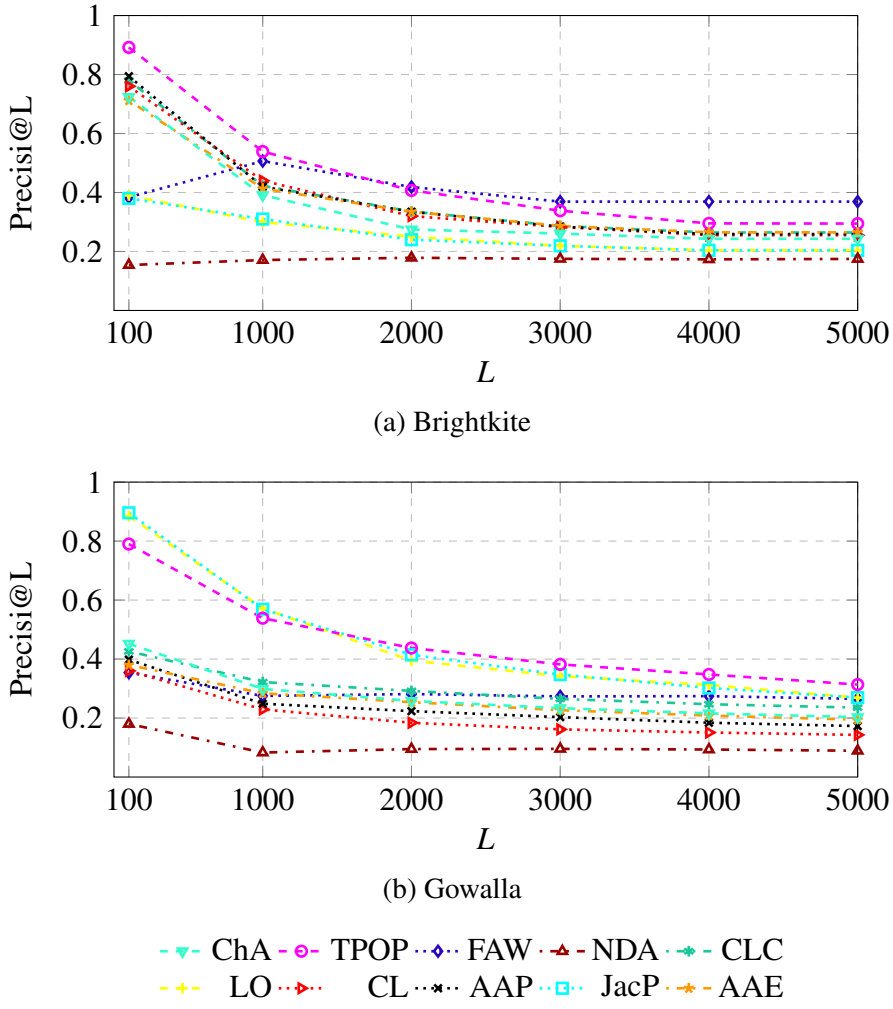


Source: Elaborated by the author.

In Figure 31, we observe that one of our proposals, FAW, outperformed all the link prediction methods. CLC ranks as the second. AAE is third and JacP is fourth. In the fifth position, LO ties with one of our proposals, TPOP. ChA, which also is one of our proposals, is sixth. AAP and CL are in seventh and eighth positions, respectively. Finally, other of our proposals, NDA, is ninth. It is important to note that most of the methods in the final ranking are based on check-in/place frequency and social strength, only one is based on information gain (AAE) and none is based on geographical distance.

Considering that, for recommending to users some links as possible new friendships, we can just select the links with the highest scores. Furthermore, whereas for recommendation task is not enough only a method with good prediction performance, also it is necessary that from the top-L predicted links it generates a high number of right predictions, good enough to be showed to users as appropriate friendship suggestions (GUNAWARDANA; SHANI, 2009; RAFAILIDIS; CRESTANI, 2016). Therefore, in Figure 32 we show different Precisi@n performance of all the top-10 methods of our final ranking. These Precisi@n results are calculated for different L values and for each LBSN analyzed.

Figure 32 – Precisi@L performance of the top-10 methods of the final ranking considering different L values for (a) Brightkite, and (b) Gowalla.



(a) Brightkite

(b) Gowalla

Source: Elaborated by the author.

From Figure 32a, we observe that in Brightkite, one of our proposals based on social strength, TPOP, outperforms all the other evaluated link prediction methods when $L = 100$ and $L = 1000$. After that, other of our our proposals also based on social strength, FAW, overcomes TPOP and performs better than the rest of methods. In Gowalla, the dominance of TPOP and FAW changes. Therefore, from Figure 32b we observe that LO and JacP, which poorer performance in

Brightkite, achieve the best performance over all the evaluated link prediction methods when $L = 100$ and $L = 1000$. After that, TPOP overcomes LO and JacP and performs better than the rest of methods.

Here it is important to note that, from Figure 32 we observe that most of the evaluated methods have their better performance when $L = 100$, i.e. they have the ability to make a small amount of accurate friendship recommendations. When link prediction methods have to make more than a thousand recommendations, i.e when $L > 1000$, their recommendation abilities decrease considerably, achieving even Precisi@L values below 0.5.

From a general view of Precisi@L results and the overall performance of top-10 methods of our final ranking, we clearly distinguish three facts: i) NDA, one of our proposals, has the worst recommendation ability among the top-10 methods and not even get to achieve a precisi@L value of 0.2 in any L value; ii) CL, CLC, AAP, AAE, and our proposal, ChA, show a similar Precisi@L performance behavior to each other; and, iii) although it is true that two of our proposals based on social strength, i.e. TPOP and FAW, outstand in Brightkite, the same does not occur in Gowalla, where highlight LO and JacP, which are two measures based on place frequency.

The struggle between FAW and TPOP with LO and JacP for the top positions can be justified by the network structure. Therefore, the high heterogeneity degree (H) and the negative assortativity coefficient (r) of Gowalla indicate that users of this network, besides being more different than users of Brightkite by the number of friends that may have, also the tendency of establish more friendships with different users. This tendency may disadvantage the performance of link prediction methods based on social strength when applied in Gowalla.

3.3.4 Remarks

This section provided a new categorization of existing link prediction methods based on structural similarity in the domain of LBSNs. This categorization is based on similarity criteria and information sources used to perform the prediction task. Furthermore, based on this categorization, we identified some gaps in existent friendship prediction methods and proposed eight new ones.

Due to that we aim to take a concrete step towards objectively quantifying the predictive power of friendship prediction methods in LBSNs as well as determine how good they work in the context of recommender systems, our evaluation process has been performed on traces of two well known real-world LBSNs, and considering a variety of evaluation measures. We discussed as the prediction methods performed and selecting the top ten for each evaluation measure used.

We empirically demonstrate that in some cases some friendship prediction methods can be ranked as the best for any evaluation measure but can perform poorly for other one. Thus, we emphasize the importance of choosing the appropriate evaluation measure according to the

objective pursued in the friendship prediction task. For instance, in general, some friendship prediction methods based on social strength perform better in F-measure than in AUC, so if in any real-world application is necessary to focus on minimize as much as possible the number of wrong predictions, the best option is to consider methods working well based on F-measure. However, if the focus is to have a considerable number of right predictions but with a high chance of become real new friendships, so the best option could be consider methods working well based on AUC. Furthermore, we also identified some methods performing in a balanced way for different metrics, such as one of our proposals, FAW.

3.4 Summary

In this chapter we have provided a variety of proposals to face the link prediction problem in both the traditional online social networks as well as the recent arrivals location-based social networks. We started by introduce the link prediction problem using community information in real world online social networks. Therefore, our contribution here is limited to the proposal of one strategy combining a fast and efficient community detection algorithm with fast and accurate link prediction methods based on community information.

After showing that community information can improve the accuracy of link prediction methods, and that its use is feasible for large-scale networks, we introduce the use of social groups in link prediction. Social groups are naturally occurring structures in online social networks; therefore, they are a more sensible choice when compared to communities, because there is no added computational cost related to the algorithm required to identify such communities. There is important to note that, to the best of our knowledge we are the first in consider the presence of social groups in the link prediction problem. Our contribution here consists in the formalization of new network properties, such as the *overlapping groups clustering coefficient*, as well as seven new link prediction methods: *common neighbors within and outside common groups* (WOCG), *common neighbors of groups* (CNG), *common neighbors with total and partial overlapping of groups* (TPOG), *group naïve Bayes* (GNB), *group naïve Bayes of common neighbors* (GNB-CN), *group naïve Bayes of Adamic-Adar* (GNB-AA), and *group naïve Bayes of Resource Allocation* (GNB-RA).

Finally, by establishing a parallel between information of users participating in social groups and information of users visiting places, we explore the friendship prediction problem in the context of location-based social networks. Due to that most of the existing friendship prediction methods focus on use only the location information to perform their predictions, we bet on combining this information source with the social strength. This combination leads to optimize considerably the prediction space size, reducing the number of wrong predictions and increasing the number of right ones. Therefore, our contribution here consists of eight new friendship prediction methods for LBSNs: *Check-in Observation* (ChO), *Check-in Allocation*

(ChA), *Within and Outside of Common Places* (WOCP), *Common Neighbors of Places* (CNP), *Total and Partial Overlapping of Places* (TPOP), *Friend Allocation Within Common Places* (FAW), *Common Neighbors of Nearby Places* (CNNP), and *Nearby Distance Allocation* (NDA).

In the next chapter, we will summarise the findings of this thesis. Furthermore, we will discuss in detail our contributions as well as directions for future work.

CHAPTER

4

CONCLUSION

This thesis aims to alleviate some of the problems related to the understanding of user behavior in online social networking services, including location-based social networks. We have conducted a comprehensive literature survey about the main topics addressed through this thesis, i.e. complex networks, link prediction and location-based social networks. Although different mining tasks can be used to understand user behavior in social networks, in this thesis we focus on link prediction since it naturally explains the network evolution process and the different phenomena that lead to the establishment of new relationships among users.

Despite the large number of studies published using link prediction task to understand user behavior, in this thesis we have focused on consider two different aspects few discussed of this task. The first aspect is concerned with the use of information related to the membership of users in communities or social groups. Therefore, we started by analyzing the feasibility of using *link prediction methods based on community information* in large scale networks. Since it is mandatory that previously communities have been discovered by any community detection algorithm, the use of this type of link prediction methods may be restrictive. However, we have presented a strategy to deal with that challenge. Obtained results on a large-scale real-world network (Twitter), lead us to prove the feasibility of using community information in the link prediction context in large-scale networks, as well as to infer that the use of this information source may greatly contributes to accurate friendship predictions and to better understand user behavior.

Since online social networks offer services to facilitate that users create and participate in many social groups by voluntary way, these social groups become a more sensible information source when compared to communities. Since there is no additional computational cost related to the algorithm required to identify such social groups. Therefore, considering that social groups are structures which naturally occur in OSNs, we have introduced the use of social groups in link prediction context. Using social groups in link prediction opens new challenges, specially

related to the consideration that users participate in multiple social groups at the same time as well as to the presence of overlapping social groups. Therefore, we addressed these challenges by proposing a set of new network properties, which efficiently extract social groups characteristics, as well as seven new *link prediction methods based on social group information*. Obtained results on four large-scale real-world networks (Flickr, LiveJournal, Orkut and Youtube), corroborate the idea that the use of social group information besides of contributing to improve friendship prediction accuracy, also conveys relevant clues to better understand user interest and behavior.

The second aspect related to link prediction and widely discussed in this thesis is concerned with the efficient way of using location information in location-based social networks. Since most of the existing link prediction methods try to maximize the number of right predictions at the cost of generating a very large number of wrong ones. Therefore, by considering our previous findings, we proposed new eight link prediction methods which combine efficiently location information with social strength and other information sources. Obtained results on two large-scale real-world networks (Brightkite and Gowalla), prove that, besides of contributing to improve friendship prediction accuracy and better understand the behaviors of users of location-based social networks, our proposals are very suitable for real world applications since they considerably reduce the number of wrong predictions and increase the number of right ones.

Finally, through this thesis we have showed how, by applying link prediction, to understand different behaviors of users in online social networks. Furhermore, the understanding of these behaviors can be very useful to improve the performance of the same mining task in a new domain of networks, i.e. location-based social networks.

During our research, some contributions to the state-of-the-art of complex networks and link prediction have been achieved. The mains contributions achieved, as well as the limitations and indications of future work are listed in the following subsections.

4.1 Contributions

This thesis contributes to two research areas: complex networks and social network analysis, with focus on link prediction, location-based social networks, and user behavior analysis. Specifically, the following contributions have been made:

- *Review of the research in link prediction*. Despite different authors have published extensive surveys on the link prediction task, most of them do not explain in detail the problem statement or do not show a clear categorization of existing methods. Therefore, our contribution here is that, in Section 2.2 we have extensively surveyed all the aspects related to link prediction. We started by positioning the link prediction as one of the most important link mining tasks, for later defining the link prediction problem from both unsupervised and supervised perspectives. Under the two perspectives, we categorize and listed a variety

of link prediction methods. Finally, we show the main link prediction applications. We believe this survey will help researchers in link prediction and link mining by providing a detailed overview of the area. An article containing this survey is in final preparation and is going to be submitted to a journal.

- *Overcome the challenges related to the use of community information in link prediction*. Most of existing link prediction methods using community information have shown no concern about the quality of the communities used as information source to perform their prediction. Moreover, since the community detection process represents a computational cost, most of the existing link prediction methods using community information are not suitable for real-world applications. Therefore, in Section 3.1 we have proposed a strategy combining a fast and efficient community detection algorithm with fast and accurate link prediction methods based on community information. Based on this strategy, we have showed that the use of community information in the link prediction context may greatly contribute to accurate friendship predictions as well as to a better understand user behavior. Moreover, by using our strategy it is possible perform link prediction using community information in large-scale networks. Part of this contribution has been previously published in a specialized journal (VALVERDE-REBAZA; LOPES, 2013).

- *Open a new research issue related to the use of social group information in link prediction*. Despite the fact of existing a variety of link prediction methods using community information, most of them (or in fact all) assume that a user belongs to a single community. However, in real-world online social networking services, users choose to participate in social groups in which other users sharing similar interests also participate. That means, social group is a naturally existing entity and users can participate in many of them at the same time. Being an important information source, Section 3.2 introduced the use of social group information in link prediction. Therefore, we have established the main challenges related to the use of social groups in link prediction, such as the consideration of the user's participation in many groups and the presence of overlapping groups. Also, we proposed new network properties to better explore group information in complex networks. Based on these properties, we proposed seven new link prediction methods classified in two types: *based purely on network topology* and *based on Naïve Bayes model*. Our proposals based purely on network topology were WOCG, CNG, and TPOG, whilst our proposals based on Naïve Bayes model were GNB, GNB-CN, GNB-AA, and GNB-RA. Besides even outperforming state-of-the-art link prediction methods, our proposals conveys relevant clues to better understand user's interest and behavior. To the best of our knowledge, we have conducted the first research considering social group as information source in link prediction. Part of this contribution has been previously published in two international conferences (VALVERDE-REBAZA; LOPES, 2014; VALVERDE-REBAZA *et al.*, 2015).

- *Overcome some challenges related to the use of location information in link prediction*.

Most of the existing link prediction methods for location-based social networks focus mainly on the use of location information to perform their predictions, giving little attention to other information sources. Moreover, these methods fall in the trap of prioritize the number of right predictions at the cost of a large amount of wrong ones. Despite the fact that this situation is recurrent for the link prediction in general, this has a gretaer impact on the context of location-based social networks since the user's location constitute a new actor in the network structure, i.e. by having a new dimension, the computational cost of all mining task on location-based social networks should consider, besides the number of nodes and links, the number of locations. Therefore, in Section 3.3 we have proposed eight new link prediction methods: ChO, ChA, WOCP, CNP, TPOP, FAW, CNNP, and NDA. These proposals combine location information with other information sources, mainly social strength, obtaining more accurate predictions and considerably reducing the number of wrong predictions compared to literature methods. Part of this contribution has been previously published in an international conference (VALVERDE-REBAZA *et al.*, 2016).

All the evaluation results showed in this thesis have been obtained by the use of two frameworks developed by us and publicly available. Our first framework is called *LPsource*[1], and contains the source code of all our contributions related to link prediction using community information and overlapping group information. Our second framework is called *GeoLPsource*[2] and contains the source code of all our contributions related to link prediction in LBSNs. Our two frameworks have been implemented using, mainly, C++ language and can be redistribute and/or modify under the terms of the GNU General Public License[3].

During our research, some collaborations in related projects have been established. These collaborations have contributed to different link mining tasks, such as classification (CHERMAN *et al.*, 2013; VALVERDE-REBAZA *et al.*, 2014; CHERMAN *et al.*, 2015; BERTON *et al.*, 2017), community detection (VALEJO *et al.*, 2014a; VALEJO *et al.*, 2014b; VALEJO; VALVERDE-REBAZA; LOPES, 2014), and others (DRURY *et al.*, 2014; DRURY; VALVERDE-REBAZA; LOPES, 2015; BERTON; VALVERDE-REBAZA; LOPES, 2015; BERTON *et al.*, 2016; BERTON *et al.*, 2017; DRURY *et al.*, 2017). Detailing these contributions is not in the scope of this thesis.

4.2 List of Publications

The above contributions lead the publication of conference papers and journal articles directly related to this research:

[1] <https://github.com/jvalverr/LPsource>
[2] <https://github.com/jvalverr/Geo-LPsource>
[3] <http://www.gnu.org/licenses/>

- **VALVERDE-REBAZA, J.**; ROCHE, M.; PONCELET, P.; LOPES, A.. *Exploiting social and mobility patterns for friendship prediction in location-based social networks*. In Proceedings of 23rd International Conference on Pattern Recognition (ICPR 2016), 2016, p. 2527-2532.

- **VALVERDE-REBAZA, J.**; VALEJO, A.; BERTON, L.; FALEIROS, T.; LOPES, A.. *A Naïve Bayes model based on overlapping groups for link prediction in online social networks*. In Proceedings of the 30th Annual ACM Symposium On Applied Computing (SAC' 15), ACM, 2015, p. 1136-1141.

- **VALVERDE-REBAZA, J.**; LOPES, A.. *Link prediction in online social networks using group information*[4]. In: Proceedings of the 14th International Conference on Computational Science and Its Applications (ICCSA 2014), Lecture Notes in Computer Science, Part VI, Springer, 2014, v. 8584, p. 31-45.

- **VALVERDE-REBAZA, J.**; LOPES, A.. *Exploiting behaviors of communities of Twitter users for link prediction*. Social Network Analysis and Mining, Springer Vienna, v. 3, n. 4, p. 1063-1074, 2013.

Other conference papers and journal articles have been published as a result of collaborations in related projects:

- DRURY, B.; **VALVERDE-REBAZA, J.**; MOURA, M. F.; LOPES, A.. *A Survey of the Applications of Bayesian Networks in Agriculture*. Engineering Applications of Artificial Intelligence, v. 65, p. 29-42, 2017.

- BERTON, L.; FALEIROS, T.; VALEJO, A.; **VALVERDE-REBAZA, J.**; LOPES, A.. *RGCLI: Robust graph that considers labeled instances for semi-supervised learning*. Neurocomputing, v. 226, p. 238-248, 2017.

- BERTON, L.; VEGA-OLIVEROS, D.; **VALVERDE-REBAZA, J.**; SILVA, A. T.; LOPES, A.. *Network sampling based on centrality measures for relational classification*. In Communications in Computer and Information Science. Cham: Springer International Publishing, 2017. Revised Selected Papers SIMBig 2015 and 2016, v. 656, p. 43–56.

- BERTON, L.; VEGA-OLIVEROS, D.; **VALVERDE-REBAZA, J.**; SILVA, A. T.; LOPES, A.. *The impact of network sampling on relational classification*. In Proceedings of The 3rd Annual International Symposium on Information Management and Big Data (SIMBig 2016), 2016, p. 62-72.

[4] Received the best paper award.

- DRURY, B.; **VALVERDE-REBAZA, J.**; LOPES, A.. *Causation generalization through the identification of equivalent nodes in causal sparse graphs constructed from text using node similarity strategies*. In Proceedings of The 2nd Annual International Symposium on Information Management and Big Data (SIMBig 2015), Web and Text Intelligence (WTI), 2015, p. 58-65.

- BERTON, L.; **VALVERDE-REBAZA, J.**; LOPES, A.. *Link prediction in graph construction for supervised and semi-supervised learning*. In Proceedings of The 2015 International Joint Conference on Neural Networks (IJCNN 2015), IEEE, 2015, p. 1818-1825.

- CHERMAN, E.; SPOLAÔR, N.; **VALVERDE-REBAZA, J.**; MONARD, M. C.. *Lazy multi-label learning algorithms based on mutuality strategies*. Journal of Intelligent & Robotic Systems, v. 80, n. 1, p. 261-276, 2015.

- **VALVERDE-REBAZA, J.**; SORIANO, A.; BERTON, L.; OLIVEIRA, M. C. F. de; LOPES, A.. *Music genre classification using traditional and relational approaches*. In Proceedings of 2014 Brazilian Conference on Intelligent Systems (BRACIS 2014), IEEE, 2014, p. 259-264.

- VALEJO, A.; **VALVERDE-REBAZA, J.**; LOPES, A.. *A multilevel approach for overlapping community detection*. In Proceedings of 2014 Brazilian Conference on Intelligent Systems (BRACIS 2014), IEEE, 2014, p. 390-395.

- DRURY, B.; CARDOSO, P.; **VALVERDE-REBAZA, J.**; VALEJO, A.; PEREIRA, F.; LOPES, A.. *An open source tool for crowd-sourcing the manual annotation of texts*. In Computational Processing of the Portuguese Language, Springer International Publishing, 2014, Lecture Notes in Computer Science, v. 8775. p. 268-273.

- VALEJO, A.; DRURY, B.; **VALVERDE-REBAZA, J.**; LOPES, A.. *Identification of related brazilian portuguese verb groups using overlapping community detection*. In Computational Processing of the Portuguese Language, Springer International Publishing, 2014, Lecture Notes in Computer Science, v. 8775, p. 292-297.

- VALEJO, A.; **VALVERDE-REBAZA, J.**; DRURY, B.; LOPES, A.. *Multilevel refinement based on neighborhood similarity*. In Proceedings of the 18th International Database Engineering & Applications Symposium (IDEAS '14), ACM, 2014, p. 67-76.

- CHERMAN, E.; SPOLAÔR, N.; **VALVERDE-REBAZA, J.**; MONARD, M. C.. *Algoritmos de Aprendizado Baseado em Grafos para Classificação Multirrótulo*. In Encontro Nacional de Inteligência Artificial e Computacional (ENIAC'2013), Sociedade Brasileira de ComputaComputação, 2013. p. 1-12.

Moreover, one paper was submitted to a journal and is in review process:

- **VALVERDE-REBAZA, J.**; ROCHE, M.; PONCELET, P.; LOPES, A.. *Friendship Prediction in Location-Based Social Networks: A Comparative Survey*. Submitted to: Social Networks, p. 1-36, July 2017.

Finally, three journal papers are in final preparation and must be submitted in the near future:

- **VALVERDE-REBAZA, J.**; LOPES, A.. *Location Prediction based on Social Strength and the Naïve Bayes Model*. To be submitted to: Knowledge-Based Systems.
- **VALVERDE-REBAZA, J.**; LOPES, A.. *Advances and Perspectives of Link Prediction: A Survey*. To be submitted to: Social Networks.
- **VALVERDE-REBAZA, J.**; LOPES, A.. *A Complete Survey of Link Mining*. To be submitted to: ACM Computing Surveys.

4.3 Limitations

One of the main limitations in our research is related to the data. Although a variety of network datasets data are publicly available, most of them are restricted to show only the linkage information, i.e. the existing links in the network. This might be sufficient for a conventional research in social networks and link prediction issues. However, as shown through this thesis, our research focuses on the use of more information. Thus, for our purpose we need user's group membership information. On the other hand, location-based social network datasets provide location information but lack linkage information, which is important for our research related to friendship prediction using location and social strength information in location-based social networks. It is important to note that both information sources necessary for our research are not unusual in the context of real-world social networks, but because there is few research exploring them, there are also few datasets making available such information resources.

The limited number of available datasets also limited the experiments carried out in this thesis. Crawling network data from online social networking sites demand more time and is not a trivial task, since it is necessary implement an amount of techniques for obtain coherent data as well as get permissions to crawl data from well-known sites, such as Facebook. In this scenario, other important limitation in our research is that the conclusions obtained from the results are indicative, but not assertive.

Other minor limitations can be observed in relation to our proposals. For instance, we performed experiments considering all the existing social groups, but not considering only some specifically selected, e.g. considering only social groups formed by a certain number of users. This consideration might be important to test the sensibility of our proposals for link prediction

using social group information. Other limitation is that none of our proposals for link prediction consider, or have been adapted to, the presence of weighted links.

4.4 Future Work

As future work, we intend to investigate the possibility of extend our proposals for friendship prediction in location-based social networks to perform the location prediction task. Moreover, since most of the efforts of community research focus on face the location prediction task using only location information, it would be interesting to instance the Naïve Bayes model, or other probabilistic frameworks, considering location and other information sources.

Another interesting future investigation is on exploring new issues of link prediction in location-based social networks, specifically by predicting location-location links, i.e. the prediction of missing parts of a trajectory. This future work could contribute with trajectory mining research.

We also intend to investigate the usage of our proposals in the context of other types of complex networks, specifically on biological networks. This intention is justified because some biological networks, such as disease interaction networks, show naturally presence of overlapping groups.

Finally, we would investigate the consideration of other types of user interactions in the form of user grouping. For instance, instead of only consider the explicit membership of a user in a group, we can consider implicit groups, such as the group of users reacting with the same sentiment for similar contents, group of users performing similar trajectories, group of user with similar social behaviors, among others.

BIBLIOGRAPHY

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the web. **Social Networks**, v. 25, n. 3, p. 211–230, 2003. Citation on page 71.

ADAMIC, L. A.; LUKOSE, R. M.; HUBERMAN, B. A. Local search in unstructured networks. In: ____. **Handbook of Graphs and Networks: From the Genome to the Internet**. [S.l.]: Wiley-VCH Verlag GmbH & Co. KGaA, 2005. chap. 13, p. 295–317. Citation on page 57.

AHMED, C.; ELKORANY, A. Enhancing link prediction in twitter using semantic user attributes. In: **Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015**. New York, NY, USA: ACM, 2015. (ASONAM '15), p. 1155–1161. ISBN 978-1-4503-3854-7. Citations on pages 27, 60, and 86.

AHUJA, G. Collaboration networks, structural holes, and innovation: A longitudinal study. **Administrative Science Quarterly**, v. 45, n. 3, p. 425–455, 2000. Citation on page 77.

AIELLO, L. M.; BARRAT, A.; SCHIFANELLA, R.; CATTUTO, C.; MARKINES, B.; MENCZER, F. Friendship prediction and homophily in social media. **ACM Trans. Web**, ACM, v. 6, n. 2, p. 9:1–9:33, 2012. Citation on page 76.

AIELLO, W.; CHUNG, F.; LU, L. Random evolution in massive graphs. In: ____. **Handbook of Massive Data Sets**. Boston, MA: Springer US, 2002. p. 97–122. Citation on page 57.

AKCORA, C. G.; CARMINATI, B.; FERRARI, E. User similarities on social networks. **Social Network Analysis and Mining**, v. 3, n. 3, p. 475–495, 2013. Citations on pages 69, 73, and 74.

ALAM, M.; KHAN, M.; VULLIKANTI, A.; MARATHE, M. An efficient and scalable algorithmic method for generating large scale random graphs. In: **Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis**. [S.l.]: IEEE Press, 2016. (SC '16), p. 32:1–32:12. ISBN 978-1-4673-8815-3. Citation on page 62.

ALAVIJEH, Z. Z. The application of link mining in social network analysis. **Advances in Computer Science: an International Journal (ACSIJ)**, v. 4, n. 15, p. 64–69, 2015. Citation on page 61.

ALHARBI, B.; ZHANG, X. Exploring the significance of human mobility patterns in social link prediction. In: **Proceedings of the 29th Annual ACM Symposium on Applied Computing**. [S.l.]: ACM, 2014. (SAC '14), p. 604–609. Citation on page 89.

ALI, Z. I. A comprehensive survey of link mining and anomalies detection. **Computer Science & Information Technology (CS & IT)**, v. 6, p. 175–189, 2016. Citation on page 61.

ALLALI, O.; MAGNIEN, C.; LATAPY, M. Internal link prediction: A new approach for predicting links in bipartite graphs. **Intell. Data Anal.**, IOS Press, v. 17, n. 1, p. 5–25, 2013. ISSN 1088-467X. Citations on pages 64, 66, and 162.

ALLAMANIS, M.; SCELLATO, S.; MASCOLO, C. Evolution of a location-based online social network: Analysis and models. In: **Proceedings of the 2012 ACM Conference on Internet Measurement Conference**. [S.l.]: ACM, 2012. (IMC '12), p. 145–158. ISBN 978-1-4503-1705-4. Citations on pages 27, 95, and 163.

ALMEIDA, L. J.; LOPES, A. An ultra-fast modularity-based graph clustering algorithm. In: **Proceedings 14th Portuguese Conference on Artificial Intelligence (EPIA) - Web and Network Intelligence Track**. [S.l.: s.n.], 2009. p. 1–9. Citation on page 110.

AMARAL, L. A.; SCALA, A.; BARTHELEMY, M.; STANLEY, H. E. Classes of small-world networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 97, n. 21, p. 11149–11152, 2000. ISSN 0027-8424. Citation on page 57.

AMERSHI, S.; FOGARTY, J.; WELD, D. Regroup: Interactive machine learning for on-demand group creation in social networks. In: **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. [S.l.]: ACM, 2012. (CHI '12), p. 21–30. ISBN 978-1-4503-1015-4. Citation on page 123.

ANDERSON, A.; HUTTENLOCHER, D.; KLEINBERG, J.; LESKOVEC, J. Effects of user similarity in social media. In: **Proceedings of the Fifth ACM International Conference on Web Search and Data Mining**. [S.l.]: ACM, 2012. (WSDM '12), p. 703–712. ISBN 978-1-4503-0747-5. Citation on page 69.

ANIL, A.; KUMAR, D.; SHARMA, S.; SINGHA, R.; SARMAH, R.; BHATTACHARYA, N.; SINGH, S. R. Link prediction using social network analysis over heterogeneous terrorist network. In: **2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)**. [S.l.: s.n.], 2015. p. 267–272. Citations on pages 26, 27, 60, and 86.

ANIL, A.; SETT, N.; SINGH, S. R. Modeling evolution of a social network using temporal graph kernels. In: **Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval**. [S.l.]: ACM, 2014. (SIGIR '14), p. 1051–1054. ISBN 978-1-4503-2257-7. Citations on pages 86 and 107.

ARNOLD, A.; COHEN, W. W. Information extraction as link prediction: Using curated citation networks to improve gene detection. In: **Proceedings of the 4th International Conference on Wireless Algorithms, Systems, and Applications**. Berlin, Heidelberg: Springer-Verlag, 2009. (WASA '09), p. 541–550. ISBN 978-3-642-03416-9. Citations on pages 27, 60, and 86.

ARTHUR, D.; MOTWANI, R.; SHARMA, A.; XU, Y. Pricing strategies for viral marketing on social networks. In: **Proceedings of the 5th International Workshop on Internet and Network Economics**. Berlin, Heidelberg: Springer-Verlag, 2009. (WINE '09), p. 101–112. ISBN 978-3-642-10840-2. Citation on page 58.

ASGHARI, M.; EMRICH, T.; DEMIRYUREK, U.; SHAHABI, C. Probabilistic estimation of link travel times in dynamic road networks. In: **Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems**. [S.l.]: ACM, 2015. (SIGSPATIAL '15), p. 47:1–47:10. Citation on page 88.

ASRATIAN, A. S.; DENLEY, T. M. J.; HÄGGKVIST, R. **Bipartite Graphs and their Applications, 1 edition**. [S.l.]: Cambridge University Press, 1998. (Cambridge Tracts in Mathematics). Citation on page 37.

BACKSTROM, L.; KLEINBERG, J. Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on facebook. In: **Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing**. [S.l.]: ACM, 2014. (CSCW '14), p. 831–841. Citations on pages 28 and 76.

BACKSTROM, L.; SUN, E.; MARLOW, C. Find me if you can: Improving geographical prediction with social and spatial proximity. In: **Proceedings of the 19th International Conference on World Wide Web**. [S.l.]: ACM, 2010. (WWW '10), p. 61–70. ISBN 978-1-60558-799-8. Citations on pages 103 and 163.

BAGCI, H.; KARAGOZ, P. Context-aware friend recommendation for location based social networks using random walk. In: **Proceedings of the 25th International Conference Companion on World Wide Web**. [S.l.]: International World Wide Web Conferences Steering Committee, 2016. (WWW '16 Companion), p. 531–536. ISBN 978-1-4503-4144-8. Citation on page 88.

BAHABADI, M. D.; GOLPAYEGANI, S. A. H.; ESMAEILI, L. A novel c2c e-commerce recommender system based on link prediction: Applying social network analysis. **CoRR**, abs/1407.8365, 2014. Citations on pages 27 and 86.

BALL, P. **Critical Mass**. London: Arrow Books Ltd, 2005. Citation on page 32.

BAO, J.; ZHENG, Y.; WILKIE, D.; MOKBEL, M. Recommendations in location-based social networks: A survey. **Geoinformatica**, Kluwer Academic Publishers, v. 19, n. 3, p. 525–565, Jul. 2015. ISSN 1384-6175. Citations on pages 88, 97, and 103.

BARABÁSI, A.-L. Taming complexity. **Nature Physics**, v. 1, p. 68–70, 2005. Citation on page 32.

BARABÁSI, A.-L. **Network Science**. [S.l.]: Cambridge University Press, 2016. Citations on pages 32, 33, 47, 51, and 108.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science 286**, n. 5439, p. 509–512, 1999. Citations on pages 26, 46, 49, 50, and 71.

BARABÁSI, A.-L.; JEONG, H.; NÉDA, Z.; RAVASZ, E.; SCHUBERT, A.; VICSEK, T. Evolution of the social network of scientific collaborations. **Physica A: Statistical Mechanics and its Applications**, v. 311, n. 3–4, p. 590 – 614, 2002. Citation on page 58.

BARBER, M. J.; CLARK, J. W. Detecting network communities by propagating labels under constraints. **Physical Review E**, v. 80, n. 2, p. 026129, 2009. Citations on pages 110 and 111.

BARBIERI, N.; BONCHI, F.; MANCO, G. Who to follow and why: Link prediction with explanations. In: **Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2014. (KDD '14), p. 1266–1275. ISBN 978-1-4503-2956-9. Citations on pages 27, 60, and 86.

BARGIGLI, L.; IASIO, G. di; INFANTE, L.; LILLO, F.; PIEROBON, F. The multiplex structure of interbank networks. **Quantitative Finance**, v. 15, n. 4, p. 673–691, 2015. Citation on page 58.

BARTEL, J. W.; DEWAN, P. Evolving friend lists in social networks. In: **Proceedings of the 7th ACM Conference on Recommender Systems**. [S.l.]: ACM, 2013. (RecSys '13), p. 435–438. ISBN 978-1-4503-2409-0. Citation on page 123.

BARTHELEMY, M.; AMARAL, L. Small-world networks: Evidence for a crossover picture. **Phys. Rev. Lett**, n. 82, p. 3180–3183, 1999. Citation on page 48.

BATTISTON, S.; RODRIGUES, J. F.; ZEYTINOGLU, H. The network of inter-regional direct investment stocks across Europe. **Advances in Complex Systems**, v. 10, n. 01, p. 29–51, 2007. Citation on page 58.

BAYRAK, A. E.; POLAT, F. Contextual feature analysis to improve link prediction for location based social networks. In: **Proceedings of the 8th Workshop on Social Network Mining and Analysis**. [S.l.]: ACM, 2014. (SNAKDD'14), p. 7:1–7:5. ISBN 978-1-4503-3192-0. Citations on pages 28, 97, 98, 100, 101, 102, 154, 162, and 163.

____. Examining place categories for link prediction in location based social networks. In: **2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**. [S.l.: s.n.], 2016. p. 976–979. Citations on pages 28, 97, and 154.

BEDI, P.; SHARMA, C. Community detection in social networks. **WIREs Data Mining Knowl Discov**, v. 6, p. 115–135, 2016. Citations on pages 62 and 110.

BENCHETTARA, N.; KANAWATI, R.; ROUVEIROL, C. A supervised machine learning link prediction approach for academic collaboration recommendation. In: **Proceedings of RecSys '10**. [S.l.: s.n.], 2010. p. 253–256. ISBN 978-1-60558-906-0. Citations on pages 67, 82, and 86.

BERNSTEIN, A.; CLEARWATER, S.; PROVOST, F. The relational vector-space model and industry classification. In: **In Proceedings of IJCAI Workshop on Statistical Models from Relational Data**. [S.l.: s.n.], 2003. p. 8–18. Citation on page 58.

BERNSTEIN, M. S.; MARCUS, A.; KARGER, D. R.; MILLER, R. C. Enhancing directed content sharing on the web. In: **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. [S.l.]: ACM, 2010. (CHI '10), p. 971–980. ISBN 978-1-60558-929-9. Citation on page 123.

BERTON, L.; FALEIROS, T. de P.; VALEJO, A.; VALVERDE-REBAZA, J.; LOPES, A. Rgcli: Robust graph that considers labeled instances for semi-supervised learning. **Neurocomputing**, v. 226, p. 238 – 248, 2017. Citations on pages 62 and 178.

BERTON, L.; VALVERDE-REBAZA, J.; LOPES, A. Link prediction in graph construction for supervised and semi-supervised learning. In: **Proceedings of The 2015 International Joint Conference on Neural Networks**. [S.l.]: IEEE, 2015. (IJCNN 2015), p. 1818–1825. Citations on pages 86, 105, and 178.

BERTON, L.; VEGA-OLIVEROS, D.; VALVERDE-REBAZA, J.; SILVA, A. da; LOPES, A. The impact of network sampling on relational classification. In: **Proceedings of The 3rd Annual International Symposium on Information Management and Big Data**. [S.l.]: CEUR-WS.org, 2016. (SIMBig 2016), p. 62–72. Citation on page 178.

BERTON, L.; VEGA-OLIVEROS, D.; VALVERDE-REBAZA, J.; SILVA, A. T. da; LOPES, A. Network sampling based on centrality measures for relational classification. In: ____. **Communications in Computer and Information Science**. Cham: Springer International Publishing, 2017. (Information Management and Big Data: Second Annual International Symposium (SIMBig 2015) and Third Annual International Symposium (SIMBig 2016), Revised Selected Papers, v. 656), p. 43–56. Citation on page 178.

BHATTACHARYYA, P.; GARG, A.; WU, S. F. Analysis of user keyword similarity in online social networks. **Social Network Analysis and Mining**, v. 1, n. 3, p. 143–158, 2011. Citations on pages 69 and 88.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. [S.l.]: Springer-Verlag New York, Inc., 2006. Citation on page 25.

BISWAS, A.; BISWAS, B. Community-based link prediction. **Multimedia Tools and Applications**, p. 1–21, 2017. Citations on pages 28, 78, 108, and 111.

BLISS, C. A.; FRANK, M. R.; DANFORTH, C. M.; DODDS, P. S. An evolutionary algorithm approach to link prediction in dynamic social networks. **Journal of Computational Science**, v. 5, n. 5, p. 750–764, 2014. Citation on page 83.

BLONDEL, V. D.; GAJARDO, A.; HEYMANS, M.; SENELLART, P.; DOOREN, P. V. A measure of similarity between graph vertices: Applications to synonym extraction and web searching. **SIAM Rev.**, Society for Industrial and Applied Mathematics, v. 46, n. 4, p. 647–666, 2004. ISSN 0036-1445. Citation on page 75.

BLUM, C.; ROLI, A. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. **ACM Comput. Surv.**, ACM, v. 35, n. 3, p. 268–308, Sep. 2003. ISSN 0360-0300. Citation on page 83.

BOCCALETTI, S.; BIANCONI, G.; CRIADO, R.; GENIO, C. del; GóMEZ-GARDEñES, J.; ROMANCE, M.; SENDIñA-NADAL, I.; WANG, Z.; ZANIN, M. The structure and dynamics of multilayer networks. **Physics Reports**, v. 544, n. 1, p. 1–122, 2014. Citation on page 60.

BOLLOBÁS, B. **Modern Graph Theory**. [S.l.]: Springer-Verlag New York, 1998. Citation on page 33.

____. **Random Graphs**. [S.l.]: Cambridge University Press, 2001. Citation on page 47.

BORGATTI, S. P.; EVERETT, M. G. Models of core/periphery structures. **Social Networks**, v. 21, n. 4, p. 375 – 395, 2000. Citation on page 47.

BOSS, M.; ELSINGER, H.; SUMMER, M.; THURNER, S. Network topology of the interbank market. **Quantitative Finance**, v. 4, n. 6, p. 677–684, 2004. Citation on page 58.

BOUILLOT, F.; PONCELET, P.; ROCHE, M. How and why exploit tweet's location information? In: **AGILE'2012: 15th International Conference on Geographic Information Science**. [S.l.: s.n.], 2012. Citation on page 89.

BRAGA, R. B.; TAHIR, A.; BERTOLOTTO, M.; MARTIN, H. A multi-layer data representation of trajectories in social networks based on points of interest. In: **Proceedings of the Twelfth International Workshop on Web Information and Data Management**. [S.l.]: ACM, 2012. (WIDM '12), p. 19–26. ISBN 978-1-4503-1720-7. Citation on page 88.

BRINGMANN, B.; BERLINGERIO, M.; BONCHI, F.; GIONIS, A. Learning and predicting the evolution of social networks. **IEEE Intelligent Systems**, p. 26–34, 2010. Citations on pages 86 and 107.

BRISABOA, N. R.; LADRA, S.; NAVARRO, G. Compact representation of web graphs with extended functionality. **Inf. Syst.**, Elsevier Science Ltd., Oxford, UK, UK, v. 39, p. 152–174, 2014. ISSN 0306-4379. Citation on page 57.

BRODER, A.; KUMAR, R.; MAGHOUL, F.; RAGHAVAN, P.; RAJAGOPALAN, S.; STATA, R.; TOMKINS, A.; WIENER, J. Graph structure in the web. **Comput. Netw.**, Elsevier North-Holland, Inc., v. 33, n. 1-6, p. 309–320, Jun. 2000. Citation on page 123.

BROWN, C.; NICOSIA, V.; SCELLATO, S.; NOULAS, A.; MASCOLO, C. The importance of being placefriends: Discovering location-focused online communities. In: **Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks**. [S.l.]: ACM, 2012. (WOSN '12), p. 31–36. Citation on page 97.

BURDA, Z.; DUDA, J.; LUCK, J. M.; WACLAW, B. Localization of the maximal entropy random walk. **Phys. Rev. Lett.**, American Physical Society, v. 102, p. 160602, 2009. Citation on page 75.

BURT, R. S. **Structural Holes: the Social Structure of Competition**. [S.l.]: Harvard University Press, 1992. Citation on page 77.

____. Structural holes versus network closure as social capital. In: **Social Capital: Theory and Research**. [S.l.]: Aldine de Gruyter, 2001. p. 31–56. Citation on page 77.

BYGRAVE, W. D. The structure of the investment networks of venture capital firms. **Journal of Business Venturing**, v. 3, n. 2, p. 137 – 157, 1988. ISSN 0883-9026. Citation on page 58.

CAIYAN, D.; CHEN, L.; LI, B. Link prediction in complex network based on modularity. **Soft Computing**, p. 1–18, 2016. Citations on pages 28, 78, and 111.

CALLAWAY, D.; NEWMAN, M.; STROGATZ, S.; WATTS, D. Network robustness and fragility: Percolation on random graphs. **Phys. Rev. Lett**, n. 85, p. 5468–5471, 2000. Citation on page 51.

CAMARINHA-MATOS, L. M.; AFSARMANESH, H. A comprehensive modeling framework for collaborative networked organizations. **Journal of Intelligent Manufacturing**, v. 18, n. 5, p. 529–542, 2007. Citation on page 58.

CANNISTRACI, C. V.; ALANIS-LOBATO, G.; RAVASI, T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. **Scientific Reports**, v. 3, p. 1613, 2013. Citations on pages 28, 78, 108, and 111.

CAO, X.; YU, Y. Joint user modeling across aligned heterogeneous sites. In: **Proceedings of the 10th ACM Conference on Recommender Systems**. [S.l.]: ACM, 2016. (RecSys '16), p. 83–90. ISBN 978-1-4503-4035-9. Citation on page 62.

CHAMANI, T.; POUREBRAHIMI, A.; SHIRAZI, B. Improving link prediction in social network with population-based metaheuristics algorithm. **International Journal of Mechatronics, Electrical and Computer Technology**, Austrian E-Journals of Universal Scientific Organization, v. 4, n. 12, p. 1202–1213, 2014. Citation on page 83.

CHANG, S.; KUMAR, V.; GILBERT, E.; TERVEEN, L. G. Specialization, homophily, and gender in a social curation site: Findings from pinterest. In: **Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing**. [S.l.]: ACM, 2014. (CSCW '14), p. 674–686. Citations on pages 28 and 76.

CHEBOTAREV, P.; SHAMIS, E. Matrix-forest theorems. **arXiv preprint math/0602575**, 2006. Citation on page 75.

CHEN, Q.; CHANG, H.; GOVINDAN, R.; JAMIN, S. The origin of power laws in internet topologies revisited. In: **Proceedings.Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies**. [S.l.: s.n.], 2002. v. 2, p. 608–617. Citation on page 56.

CHEN, X.; PANG, J.; XUE, R. Constructing and comparing user mobility profiles. **ACM Trans. Web**, ACM, v. 8, n. 4, p. 21:1–21:25, Nov. 2014. Citation on page 88.

CHEN, Z.; SHEN, H. T.; ZHOU, X. Discovering popular routes from trajectories. In: **Proceedings of the 2011 IEEE 27th International Conference on Data Engineering**. [S.l.]: IEEE Computer Society, 2011. (ICDE '11), p. 900–911. ISBN 978-1-4244-8959-6. Citation on page 103.

CHEN, Z.; SHEN, H. T.; ZHOU, X.; ZHENG, Y.; XIE, X. Searching trajectories by locations: An efficiency study. In: **Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data**. [S.l.]: ACM, 2010. (SIGMOD '10), p. 255–266. Citation on page 88.

CHENG, H.; AREFIN, M.; CHEN, Z.; MORIMOTO, Y. Place recommendation based on users check-in history for location-based services. **International Journal of Networking and Computing**, v. 3, n. 2, p. 228–243, 2013. ISSN 2185-2847. Citation on page 103.

CHENG, H.-M.; ZHANG, Z. Community detection based on link prediction methods. **CoRR abs/1611.00254**, 2016. Citation on page 86.

CHENG, R.; PANG, J.; ZHANG, Y. Inferring friendship from check-in data of location-based social networks. In: **Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015**. [S.l.]: ACM, 2015. (ASONAM '15), p. 1284–1291. Citation on page 88.

CHENG, Z.; CAVERLEE, J.; BARTHWAL, H.; BACHANI, V. Finding local experts on twitter. In: **Proceedings of the 23rd International Conference on World Wide Web**. [S.l.]: ACM, 2014. (WWW '14 Companion), p. 241–242. ISBN 978-1-4503-2745-9. Citation on page 97.

____. Who is the barbecue king of texas?: A geo-spatial approach to finding local experts on twitter. In: **Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval**. [S.l.]: ACM, 2014. (SIGIR '14), p. 335–344. ISBN 978-1-4503-2257-7. Citation on page 97.

CHENG, Z.; CAVERLEE, J.; LEE, K.; SUI, D. Exploring millions of footprints in location sharing services. In: **Proceedings of the Fifth International AAAI Conference on Web and Social Media**. [S.l.: s.n.], 2011. p. 81–88. Citations on pages 95 and 163.

CHERMAN, E.; SPOLAÔR, N.; VALVERDE-REBAZA, J.; MONARD, M. C. Algoritmos de Aprendizado Baseado em Grafos para Classificação Multirrótulo. In: **ENIAC'2013: Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.]: Sociedade Brasileira de Computação, 2013. p. 1–12. Citation on page 178.

____. Lazy multi-label learning algorithms based on mutuality strategies. **Journal of Intelligent & Robotic Systems**, v. 80, n. 1, p. 261–276, 2015. Citation on page 178.

CHIANG, K.-Y.; NATARAJAN, N.; TEWARI, A.; DHILLON, I. S. Exploiting longer cycles for link prediction in signed networks. In: **Proceedings of the 20th ACM International Conference on Information and Knowledge Management**. [S.l.]: ACM, 2011. (CIKM '11), p. 1157–1162. ISBN 978-1-4503-0717-8. Citation on page 81.

CHIANG, M.-F.; HOANG, T.-A.; LIM, E.-P. Where are the passengers?: A grid-based gaussian mixture model for taxi bookings. In: **Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems**. [S.l.]: ACM, 2015. (SIGSPATIAL '15), p. 32:1–32:10. Citation on page 88.

CHIKHAOUI, B.; WANG, S.; XIONG, T.; PIGOT, H. Pattern-based causal relationships discovery from event sequences for modeling behavioral user profile in ubiquitous environments. **Inf. Sci.**, Elsevier Science Inc., v. 285, n. C, p. 204–222, 2014. Citations on pages 27, 88, and 89.

CHILUKA, N.; ANDRADE, N.; POUWELSE, J. A link prediction approach to recommendations in large-scale user-generated content systems. In: ____. **Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 189–200. ISBN 978-3-642-20161-5. Citations on pages 27, 60, and 86.

CHO, E.; MYERS, S. A.; LESKOVEC, J. Friendship and mobility: User movement in location-based social networks. In: **ACM KDD**. [S.l.: s.n.], 2011. p. 1082–1090. ISBN 978-1-4503-0813-7. Citations on pages 27, 28, 88, 89, 93, 95, 96, 97, 153, 154, 160, and 163.

CHOI, T. Y.; DOOLEY, K. J.; RUNGTUSANATHAM, M. Supply networks and complex adaptive systems: control versus emergence. **Journal of Operations Management**, v. 19, n. 3, p. 351–366, 2001. Citation on page 58.

CHOI, T. Y.; HONG, Y. Unveiling the structure of supply networks: case studies in honda, acura, and daimlerchrysler. **Journal of Operations Management**, v. 20, n. 5, p. 469–493, 2002. Citation on page 58.

CHORLEY, M. J.; WHITAKER, R. M.; ALLEN, S. M. Personality and location-based social networks. **Computers in Human Behavior**, v. 46, p. 45–56, 2015. Citations on pages 27, 87, 89, 96, and 153.

CHOROMANSKI, K.; MATUSZAK, M.; MIEKISZ, J. Scale-free graph with preferential attachment and evolving internal vertex structure. **Journal of Statistical Physics**, v. 151, n. 6, p. 1175–1183, 2013. Citation on page 51.

CHUA, H. N.; SUNG, W.-K.; WONG, L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. **Bioinformatics**, Oxford University Press, v. 22, n. 13, p. 1623–1630, Jul. 2006. ISSN 1367-4803. Citation on page 73.

CHUNG, F. **Spectral Graph Theory**. [S.l.]: American Mathematical Society, 1997. Citation on page 46.

CINGOLANI, I.; PICCARDI, C.; TAJOLI, L. Discovering preferential patterns in sectoral trade networks. **PLoS ONE**, v. 10, n. 10, 2015. Citation on page 58.

CIOTTI, V.; BONAVENTURA, M.; NICOSIA, V.; PANZARASA, P.; LATORA, V. Homophily and missing links in citation networks. **EPJ Data Sci.**, v. 5, n. 1, p. 7, 2016. Citation on page 77.

CLAUSET, A.; MOORE, C.; NEWMAN, M. E. J. Hierarchical structure and the prediction of missing links in networks. **Nature**, v. 453, p. 98–101, 2008. Citations on pages 26, 27, 80, and 86.

CLAUSET, A.; NEWMAN, M.; MOORE, C. Finding community structure in very large networks. **Physical Review E**, p. 1–6, 2004. Citations on pages 51, 77, 108, 110, and 111.

COHEN, R.; EREZ, K.; AVRAHAM, D. ben; HAVLIN, S. Breakdown of the internet under intentional attack. **Phys. Rev. Lett**, n. 86, p. 3682–3685, 2001. Citation on page 51.

COSTA, L.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. **Advances in Physics**, v. 56, p. 167, 2007. Citations on pages 26, 33, and 47.

CRANSHAW, J.; TOCH, E.; HONG, J.; KITTUR, A.; SADEH, N. Bridging the gap between physical location and online social networks. In: **Proceedings of the 12th ACM International Conference on Ubiquitous Computing**. [S.l.]: ACM, 2010. (UbiComp '10), p. 119–128. ISBN 978-1-60558-843-8. Citations on pages 27, 28, 88, 89, 95, 96, 97, 98, 100, and 154.

CUI, A.-X.; YAN, F.; MING-SHENG, S.; DUAN-BING, C.; TAO, Z. Emergence of local structures in complex network: common neighborhood drives the network evolution. **Acta Physica Sinica**, v. 60, n. 3, p. 38901, 2011. Citations on pages 86 and 107.

CUKIERSKI, W.; HAMNER, B.; YANG, B. Graph-based features for supervised link prediction. In: **The 2011 International Joint Conference on Neural Networks**. [S.l.: s.n.], 2011. (IJCNN 2011), p. 1237–1244. Citation on page 82.

DAI, J.; YANG, B.; GUO, C.; DING, Z. Personalized route recommendation using big trajectory data. In: **31st IEEE International Conference on Data Engineering (ICDE 2015)**. [S.l.: s.n.], 2015. p. 543–554. Citation on page 103.

DALL'ASTA, L.; MARSILI, M.; PIN, P. Collaboration in social networks. **Proceedings of the National Academy of Sciences**, v. 109, n. 12, p. 4395–4400, 2012. Citation on page 58.

DAMINELLI, S.; THOMAS, J. M.; DURáN, C.; CANNISTRACI, C. V. Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. **New Journal of Physics**, v. 17, n. 11, p. 113037, 2015. Citations on pages 28, 78, and 111.

DAVIS, M.; LIU, W.; MILLER, P.; HUNTER, R.; KEE, F. Agwan: A generative model for labelled, weighted graphs. In: **Appice A., New Frontiers in Mining Complex Patterns, Lecture Notes in Computer Science**. [S.l.]: Springer, 2014. (NFMCP 2013, v. 8399). Citation on page 62.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **JMLR**, v. 7, p. 1–30, 2006. ISSN 1532-4435. Citations on pages 142, 148, and 167.

DEYNE, S. D.; G., S. Word associations: network and semantic properties. **Behav Res Methods**, v. 40, n. 1, p. 213–231, 2008. Citation on page 58.

DIAKONOVA, M.; NICOSIA, V.; LATORA, V.; MIGUEL, M. S. Irreducibility of multilayer network dynamics: the case of the voter model. **New Journal of Physics**, v. 18, n. 2, p. 023010, 2016. Citation on page 60.

DING, J.; JIAO, L.; WU, J.; LIU, F. Prediction of missing links based on community relevance and ruler inference. **Knowledge-Based Systems**, v. 98, p. 200–215, 2016. Citations on pages 28, 78, and 111.

DINH, T. N.; ZHANG, H.; NGUYEN, D. T.; THAI, M. T. Cost-effective viral marketing for time-critical campaigns in large-scale social networks. **IEEE/ACM Trans. Netw.**, IEEE Press, Piscataway, NJ, USA, v. 22, n. 6, p. 2001–2011, 2014. ISSN 1063-6692. Citation on page 58.

DIXON, S. J.; COSTANZO, M.; BARYSHNIKOVA, A.; ANDREWS, B.; BOONE, C. Systematic mapping of genetic interaction networks. **Annual Review of Genetics**, v. 43, n. 1, p. 601–625, 2009. Citation on page 56.

DODDS, P. S. **Geometry of river networks**. Phd Thesis (PhD Thesis) — Massachusetts Institute of Technology, Dept. of Mathematics, USA, 1969. Citation on page 57.

DODDS, P. S.; MUHAMAD, R.; WATTS, D. An experimental study of search in global social networks. **Science**, v. 301, p. 827–829, 2003. Citation on page 48.

DOMENICO, M. D.; NICOSIA, V.; ARENAS, A.; LATORA, V. Structural reducibility of multilayer networks. **Nature Communications**, v. 6, n. 6864, 2015. Citation on page 60.

DOMÍNGUEZ-MUJICA, J. **Global Change and Human Mobility**. [S.l.]: Springer Singapore, 2016. (Advances in Geographical and Environmental Sciences). ISBN 2198-3542. Citation on page 89.

DONG, E.; LI, J.; XIE, Z. Link prediction via convex nonnegative matrix factorization on multiscale blocks. **Journal of Applied Mathematics**, n. 786156, 2014. Citation on page 85.

DONG, Y.; KE, Q.; WANG, B.; WU, B. Link prediction based on local information. In: **Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining**. [S.l.]: IEEE Computer Society, 2011. (ASONAM '11), p. 382–386. ISBN 978-0-7695-4375-8. Citation on page 73.

DOREIAN, P.; BATAGELJ, V.; FERLIGOJ, A. **Generalized Blockmodeling**. [S.l.]: Cambridge University Press, 2005. (Structural Analysis in the Social Sciences). Citation on page 80.

DOYTSHER, Y.; GALON, B.; KANZA, Y. Querying geo-social data by bridging spatial networks and social networks. In: **Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks**. [S.l.]: ACM, 2010. (LBSN '10), p. 39–46. Citation on page 88.

_____. Storing routes in socio-spatial networks and supporting social-based route recommendation. In: **Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks**. [S.l.]: ACM, 2011. (LBSN '11), p. 49–56. ISBN 978-1-4503-1033-8. Citation on page 103.

DRABEK, T. E. **Managing Multiorganizational Emergency Responses: Emergent Search and Rescue Networks in Natural Disaster and Remote Area Settings**. [S.l.]: Univ. of Colorado Natural Hazards, 1981. (Program Environm). Citation on page 58.

DRURY, B.; CARDOSO, P.; VALVERDE-REBAZA, J.; VALEJO, A.; PEREIRA, F.; LOPES, A. An open source tool for crowd-sourcing the manual annotation of texts. In: **Computational Processing of the Portuguese Language**. [S.l.]: Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8775). p. 268–273. Citation on page 178.

DRURY, B.; VALVERDE-REBAZA, J.; LOPES, A. Causation generalization through the identification of equivalent nodes in causal sparse graphs constructed from text using node similarity strategies. In: **Proceedings of The 2nd Annual International Symposium on Information Management and Big Data**. [S.l.]: CEUR-WS.org, 2015. (SIMBig 2015 - Web and Text Intelligence (WTI)), p. 58–65. Citations on pages 86, 105, and 178.

DRURY, B.; VALVERDE-REBAZA, J.; MOURA, M.-F.; LOPES, A. A survey of the applications of bayesian networks in agriculture. **Engineering Applications of Artificial Intelligence**, v. 65, p. 29–42, 2017. Citation on page 178.

EAGLE, N.; MONTJOYE, Y. A. de; BETTENCOURT, L. M. A. Community computing: Comparisons between rural and urban societies using mobile phone data. In: **2009 International Conference on Computational Science and Engineering**. [S.l.: s.n.], 2009. v. 4, p. 144–150. Citation on page 88.

EAGLE, N.; PENTLAND, A.; LAZER, D. Mobile phone data for inferring social network structure. In: ____. **Social Computing, Behavioral Modeling, and Prediction**. Boston, MA: Springer US, 2008. p. 79–88. Citations on pages 87 and 88.

ERDÖS, P.; RÉNYI, A. On random graphs. **Publications Mathematicae 6**, p. 290–297, 1959. Citations on pages 26, 46, and 47.

____. On the evolution of random graphs. In: **Publications of the Mathematical Institute of the Hungarian Academy of Sciences**. [S.l.: s.n.], 1960. v. 5, p. 17–61. Citation on page 47.

____. On the strength of connectedness of a random graph. **Acta Mathematica Scientia Hungary**, v. 12, p. 261–267, 1961. Citation on page 47.

ESLAMI, M.; ALEYASEN, A.; MOGHADDAM, R. Z.; KARAHALIOS, K. G. Evaluation of automated friend grouping in online social networks. In: **CHI '14 Extended Abstracts on Human Factors in Computing Systems**. [S.l.]: ACM, 2014. (CHI EA '14), p. 2119–2124. Citation on page 124.

ESLAMI, M.; ALEYASEN, A.; MOGHADDAM, R. Z.; KARAHALIOS, K. Friend grouping algorithms for online social networks: Preference, bias, and implications. In: ____. **Proceedings of 6th International Conference on Social Informatics, SocInfo 2014**. Cham: Springer International Publishing, 2014. p. 34–49. Citations on pages 27 and 124.

ESSEN, U.; STEINBISS, V. Cooccurrence smoothing for stochastic language modeling. In: **Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1**. [S.l.]: IEEE Computer Society, 1992. (ICASSP'92), p. 161–164. Citation on page 85.

ESSLIMANI, I.; BRUN, A.; BOYER, A. Densifying a behavioral recommender system by social networks link prediction methods. **Social Network Analysis and Mining**, Springer Vienna, v. 1, p. 159–172, 2011. Citation on page 27.

FAGIOLO, G.; REYES, J.; SCHIAVO, S. On the topological properties of the world trade web: A weighted network analysis. **Physica A: Statistical Mechanics and its Applications**, v. 387, n. 15, p. 3868 – 3873, 2008. Citation on page 58.

____. World-trade web: Topological properties, dynamics, and evolution. **Phys. Rev. E**, American Physical Society, v. 79, p. 036115, Mar 2009. Citation on page 58.

FALEIROS, T. de P.; ROSSI, R. G.; LOPES, A. Optimizing the class information divergence for transductive classification of texts using propagation in bipartite graphs. **Pattern Recognition Letters**, v. 87, p. 127 – 138, 2017. Advances in Graph-based Pattern Recognition. Citation on page 62.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. **SIGCOMM Comput. Commun. Rev.**, ACM, v. 29, n. 4, p. 251–262, Aug. 1999. ISSN 0146-4833. Citation on page 56.

FARALLI, S.; STILO, G.; VELARDI, P. Large scale homophily analysis in twitter using a twixonomy. In: **Proceedings of the 24th International Conference on Artificial Intelligence**. [S.l.]: AAAI Press, 2015. (IJCAI'15), p. 2334–2340. Citations on pages 28 and 76.

FATOURECHI, M.; WARD, R. K.; MASON, S. G.; HUGGINS, J.; SCHLOGL, A.; BIRCH, G. E. Comparison of evaluation metrics in classification applications with imbalanced datasets. In: **Seventh International Conference on Machine Learning and Applications**. [S.l.: s.n.], 2008. (ICMLA '08), p. 777–782. Citation on page 65.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Advances in knowledge discovery and data mining. In: . [S.l.]: American Association for Artificial Intelligence, 1996. chap. From Data Mining to Knowledge Discovery: An Overview, p. 1–34. Citations on pages 25 and 60.

FELLEGARA, R.; FUGACCI, U.; IURICICH, F.; FLORIANI, L. D. Analysis of geolocalized social networks based on simplicial complexes. In: **Proceedings of the 9th ACM SIGSPATIAL International Workshop on Location-Based Social Networks**. [S.l.]: ACM, 2016. (LBSN '16). Citations on pages 87 and 93.

FENG, X.; ZHAO, J. C.; XU, K. Link prediction in complex networks: a clustering perspective. **The European Physical Journal B**, v. 85, n. 1, p. 3, 2012. Citations on pages 28, 77, 108, and 111.

FENG, Z.; ZHU, Y. A survey on trajectory data mining: Techniques and applications. **IEEE Access**, v. 4, p. 2056–2067, 2016. Citation on page 91.

FERRARI, L.; ROSI, A.; MAMEI, M.; ZAMBONELLI, F. Extracting urban patterns from location-based social networks. In: **Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks**. [S.l.]: ACM, 2011. (LBSN '11), p. 9–16. Citation on page 89.

FIRE, M.; TENENBOIM-CHEKINA, L.; PUZIS, R.; LESSER, O.; ROKACH, L.; ELOVICI, Y. Computationally efficient link prediction in a variety of social networks. **ACM Trans. Intell. Syst. Technol.**, ACM, v. 5, n. 1, p. 10:1–10:25, 2014. ISSN 2157-6904. Citation on page 82.

FIRE, M.; TENENBOIM, L.; LESSER, O.; PUZIS, R.; ROKACH, L.; ELOVICI, Y. Link prediction in social networks using computationally efficient topological features. In: **2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing**. [S.l.: s.n.], 2011. p. 73–80. Citation on page 82.

FORTUNATO, S. Community detection in graphs. **Physics Reports**, v. 486, n. 3-5, p. 75–174, 2010. Citations on pages 77, 108, and 109.

FOUSS, F.; PIROTTE, A.; RENDERS, J.-M.; SAERENS, M. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. **IEEE Trans. on Knowl. and Data Eng.**, IEEE Educational Activities Department, v. 19, n. 3, p. 355–369, Mar. 2007. ISSN 1041-4347. Citation on page 75.

FREEMAN, L. C. A Set of Measures of Centrality Based on Betweenness. **Sociometry**, American Sociological Association, v. 40, n. 1, p. 35–41, 1977. Citation on page 51.

GALLAGHER, B.; TONG, H.; ELIASSI-RAD, T.; FALOUTSOS, C. Using ghost edges for classification in sparsely labeled networks. In: **Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2008. (KDD '08), p. 256–264. Citation on page 86.

GANTZ, J. F.; REINSEL, D. **The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east**. 2012. External Publication of IDC (Analyse the Future) Information and Data. Citation on page 25.

GAO, H.; LIU, H. **Mining Human Mobility in Location-Based Social Networks**. [S.l.]: Morgan & Claypool, 2015. Citations on pages 87 and 89.

GETOOR, L.; DIEHL, C. P. Link mining: a survey. **ACM SIGKDD Explorations Newsletter**, ACM Press, v. 7, n. 2, p. 3–12, 2005. Citations on pages 26, 60, and 61.

GETOOR, L.; FRIEDMAN, N.; KOLLER, D.; TASKAR, B. Learning probabilistic models of link structure. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 679–707, Mar. 2003. ISSN 1532-4435. Citation on page 62.

GIMENES, G. P.; GUALDRON, H.; RADDO, T. R.; RODRIGUES, J. F. Supervised-learning link recommendation in the dblp co-authoring network. In: **2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)**. [S.l.: s.n.], 2014. p. 563–568. Citation on page 82.

GIRVAN, M.; NEWMAN, M. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 12, p. 7821–7826, 2002. Citations on pages 47, 51, and 110.

GOH, K.-I.; CUSICK, M. E.; VALLE, D.; CHILDS, B.; VIDAL, M.; BARABáSI, A.-L. The human disease network. **Proceedings of the National Academy of Sciences**, v. 104, n. 21, p. 8685–8690, 2007. Citation on page 56.

GOLDBERG, S. R.; ANTHONY, H.; EVANS, T. S. Modelling citation networks. **Scientometrics**, v. 105, n. 3, p. 1577–1604, 2015. Citation on page 57.

GONG, N. Z.; TALWALKAR, A.; MACKEY, L.; HUANG, L.; SHIN, E. C. R.; STEFANOV, E.; SHI, E. R.; SONG, D. Joint link prediction and attribute inference using a social-attribute network. **ACM Trans. Intell. Syst. Technol.**, ACM, v. 5, n. 2, p. 27:1–27:20, Apr. 2014. ISSN 2157-6904. Citation on page 69.

GOODWIN, H. B. The haversine in nautical astronomy. **Naval Institute Proceedings**, v. 36, n. 3, p. 735–746, 1910. Citation on page 95.

GRABOWICZ, P. A.; RAMASCO, J. J.; GONÇALVES, B.; EGUÍLUZ, V. M. Entangling mobility and interactions in social media. **PLoS ONE**, v. 9, n. 3, p. 1–12, 2014. Citations on pages 27, 88, 95, and 163.

GRANOVETTER, M. The impact of social structure on economic outcomes. **The Journal of Economic Perspectives**, American Economic Association, v. 19, n. 1, p. 33–50, 2005. Citations on pages 76 and 77.

GREENWOOD, M. J. Research on internal migration in the united states: A survey. **Journal of Economic Literature**, American Economic Association, v. 13, n. 2, p. 397–433, 1975. Citation on page 89.

GROH, G.; STRAUB, F.; EICHER, J.; GROB, D. Geographic aspects of tie strength and value of information in social networking. In: **Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks**. [S.l.]: ACM, 2013. (LBSN '13). Citation on page 87.

GUIMERÀ, R.; SALES-PARDO, M. Missing and spurious interactions and the reconstruction of complex networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 52, p. 22073–22078, 2009. Citations on pages 81 and 86.

GUNAWARDANA, A.; SHANI, G. A survey of accuracy evaluation metrics of recommendation tasks. **JMLR**, JMLR.org, v. 10, p. 2935–2962, 2009. ISSN 1532-4435. Citation on page 170.

GÜNEŞ, İ.; GÜNDÜZ-ÖĞÜDÜCÜ, Ş.; ÇATALTEPE, Z. Link prediction using time series of neighborhood-based node similarity scores. **Data Mining and Knowledge Discovery**, v. 30, n. 1, p. 147–180, 2016. Citation on page 67.

HADIAN, A.; NOBARI, S.; MINAEI-BIDGOLI, B.; QU, Q. Roll: Fast in-memory generation of gigantic scale-free networks. In: **Proceedings of the 2016 International Conference on Management of Data**. [S.l.]: ACM, 2016. (SIGMOD '16), p. 1829–1842. ISBN 978-1-4503-3531-7. Citation on page 62.

HAI, P. N.; IENCO, D.; PONCELET, P.; TEISSEIRE, M. Extracting Trajectories through an Efficient and Unifying Spatio-Temporal Patten Mining System. In: **ECML PKDD'2012: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases**. [S.l.]: Springer Verlag, 2012. p. 820–823. Citation on page 88.

____. Mining representative movement patterns through compression. In: **Advances in Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2013. (Lecture Notes in Computer Science, v. 7818), p. 314–326. Citation on page 88.

HAN, B.; COOK, P.; BALDWIN, T. Text-based twitter user geolocation prediction. **J. Artif. Int. Res.**, AI Access Foundation, v. 49, n. 1, p. 451–500, 2014. ISSN 1076-9757. Citation on page 89.

HAN, X.; WANG, L.; CRESPI, N.; PARK, S.; CUEVAS, A. Alike people, alike interests? inferring interest similarity in online social networks. **Decis. Support Syst.**, Elsevier Science Publishers B. V., v. 69, n. C, p. 92–106, Jan. 2015. ISSN 0167-9236. Citation on page 69.

HAN, X.; WANG, L.; HAN, S. N.; CHEN, C.; CRESPI, N.; FARAHBAKHSH, R. Link prediction for new users in social networks. In: **2015 IEEE International Conference on Communications (ICC)**. [S.l.: s.n.], 2015. p. 1250–1255. Citation on page 69.

HAND, D. J.; SMYTH, P.; MANNILA, H. **Principles of Data Mining**. Cambridge, MA, USA: MIT Press, 2001. Citations on pages 25 and 60.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982. Citation on page 66.

HARENBERG, S.; BELLO, G.; GJELTEMA, L.; RANSHOUS, S.; HARLALKA, J.; SEAY, R.; PADMANABHAN, K.; SAMATOVA, N. Community detection in large-scale networks: a survey and empirical evaluation. **WIREs Comput Stat**, v. 6, p. 426–439, 2014. Citation on page 110.

HASAN, M. A.; CHAOJI, V.; SALEM, S.; ZAKI, M. Link prediction using supervised learning. In: **In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security**. [S.l.: s.n.], 2006. Citations on pages 26, 27, 60, 67, 82, and 86.

HASHEM, T.; HASHEM, T.; ALI, M. E.; KULIK, L. Group trip planning queries in spatial databases. In: **Proceedings of the 13th International Symposium on Advances in Spatial and Temporal Databases - Volume 8098**. [S.l.]: Springer-Verlag, 2013. (SSTD 2013), p. 259–276. Citation on page 88.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference and prediction**. 2. ed. [S.l.]: Springer, 2009. Citation on page 129.

HERLOCKER, J. L.; KONSTAN, J. A.; TERVEEN, L. G.; RIEDL, J. T. Evaluating collaborative filtering recommender systems. **ACM Trans. Inf. Syst.**, ACM, v. 22, n. 1, p. 5–53, 2004. Citation on page 66.

HIDALGO, C. A.; BLUMM, N.; BARABÁSI, A.-L.; CHRISTAKIS, N. A. A dynamic network approach for the study of human phenotypes. **PLOS Computational Biology**, Public Library of Science, v. 5, n. 4, p. 1–11, 04 2009. Citation on page 56.

HOLME, P. Modern temporal network theory: a colloquium. **The European Physical Journal B**, v. 88, n. 9, p. 234, 2015. Citation on page 59.

HOLME, P.; SARAMÄKI, J. Temporal networks. **Physics Reports**, v. 519, n. 3, p. 97–125, 2012. Citation on page 59.

____. **Temporal Networks**. [S.l.]: Springer-Verlag Berlin Heidelberg, 2013. (Understanding Complex Systems). Citation on page 59.

HONG, W.; YANSHEN, S.; XIAOMEI, Y. Personalized recommendation based on link prediction in dynamic super-networks. In: **Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT)**. [S.l.: s.n.], 2014. p. 1–7. Citation on page 86.

HOSEINI, E.; HASHEMI, S.; HAMZEH, A. Link prediction in social network using co-clustering based approach. In: **Proceedings of the 2012 26th International Conference on Advanced Information Networking and Applications Workshops**. [S.l.]: IEEE Computer Society, 2012. (WAINA '12), p. 795–800. ISBN 978-0-7695-4652-0. Citations on pages 28, 78, and 111.

HSIEH, H.-P.; LI, C.-T.; LIN, S.-D. Measuring and recommending time-sensitive routes from location-based data. In: **Proceedings of the 24th International Conference on Artificial Intelligence**. [S.l.]: AAAI Press, 2015. (IJCAI'15), p. 4193–4196. ISBN 978-1-57735-738-4. Citations on pages 97 and 103.

HU, G.; AGARWAL, P. Human disease-drug network based on genomic expression profiles. **PLOS ONE**, Public Library of Science, v. 4, n. 8, p. 1–11, 08 2009. Citation on page 56.

HUANG, Z. Link prediction based on graph topology: The predictive value of the generalized clustering coefficient. In: **The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), Workshop on Link Analysis: Dynamics and Static of Large Networks (LinkKDD' 06)**. [S.l.: s.n.], 2006. Citation on page 81.

HUANG, Z.; LIN, D. K. J. The time-series link prediction problem with applications in communication surveillance. **INFORMS Journal on Computing**, v. 21, n. 2, p. 286–303, 2009. Citation on page 67.

HUANG, Z.; ZENG, D. D. A link prediction approach to anomalous email detection. In: **2006 IEEE International Conference on Systems, Man and Cybernetics**. [S.l.: s.n.], 2006. v. 2, p. 1131–1136. Citation on page 86.

HUBERMAN, B. A. **The Laws of the Web: Patterns in the Ecology of Information**. Cambridge, MA, USA: MIT Press, 2001. ISBN 0262083035. Citation on page 57.

HUETTEL, S.; MACK, P.; MCCARTHY, G. Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. **Nature Neuroscience**, v. 5, n. 5, p. 485–90, 2002. Citation on page 25.

IAMNITCHI, A.; RIPEANU, M.; FOSTER, I. T. Locating data in (small-world?) peer-to-peer scientific collaborations. In: **Revised Papers from the First International Workshop on Peer-to-Peer Systems**. London, UK, UK: Springer-Verlag, 2002. (IPTPS '01), p. 232–241. ISBN 3-540-44179-4. Citation on page 57.

INOKUCHI, A.; WASHIO, T.; MOTODA, H. An apriori-based algorithm for mining frequent substructures from graph data. In: **Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery**. [S.l.]: Springer-Verlag, 2000. (PKDD '00), p. 13–23. Citation on page 62.

ITAKURA, K. Y.; CLARKE, C. L. A.; GEVA, S.; TROTMAN, A.; HUANG, W. C. Topical and structural linkage in wikipedia. In: **Proceedings of ECIR'11**. [S.l.: s.n.], 2011. p. 460–465. ISBN 978-3-642-20160-8. Citations on pages 27, 60, and 86.

JACCARD, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. **Bulletin del la Société Vaudoise des Sciences Naturelles**, v. 37, p. 547–579, 1901. Citation on page 71.

JAMAKOVIĆ, A. **Characterization of Complex Networks: Application to Robustness Analysis**. Phd Thesis (PhD Thesis) — Technische Universiteit Delft, The Netherlands, 2008. Citations on pages 26 and 47.

JEH, G.; WIDOM, J. Simrank: A measure of structural-context similarity. In: **Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2002. (KDD '02), p. 538–543. ISBN 1-58113-567-X. Citation on page 74.

JENSEN, D. Statistical challenges to inductive inference in linked data. In: **In Proceedings of the 17th International Workshop on Artificial Intelligence and Statistics**. [S.l.: s.n.], 1999. Citations on pages 26 and 61.

JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABáSI, A. L. The large-scale organization of metabolic networks. **Nature**, Nature Publishing Group, v. 407, n. 6804, p. 651–654, 2000. Citation on page 56.

JI, S.; ZHENG, Y.; LI, T. Urban sensing based on human mobility. In: **Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing**. [S.l.]: ACM, 2016. (UbiComp '16), p. 1040–1051. Citations on pages 88 and 89.

JIN, N.; YOUNG, C.; WANG, W. Gaia: Graph classification using evolutionary computation. In: **Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: ACM, 2010. (SIGMOD '10), p. 879–890. ISBN 978-1-4503-0032-2. Citation on page 62.

JURDAK, R.; ZHAO, K.; LIU, J.; ABOUJAOUDE, M.; CAMERON, M.; NEWTH, D. Understanding human mobility from twitter. **PLOS ONE**, Public Library of Science, v. 10, n. 7, p. 1–16, 07 2015. Citation on page 89.

KALAPALA, V.; SANWALANI, V.; CLAUSET, A.; MOORE, C. Scale invariance in road networks. **Phys. Rev. E**, v. 73, p. 026130, 2006. Citation on page 57.

KATZ, L. A new status index derived from sociometric analysis. **Psychometrika**, v. 18, n. 1, p. 39–43, 1953. Citation on page 74.

KAUTZ, H.; SELMAN, B.; SHAH, M. Referral web: Combining social networks and collaborative filtering. **Commun. ACM**, ACM, New York, NY, USA, v. 40, n. 3, p. 63–65, Mar. 1997. ISSN 0001-0782. Citation on page 58.

KEMAL, M. K.; TSUYOSHI, M. *et al.* Predicting within-and inter-community links in networks. **SIG-KBS**, v. 4, n. 1, p. 8–12, 2014. Citations on pages 28, 78, 108, and 111.

KIVELA, M.; ARENAS, A.; BARTHELEMY, M.; GLEESON, J. P.; MORENO, Y.; PORTER, M. A. Multilayer networks. **Journal of Complex Networks**, 2014. Citations on pages 38 and 60.

KLEINBERG, J. M.; KUMAR, R.; RAGHAVAN, P.; RAJAGOPALAN, S.; TOMKINS, A. S. The web as a graph: Measurements, models, and methods. In: **Proceedings of the 5th Annual International Conference on Computing and Combinatorics**. Berlin, Heidelberg: Springer-Verlag, 1999. (COCOON'99), p. 1–17. ISBN 3-540-66200-6. Citation on page 57.

KLEINBERG, J. M.; SURI, S.; TARDOS, E.; WEXLER, T. Strategic network formation with structural holes. In: **Proceedings of the 9th ACM Conference on Electronic Commerce**. [S.l.]: ACM, 2008. (EC '08), p. 284–293. Citation on page 77.

KOLLI, N.; NARAYANASWAMY, B. Analysis of e-mail communication using a social network framework for crisis detection in an organization. **Procedia - Social and Behavioral Sciences**, v. 100, p. 57 – 67, 2013. Citation on page 58.

KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix factorization techniques for recommender systems. **Computer**, IEEE Computer Society Press, v. 42, n. 8, p. 30–37, Aug. 2009. Citation on page 84.

KOSCHÜTZKI, D.; LEHMANN, K. A.; PEETERS, L.; RICHTER, S.; TENFELDE-PODEHL, D.; ZLOTOWSKI, O. Centrality indices. In: ____. **Network Analysis: Methodological Foundations**. [S.l.]: Springer Berlin Heidelberg, 2005. p. 16–61. Citation on page 51.

KOSSINETS, G.; WATTS, D. J. Empirical analysis of an evolving social network. **Science**, American Association for the Advancement of Science, v. 311, n. 5757, p. 88–90, 2006. Citations on pages 27, 77, 86, and 107.

KOTERA, M.; YAMANISHI, Y.; MORIYA, Y.; KANEHISA, M.; GOTO, S. Genies: gene network inference engine based on supervised analysis. **Nucleic Acids Research**, v. 40, n. Web-Server-Issue, p. 162–167, 2012. Citations on pages 26, 56, and 60.

KUANG, R.; LIU, Q.; YU, H. Community-based link prediction in social networks. In: ____. **Advances in Swarm Intelligence: 7th International Conference, ICSI 2016, Bali, Indonesia, June 25-30, 2016, Proceedings, Part II**. Cham: Springer International Publishing, 2016. p. 341–348. Citations on pages 28, 78, and 111.

KUMAR, R.; NOVAK, J.; TOMKINS, A. Structure and evolution of online social networks. In: **Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2006. (KDD '06), p. 611–617. ISBN 1-59593-339-5. Citations on pages 27 and 123.

KUNEGIS, J.; LOMMATZSCH, A. Learning spectral graph transformations for link prediction. In: **Proceedings of the 26th Annual International Conference on Machine Learning**. [S.l.]: ACM, 2009. (ICML '09), p. 561–568. ISBN 978-1-60558-516-1. Citation on page 84.

KWAK, H.; LEE, C.; PARK, H.; MOON, S. What is twitter, a social network or a news media? In: **Proceedings of WWW '10**. [S.l.: s.n.], 2010. p. 591–600. ISBN 978-1-60558-799-8. Citation on page 116.

KYLASA, S. B.; KOLLIAS, G.; GRAMA, A. Social ties and checkin sites: connections and latent structures in location-based social networks. **Social Network Analysis and Mining**, v. 6, n. 1, p. 95, 2016. ISSN 1869-5469. Citations on pages 28, 97, 98, and 154.

LAK, P.; CAGLAYAN, B.; BENER, A. B. The impact of basic matrix factorization refinements on recommendation accuracy. In: **Proceedings of the 2014 IEEE/ACM International Symposium on Big Data Computing**. [S.l.]: IEEE Computer Society, 2014. (BDC '14), p. 105–112. ISBN 978-1-4799-1897-3. Citation on page 84.

LEE, K.-W. R.; LIM, E.-P. Friendship maintenance and prediction in multiple social networks. In: **Proceedings of the 27th ACM Conference on Hypertext and Social Media**. New York, NY, USA: ACM, 2016. (HT '16), p. 83–92. ISBN 978-1-4503-4247-6. Citation on page 58.

LEE, L. Measures of distributional similarity. In: **Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics**. [S.l.]: Association for Computational Linguistics, 1999. (ACL '99), p. 25–32. ISBN 1-55860-609-3. Citation on page 85.

LEE, M.-J.; CHUNG, C.-W. A user similarity calculation based on the location for social network services. In: **Proceedings of the 16th International Conference on Database Systems for Advanced Applications - Volume Part I**. [S.l.]: Springer-Verlag, 2011. (DASFAA'11), p. 38–52. Citation on page 97.

LEICHT, E. A.; HOLME, P.; NEWMAN, M. E. J. Vertex similarity in networks. **Phys. Rev. E**, American Physical Society, v. 73, n. 2, p. 026120, Feb 2006. Citations on pages 72 and 74.

LESKOVEC, J.; ADAMIC, L. A.; HUBERMAN, B. A. The dynamics of viral marketing. **ACM Trans. Web**, ACM, New York, NY, USA, v. 1, n. 1, May 2007. ISSN 1559-1131. Citation on page 58.

LESKOVEC, J.; CHAKRABARTI, D.; KLEINBERG, J.; FALOUTSOS, C.; GHAHRAMANI, Z. Kronecker graphs: An approach to modeling networks. **J. Mach. Learn. Res.**, JMLR.org, v. 11, p. 985–1042, Mar. 2010. ISSN 1532-4435. Citation on page 62.

LESKOVEC, J.; HUTTENLOCHER, D.; KLEINBERG, J. Predicting positive and negative links in online social networks. In: **Proceedings of the 19th International Conference on World Wide Web**. [S.l.]: ACM, 2010. (WWW '10), p. 641–650. ISBN 978-1-60558-799-8. Citation on page 81.

LEUNG, I.; HUI, P.; LIO, P.; CROWCROFT, J. Towards real-time community detection in large networks. **Physical Review E**, v. 79, n. 6, p. 066107, 2009. Citation on page 110.

LI, C.-T.; HSIEH, H.-P. Geo-social media analytics. In: **Proceedings of the 24th International Conference on World Wide Web**. [S.l.]: ACM, 2015. (WWW '15 Companion), p. 1533–1534. ISBN 978-1-4503-3473-0. Citation on page 88.

LI, D.; ZHANG, Y.; XU, Z.; CHU, D.; LI, S. Exploiting information diffusion feature for link prediction in sina weibo. **Scientific Reports**, v. 6, p. srep20058, 2016. Citations on pages 27, 60, and 86.

LI, G.; SEMERCI, M.; YENER, B.; ZAKI, M. J. Effective graph classification based on topological and label attributes. **Stat. Anal. Data Min.**, John Wiley & Sons, Inc., v. 5, n. 4, p. 265–283, Aug. 2012. ISSN 1932-1864. Citation on page 62.

LI, J.; ZHANG, L.; MENG, F.; LI, F. Recommendation algorithm based on link prediction and domain knowledge in retail transactions. **Procedia Computer Science**, v. 31, p. 875 – 881, 2014. Citations on pages 27, 60, and 86.

LI, Q.; ZHENG, Y.; XIE, X.; CHEN, Y.; LIU, W.; MA, W.-Y. Mining user similarity based on location history. In: **Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems**. [S.l.]: ACM, 2008. (GIS '08), p. 34:1–34:10. ISBN 978-1-60558-323-5. Citations on pages 88, 97, and 103.

LI, X.; CHEN, H. Recommendation as link prediction: A graph kernel-based machine learning approach. In: **Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries**. New York, NY, USA: ACM, 2009. (JCDL '09), p. 213–216. ISBN 978-1-60558-322-8. Citations on pages 27 and 86.

LI, X.; JIN, Y. Y.; CHEN, G. Complexity and synchronization of the world trade web. **Physica A: Statistical Mechanics and its Applications**, v. 328, n. 1–2, p. 287 – 296, 2003. Citation on page 58.

LI, Y.; NIU, K.; TIAN, B. Link prediction in sina microblog using comprehensive features and improved SVM algorithm. In: **2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems**. [S.l.: s.n.], 2014. p. 18–22. Citation on page 83.

LIANG, H.; WANG, K.; ZHU, F. Mining social ties beyond homophily. **2016 IEEE 32nd International Conference on Data Engineering, ICDE 2016**, p. 421–432, 2016. Citation on page 58.

LIAO, Y.; LAM, W.; JAMEEL, S.; SCHOCKAERT, S.; XIE, X. Who wants to join me?: Companion recommendation in location based social networks. In: **Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval**. [S.l.]: ACM, 2016. (ICTIR '16), p. 271–280. ISBN 978-1-4503-4497-5. Citations on pages 27, 88, and 97.

LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. **Journal of the American Society for Information Science and Technology**, v. 58, n. 7, p. 1019–1031, May 2007. Citations on pages 26, 60, 62, 64, 65, 69, 70, 72, 74, 85, and 86.

LICHTENWALTER, R. N.; LUSSIER, J. T.; CHAWLA, N. V. New perspectives and methods in link prediction. In: **Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2010. (KDD '10), p. 243–252. ISBN 978-1-4503-0055-1. Citations on pages 64, 65, 67, 70, 75, and 82.

LICHTNWALTER, R.; CHAWLA, N. V. Link prediction: Fair and effective evaluation. In: **2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**. [S.l.: s.n.], 2012. p. 376–383. Citations on pages 67 and 162.

LIN, D. An information-theoretic definition of similarity. In: **Proceedings of the Fifteenth International Conference on Machine Learning**. [S.l.]: Morgan Kaufmann Publishers Inc., 1998. (ICML '98), p. 296–304. ISBN 1-55860-556-8. Citation on page 68.

LIN, Z.; ZHENG, X.; XIN, N.; CHEN, D. Ck-lpa: Efficient community detection algorithm based on label propagation with community kernel. **Physica A: Statistical Mechanics and its Applications**, v. 416, p. 386 – 399, 2014. Citations on pages 110 and 111.

LIU, G.; LI, J.; WONG, L. Assessing and predicting protein interactions using both local and global network topological metrics. **Genome Informatics**, n. 21, p. 138–149, 2008. Citation on page 73.

LIU, H.; HU, Z.; HADDADI, H.; TIAN, H. Hidden link prediction based on node centrality and weak ties. **EPL (Europhysics Letters)**, v. 101, n. 1, p. 18004, 2013. Citation on page 76.

LIU, J.; LI, Y.; LING, G.; LI, R.; ZHENG, Z. Community detection in location-based social networks: An entropy-based approach. In: **2016 IEEE International Conference on Computer and Information Technology (CIT)**. [S.l.: s.n.], 2016. p. 452–459. Citation on page 97.

LIU, J.; ZHANG, F.; SONG, X.; SONG, Y.-I.; LIN, C.-Y.; HON, H.-W. What's in a name?: An unsupervised approach to link users across communities. In: **Proceedings of the Sixth ACM International Conference on Web Search and Data Mining**. New York, NY, USA: ACM, 2013. (WSDM '13), p. 495–504. ISBN 978-1-4503-1869-3. Citation on page 62.

LIU, Q.; TANG, S.; ZHANG, X.; ZHAO, X.; ZHAO, B. Y.; ZHENG, H. Network growth and link prediction through an empirical lens. In: **Proceeding of the ACM Internet Measurement Conference**. [S.l.: s.n.], 2016. (IMC '16). To be published. Citation on page 67.

LIU, S.; JI, X.; LIU, C.; BAI, Y. Extended resource allocation index for link prediction of complex network. **Physica A: Statistical Mechanics and its Applications**, 2017. To appear. Citation on page 73.

LIU, W.; JIANG, X.; PELLEGRINI, M.; WANG, X. Discovering communities in complex networks by edge label propagation. **Scientific Reports**, v. 6, p. 22470, 2016. Citations on pages 110 and 111.

LIU, W.; LÜ, L. Link prediction based on local random walk. **EPL (Europhysics Letters)**, v. 89, n. 5, p. 58007, 2010. Citations on pages 74 and 75.

LIU, W.; ZHENG, Y.; CHAWLA, S.; YUAN, J.; XING, X. Discovering spatio-temporal causal interactions in traffic data streams. In: **Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2011. (KDD '11), p. 1010–1018. ISBN 978-1-4503-0813-7. Citation on page 88.

LIU, X.; LIU, Y.; ABERER, K.; MIAO, C. Personalized point-of-interest recommendation by mining users' preference transition. In: **Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management**. [S.l.]: ACM, 2013. (CIKM '13), p. 733–738. ISBN 978-1-4503-2263-8. Citation on page 103.

LIU, Y.; LI, Z. A novel algorithm of low sampling rate gps trajectories on map-matching. **EURASIP Journal on Wireless Communications and Networking**, v. 2017, n. 1, p. 30, 2017. Citation on page 91.

LIU, Z.; HE, J.-L.; KAPOOR, K.; SRIVASTAVA, J. Correlations between community structure and link formation in complex networks. **PLOS ONE**, Public Library of Science, v. 8, n. 9, p. 1–10, 09 2013. Citations on pages 28, 77, 81, 108, and 111.

LIU, Z.; HUANG, Y. Where are you tweeting?: A context and user movement based approach. In: **Proceedings of the 25th ACM International on Conference on Information and Knowledge Management**. [S.l.]: ACM, 2016. (CIKM '16), p. 1949–1952. Citation on page 89.

LIU, Z.; ZHANG, Q.-M.; LÜ, L.; ZHOU, T. Link prediction in complex networks: A local naïve bayes model. **EPL (Europhysics Letters)**, v. 96, n. 4, p. 48007, 2011. Citations on pages 72, 114, 134, 136, and 140.

LOGLISCI, C.; IENCO, D.; ROCHE, M.; TEISSEIRE, M.; MALERBA, D. Toward geographic information harvesting: Extraction of spatial relational facts from web documents. In: **Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on**. [S.l.: s.n.], 2012. p. 789–796. Citation on page 89.

____. An unsupervised framework for topological relations extraction from geographic documents. In: . [S.l.: s.n.], 2012. (Lecture Notes in Computer Science, v. 7447), p. 48–55. Citation on page 89.

LONG, X. **Location-based Social Networks: Latent Topics Mining and Hybrid Trust-based Recommendation**. Phd Thesis (PhD Thesis) — University of Pittsburgh, USA, 2015. Citations on pages 91 and 96.

LOPES, A.; BERTINI, J. R.; MOTTA, R.; ZHAO, L. Classification based on the optimal k-associated network. In: **Complex Sciences: First International Conference, Complex 2009, Revised Papers, Part 1**. [S.l.]: Springer Berlin Heidelberg, 2009. p. 1167–1177. Citation on page 113.

LORRAIN, F.; WHITE, H. Structural equivalence of individuals in social networks. **Journal of Mathematical Sociology**, v. 1, p. 49–80, 1971. Citation on page 70.

LÜ, L.; JIN, C.-H.; ZHOU, T. Similarity index based on local paths for link prediction of complex networks. **Phys. Rev. E**, American Physical Society, v. 80, p. 046122, 2009. Citation on page 75.

LÜ, L.; ZHOU, T. Link prediction in weighted networks: The role of weak ties. **EPL (Europhysics Letters)**, v. 89, n. 1, p. 18001, 2010. Citations on pages 28 and 76.

____. Link prediction in complex networks: A survey. **Physica A: Statistical Mechanics and its Applications**, v. 390, n. 6, p. 1150–1170, 2011. Citations on pages 26, 60, 62, 64, 66, 67, 68, 69, 70, 72, 74, 75, 78, 86, and 97.

LU, Q.; GETOOR, L. Link-based classification. In: **Proceedings of the Twentieth International Conference on International Conference on Machine Learning**. [S.l.]: AAAI Press, 2003. (ICML'03), p. 496–503. ISBN 1-57735-189-4. Citation on page 62.

LUO, H.; GUO, B.; ZHIWENYU; WANG, Z.; FENG, Y. Friendship Prediction Based on the Fusion of Topology and Geographical Features in LBSN. In: **10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC-EUC)**. [S.l.]: IEEE, 2013. p. 2224–2230. Citations on pages 97, 100, 153, 162, and 163.

LV, M.; CHEN, L.; CHEN, G. Mining user similarity based on routine activities. **Inf. Sci.**, Elsevier Science Inc., v. 236, p. 17–32, Jul. 2013. ISSN 0020-0255. Citation on page 97.

MA, H. On measuring social friend interest similarities in recommender systems. In: **Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval**. [S.l.]: ACM, 2014. (SIGIR '14), p. 465–474. ISBN 978-1-4503-2257-7. Citation on page 69.

MA, H.; CAO, H.; YANG, Q.; CHEN, E.; TIAN, J. A habit mining approach for discovering similar mobile users. In: **Proceedings of the 21st International Conference on World Wide Web**. [S.l.]: ACM, 2012. (WWW '12), p. 231–240. Citations on pages 88 and 89.

MA, Y.; CHENG, G.; LIU, Z.; LIANG, X. Link prediction based on clustering information in scientific coauthorship networks. In: **2016 IEEE First International Conference on Data Science in Cyberspace (DSC)**. [S.l.: s.n.], 2016. p. 668–672. Citations on pages 28, 78, and 111.

MACSKASSY, S. A.; PROVOST, F. Classification in networked data: A toolkit and a univariate case study. **J. Mach. Learn. Res.**, JMLR.org, v. 8, p. 935–983, May 2007. ISSN 1532-4435. Citation on page 62.

MALLEK, S.; BOUKHRIS, I.; ELOUEDI, Z.; LEFEVRE, E. Evidential link prediction based on group information. In: ____. **Mining Intelligence and Knowledge Exploration: Third International Conference, MIKE 2015, Hyderabad, India, December 9-11, 2015, Proceedings**. Cham: Springer International Publishing, 2015. p. 482–492. Citations on pages 28, 78, 108, and 111.

MARIANI, M.; MEDO, M.; ZHANG, Y.-C. Ranking nodes in growing networks: When pagerank fails. **Scientific Reports**, n. 5, p. 16181, 2015. Citation on page 62.

MARITAN, A.; RINALDO, A.; RIGON, R.; GIACOMETTI, A.; RODRÍGUEZ-ITURBE, I. Scaling laws for river networks. **Phys. Rev. E**, p. 1510, 1996. Citation on page 57.

MARR, B. **Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance**. UK: John Wiley & Sons Ltd, 2015. Citation on page 26.

MARTÍNEZ, V.; BERZAL, F.; CUBERO, J.-C. A survey of link prediction in complex networks. **ACM Comput. Surv.**, ACM, v. 49, n. 4, p. 69:1–69:33, 2016. Citations on pages 26, 68, 69, 70, 71, 72, 74, 75, 78, 79, and 85.

MAYER-SCHÖNBERGER, V.; CUKIER, K. **Big Data: A Revolution That Will Transform How We Live, Work and Think**. UK: John Murray Publishers, 2013. Citation on page 26.

MCGEE, J.; CAVERLEE, J.; CHENG, Z. Location prediction in social media based on tie strength. In: **Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management**. [S.l.]: ACM, 2013. (CIKM '13), p. 459–468. ISBN 978-1-4503-2263-8. Citations on pages 103 and 163.

MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. **Annual Review of Sociology**, v. 27, n. 1, p. 415–444, 2001. Citations on pages 28 and 76.

MEDO, M. Network-based information filtering algorithms: Ranking and recommendation. In: ____. **Dynamics On and Of Complex Networks, Volume 2: Applications to Time-Varying Dynamical Systems**. New York, NY: Springer New York, 2013. p. 315–334. Citation on page 62.

MENCHE, J.; SHARMA, A.; KITSAK, M.; GHIASSIAN, S. D.; VIDAL, M.; LOSCALZO, J.; BARABÁSI, A.-L. Uncovering disease-disease relationships through the incomplete interactome. **Science**, American Association for the Advancement of Science, v. 347, n. 6224, 2015. ISSN 0036-8075. Citations on pages 26, 56, and 60.

MENGSHOEL, O.; DESAIL, R.; CHEN, A.; TRAN, B. Will we connect again? machine learning for link prediction in mobile social networks. In: **ACM MLG 2013**. [S.l.: s.n.], 2013. Citations on pages 28, 94, 97, 98, 154, and 163.

MENON, A. K.; ELKAN, C. Link prediction via matrix factorization. In: **Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II**. [S.l.]: Springer-Verlag, 2011. (ECML PKDD'11), p. 437–452. ISBN 978-3-642-23782-9. Citation on page 84.

MESSIAS, J.; MAGNO, G.; BENEVENUTO, F.; VELOSO, A.; ALMEIDA, V. Brazil around the world: Characterizing and detecting brazilian emigrants using google+. In: **Proceedings of the 21st Brazilian Symposium on Multimedia and the Web**. [S.l.]: ACM, 2015. (WebMedia '15), p. 85–91. Citation on page 89.

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. **The Journal of Chemical Physics**, v. 21, n. 6, p. 1087–1092, 1953. Citation on page 81.

MILGRAM, S. The small world problem. **Psychol. Today**, v. 2, p. 60–67, 1967. Citations on pages 26 and 48.

MIRZAL, A. On the relationship between trading network and www network: a preferential attachment perspective. **Int. J. Bus. Intell. Data Min.**, Inderscience Publishers, v. 5, n. 3, p. 247–268, 2010. ISSN 1743-8195. Citation on page 57.

MISLOVE, A.; MARCON, M.; GUMMADI, K. P.; DRUSCHEL, P.; BHATTACHARJEE, B. Measurement and analysis of online social networks. In: **ACM SIGCOMM IMC '07**. [S.l.]: ACM, 2007. p. 29–42. ISBN 978-1-59593-908-1. Citations on pages 88, 124, 127, and 137.

MISLOVE, A. E. **Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems**. Phd Thesis (PhD Thesis) — Rice University, USA, 2009. Citations on pages 27, 123, 124, and 127.

MOHAMED, S.; ABDELMOTY, A. Computing similarity between users on location-based social networks. **International Journal on Advances in Intelligent Systems**, v. 9, n. 3 & 4, p. 542–553, 2016. Citation on page 97.

MORADABADI, B.; MEYBODI, M. R. A novel time series link prediction method: Learning automata approach. **Physica A: Statistical Mechanics and its Applications**, v. 482, p. 422–432, 2017. Citation on page 67.

MOTODA, H. Pattern discovery from graph-structured data: A data mining perspective. In: **Proceedings of the 20th International Conference on Industrial, Engineering, and Other Applications of Applied Intelligent Systems**. [S.l.]: Springer-Verlag, 2007. (IEA/AIE'07), p. 12–22. ISBN 978-3-540-73322-5. Citation on page 62.

MURZIN, A. G.; BRENNER, S. E.; HUBBARD, T.; CHOTHIA, C. Scop: A structural classification of proteins database for the investigation of sequences and structures. **Journal of Molecular Biology**, v. 247, n. 4, p. 536–540, 1995. Citation on page 56.

NANDURI, A.; RANGWALA, H. Predicting new friendships in social networks. In: **2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)**. [S.l.: s.n.], 2015. p. 521–526. Citation on page 83.

NAUDÉ, K. A.; GREYLING, J. H.; VOGTS, D. When similarity measures lie. In: **Proceedings of the 8th International Conference on Similarity Search and Applications - Volume 9371**. [S.l.]: Springer-Verlag New York, Inc., 2015. (SISAP 2015), p. 113–124. ISBN 978-3-319-25086-1. Citation on page 64.

NAVARRO, G. A guided tour to approximate string matching. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 33, n. 1, p. 31–88, 2001. ISSN 0360-0300. Citation on page 69.

NEVILLE, J.; JENSEN, D. Iterative classification in relational data. In: **Proceedings of the Workshop on Learning Statistical Models from Relational Data, Seventeenth National Conference on Artificial Intelligence**. [S.l.]: AAAI Press, Menlo Park, CA, 2000. p. 42–49. Citation on page 58.

NEWMAN, M. Clustering and preferential attachment in growing networks. **Physical Review E**, v. 64, n. 2, p. 025102, 2001. Citation on page 70.

____. Assortative mixing in networks. **Phys. Rev. Lett.**, v. 89, n. 20, p. 208701, 2002. Citation on page 51.

____. Mixing patterns in networks. **Phys. Rev. E**, American Physical Society, v. 67, n. 2, p. 026126, 2003. Citation on page 51.

____. Analysis of weighted networks. **Physical Review E**, American Physical Society, v. 70, n. 5, p. 056131, 2004. Citation on page 51.

____. Modularity and community structure in networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 103, n. 23, p. 8577–8582, 2006. Citation on page 51.

____. **Networks: an introduction**. [S.l.]: Oxford University Press, 2010. Citations on pages 25, 26, 32, 33, 44, 46, 47, 51, 55, and 108.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM Review**, n. 45, p. 167–256, 2003. Citations on pages 26, 32, and 47.

____. Fast algorithm for detecting community structure in networks. **Phys. Rev. E**, American Physical Society, v. 69, p. 066133, 2004. Citations on pages 62 and 110.

NIE, Y.; JIA, Y.; LI, S.; ZHU, X.; LI, A.; ZHOU, B. Identifying users across social networks based on dynamic core interests. **Neurocomputing**, v. 210, p. 107 – 115, 2016. SI: Behavior Analysis in SN. Citation on page 62.

NOULAS, A. **Human Urban Mobility in Location-based Social Networks: Analysis, Models and Applications**. Phd Thesis (PhD Thesis) — St. Edmund's College, Computer Laboratory, University of Cambridge, United Kingdom, 2013. Citations on pages 89, 91, and 96.

O'MADADHAIN, J.; HUTCHINS, J.; SMYTH, P. Prediction and ranking algorithms for event-based network data. **SIGKDD Explor. Newsl.**, ACM, v. 7, n. 2, p. 23–30, Dec. 2005. ISSN 1931-0145. Citation on page 86.

ONCE-CS. **Living Roadmap for Complex Systems Science**. 2006. 71 p. IST-FET Coordination Action. Open Network of Centres of Excellence in Complex Systems. Project FP6-IST 29814. Version 1.22. Citation on page 32.

OPSAHL, T.; PANZARASA, P. Clustering in weighted networks. **Social Networks**, v. 31, p. 155–163, 2009. Citation on page 51.

OSPINA, E.; MORENO, F.; URIBE, I. A. Using criteria reconstruction for low-sampling trajectories as a tool for analytics. **Procedia Computer Science**, v. 51, p. 366 – 373, 2015. Citation on page 91.

PAN, Y.; LI, D.-H.; LIU, J.-G.; LIANG, J.-Z. Detecting community structure in complex networks via node similarity. **Physica A: Statistical Mechanics and its Applications**, v. 389, n. 14, p. 2849 – 2857, 2010. Citations on pages 62, 71, and 110.

PAPADIMITRIOU, A.; SYMEONIDIS, P.; MANOLOPOULOS, Y. Fast and accurate link prediction in social networking systems. **J. Syst. Softw.**, Elsevier Science Inc., v. 85, n. 9, p. 2119–2132, Sep. 2012. ISSN 0164-1212. Citation on page 75.

PARK, M.-H.; HONG, J.-H.; CHO, S.-B. Location-based recommendation system using bayesian user's preference model in mobile devices. In: **Proceedings of the 4th International Conference on Ubiquitous Intelligence and Computing**. [S.l.]: Springer-Verlag, 2007. (UIC'07), p. 1130–1139. ISBN 3-540-73548-8, 978-3-540-73548-9. Citation on page 103.

PASTOR-SATORRAS, R.; VESPIGNANI, A. **Evolution and Structure of the Internet: A statistical physics approach**. [S.l.]: Cambridge University Press, 2004. Citation on page 56.

PEARSON, K. The Problem of the Random Walk. **Nature**, Nature Publishing Group, v. 72, n. 1865, p. 294, 1905. Citation on page 74.

PECH, R.; HAO, D.; PAN, L.; CHENG, H.; ZHOU, T. Link prediction via matrix completion. **EPL (Europhysics Letters)**, v. 117, n. 3, p. 38002, 2017. Citation on page 85.

PEREZ-CERVANTES, E.; MENA-CHALCO, J. P.; OLIVEIRA, M. C. F. D.; JR., R. M. C. Using link prediction to estimate the collaborative influence of researchers. In: **Proceedings of the 2013 IEEE 9th International Conference on e-Science**. Washington, DC, USA: IEEE Computer Society, 2013. (ESCIENCE '13), p. 293–300. ISBN 978-0-7695-5083-1. Citation on page 86.

PÉREZ-SOLÀ, C.; HERRERA-JOANCOMARTÍ, J. Improving relational classification using link prediction techniques. In: ____. **Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Proceedings, Part I**. [S.l.]: Springer Berlin Heidelberg, 2013. p. 590–605. Citation on page 86.

PEROZZI, B.; SCHUEPPERT, M.; SAALWEACHTER, J.; THAKUR, M. When recommendation goes wrong: Anomalous link discovery in recommendation networks. In: **Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2016. (KDD '16), p. 569–578. Citations on pages 27 and 86.

PHAM, H.; SHAHABI, C.; LIU, Y. Ebm: An entropy-based model to infer social strength from spatiotemporal data. In: **Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data**. [S.l.]: ACM, 2013. (SIGMOD '13), p. 265–276. ISBN 978-1-4503-2037-5. Citations on pages 28, 65, 97, and 154.

POLKINGHORNE, J. **Reductionism**. 2002. Interdisciplinary Encyclopedia of Religion and Science. Advanced School for Interdisciplinary Research, Pontifical University of the Holy Cross. Citation on page 32.

PONS, P.; LATAPY, M. Computing communities in large networks using random walks. **J. Graph Algorithms Appl.**, v. 10, n. 2, p. 191–218, 2006. Citation on page 110.

PRICE, D. J. de S. Networks of scientific papers. **Science**, American Association for the Advancement of Science, v. 149, n. 3683, p. 510–515, 1965. ISSN 0036-8075. Citation on page 57.

PRIEDHORSKY, R.; CULOTTA, A.; VALLE, S. Y. D. Inferring the origin locations of tweets with quantitative confidence. In: **Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing**. [S.l.]: ACM, 2014. (CSCW '14), p. 1523–1536. Citation on page 89.

QUERCIA, D.; LATHIA, N.; CALABRESE, F.; LORENZO, G. D.; CROWCROFT, J. Recommending social events from mobile phone location data. In: **Proceedings of the 2010 IEEE International Conference on Data Mining**. [S.l.]: IEEE Computer Society, 2010. (ICDM '10), p. 971–976. ISBN 978-0-7695-4256-0. Citation on page 88.

RAFAILIDIS, D.; CRESTANI, F. Collaborative ranking with social relationships for top-n recommendations. In: **Proceedings of the 39th International ACM Conference on Research and Development in Information Retrieval**. [S.l.]: ACM, 2016. (SIGIR '16), p. 785–788. Citations on pages 27 and 170.

RAGHAVAN, U. N.; ALBERT, R.; KUMARA, S. Near linear time algorithm to detect community structures in large-scale networks. **Phys. Rev. E**, American Physical Society, v. 76, p. 036106, 2007. Citations on pages 110 and 111.

RAHMAN, M.; HASAN, M. A. Link prediction in dynamic networks usingźgraphlet. In: **European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851**. [S.l.]: Springer-Verlag, 2016. (ECML PKDD 2016), p. 394–409. Citation on page 67.

RANVIER, J.-E.; CATASTA, M.; VASIRANI, M.; ABERER, K. Routinesense: A mobile sensing framework for the reconstruction of user routines. In: **Proceedings of the 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services on 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services**. [S.l.]: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2015. (MOBIQUITOUS'15), p. 150–159. Citations on pages 88 and 89.

RAVASZ, E.; BARABÁSI. Hierarchical organization in complex networks. **Phys. Rev. Lett**, n. 67, p. 026112, 2003. Citation on page 51.

RAVASZ, E.; SOMERA, A. L.; MONGRU, D. A.; OLTVAI, Z. N.; BARABáSI, A. L. Hierarchical Organization of Modularity in Metabolic Networks. **Science**, American Association for the Advancement of Science, v. 297, n. 5586, p. 1551–1555, 2002. ISSN 1095-9203. Citations on pages 72 and 80.

REDNER, S. Networks: Teasing out the missing links. **Nature**, v. 453, p. 47–48, 2008. Citation on page 86.

REHMAN, S. U.; ASGHAR, S.; ZHUANG, Y.; FONG, S. Performance evaluation of frequent subgraph discovery techniques. **Mathematical Problems in Engineering**, v. 2014, 2014. Citation on page 62.

RIJSBERGEN, C. J. V. **Information Retrieval**. 2nd. ed. Newton, MA, USA: Butterworth-Heinemann, 1979. ISBN 0408709294. Citation on page 65.

ROA-VALVERDE, A. J.; SICILIA, M.-A. A survey of approaches for ranking on the web of data. **Information Retrieval**, v. 17, n. 4, p. 295–325, 2014. Citation on page 62.

ROBUSTO, C. The cosine-haversine formula. **The American Mathematical Monthly**, JSTOR, v. 64, n. 1, p. 38–40, 1957. Citation on page 95.

ROICK, O.; HEUSER, S. Location based social networks – definition, current state of the art and research agenda. **Transactions in GIS**, v. 17, n. 5, p. 763–784, 2013. ISSN 1467-9671. Citations on pages 87, 88, and 93.

ROMERO, D. M.; KLEINBERG, J. M. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In: **Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM' 2010**. [S.l.: s.n.], 2010. Citation on page 77.

ROSSI, L.; MUSOLESI, M. It's the way you check-in: Identifying users in location-based social networks. In: **Proceedings of the Second ACM Conference on Online Social Networks**. [S.l.]: ACM, 2014. (COSN '14), p. 215–226. Citation on page 97.

ROSSI, L.; WILLIAMS, M. J.; STICH, C.; MUSOLESI, M. Privacy and the city: User identification and location semantics in location-based social networks. In: **Proceedings of the 9th AAAI International Conference on Weblogs and Social Media (ICWSM 2015)**. [S.l.: s.n.], 2015. Citation on page 91.

ROSSI, R. A.; ZHOU, R.; AHMED, N. K. Relational similarity machines. In: **Proceedings of the 12th International Workshop on Mining and Learning with Graphs (MLG)**. [S.l.: s.n.], 2016. p. 1–8. Citation on page 62.

ROSVALL, M.; BERGSTROM, C. T. Mapping change in large networks. **PLOS ONE**, Public Library of Science, v. 5, n. 1, p. 1–7, 01 2010. Citation on page 110.

SADILEK, A.; KAUTZ, H.; BIGHAM, J. P. Finding your friends and following them to where you are. In: **Proceedings of the Fifth ACM International Conference on Web Search and Data Mining**. [S.l.]: ACM, 2012. (WSDM '12), p. 723–732. Citation on page 103.

SAIKAEW, K. R.; JIRANUWATTANAWONG, P.; TAEARAK, P. Place recommendation using location-based services and real-time social network data. **International Journal of Computer, Electrical, Automation, Control and Information Engineering**, World Academy of Science, Engineering and Technology, v. 9, n. 1, p. 300 – 305, 2015. ISSN PISSN:2010-376X, EISSN:2010-3778. Citation on page 103.

SALES-PARDO, M.; GUIMERÀ, R.; MOREIRA, A. A.; AMARAL, L. A. N. Extracting the hierarchical organization of complex systems. **Proceedings of the National Academy of Sciences**, v. 104, n. 39, p. 15224–15229, 2007. Citation on page 79.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. [S.l.]: McGraw-Hill, 1983. (McGraw-Hill computer science series). Citation on page 72.

SARKAR, P.; CHAKRABARTI, D.; JORDAN, M. I. Nonparametric link prediction in dynamic networks. In: **Proceedings of the 29th International Coference on International Conference on Machine Learning**. [S.l.]: Omnipress, 2012. (ICML'12), p. 1897–1904. Citation on page 67.

SARKAR, P.; CHAKRABARTI, D.; MOORE, A. W. Theoretical justification of popular link prediction heuristics. In: **Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three**. [S.l.]: AAAI Press, 2011. (IJCAI'11), p. 2722–2727. Citation on page 74.

SARWAR, B.; KARYPIS, G.; KONSTAN, J.; RIEDL, J. Application of dimensionality reduction in recommender systems: a case study. In: **Proceedings of ACM WebKDD Workshop**. [S.l.: s.n.], 2000. Citation on page 84.

SCELLATO, S.; MASCOLO, C.; MUSOLESI, M.; LATORA, V. Distance matters: Geo-social metrics for online social networks. In: **Proceedings of the 3rd Wonference on Online Social Networks**. [S.l.]: USENIX Association, 2010. (WOSN'10). Citations on pages 88, 95, and 101.

SCELLATO, S.; NOULAS, A.; MASCOLO, C. Exploiting place features in link prediction on location-based social networks. In: **ACM KDD**. [S.l.: s.n.], 2011. p. 1046–1054. ISBN 978-1-4503-0813-7. Citations on pages 64, 65, 94, 95, 96, 98, 100, 101, 154, and 164.

SCHACHERER, F.; CHOI, C.; GÖTZE, U.; KRULL, M.; PISTOR, S.; WINGENDER, E. The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. **Bioinformatics**, v. 17, n. 11, p. 1053–1057, 2001. Citation on page 56.

SHEN-ORR, S. S.; MILO, R.; MANGAN, S.; ALO, U. Network motifs in the transcriptional regulation network of Escherichia coli. **Nature Genetics**, v. 31, n. 1, p. 64–68, 2002. Citation on page 56.

SHERKAT, E.; RAHGOZAR, M.; ASADPOUR, M. Structural link prediction based on ant colony approach in social networks. **Physica A: Statistical Mechanics and its Applications**, v. 419, p. 80–94, 2015. Citation on page 83.

SHIN, D.; SI, S.; DHILLON, I. S. Multi-scale link prediction. In: **Proceedings of the 21st ACM International Conference on Information and Knowledge Management**. [S.l.]: ACM, 2012. (CIKM '12), p. 215–224. Citation on page 85.

SHOKRI, R.; THEODORAKOPOULOS, G.; DANEZIS, G.; HUBAUX, J.-P.; BOUDEC, J.-Y. L. Quantifying location privacy: The case of sporadic location exposure. In: **Proceedings of the 11th International Conference on Privacy Enhancing Technologies**. [S.l.]: Springer-Verlag, 2011. (PETS'11), p. 57–76. Citation on page 91.

SIGMAN, M.; CECCHI, G. A. Global organization of the wordnet lexicon. **Proceedings of the National Academy of Sciences**, v. 99, n. 3, p. 1742–1747, 2002. Citation on page 58.

SILVA, T. C.; ZHAO, L. **Machine Learning in Complex Networks**. [S.l.]: Springer International Publishing, 2016. Citations on pages 26, 32, 33, 44, 47, and 51.

SINGH-BLOM, U. M.; NATARAJAN, N.; TEWARI, A.; WOODS, J. O.; DHILLON, I. S.; MARCOTTE, E. M. Prediction and validation of gene-disease associations using methods inspired by social network analyses. **PLoS ONE**, v. 8, n. 5, p. e58977, 2013. Citations on pages 26, 56, and 60.

SINTOS, S.; TSAPARAS, P. Using strong triadic closure to characterize ties in social networks. In: **Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2014. (KDD '14), p. 1466–1475. ISBN 978-1-4503-2956-9. Citation on page 77.

SJAASTAD, L. A. The costs and returns of human migration. **The Journal of Political Economy**, v. 70, n. 5, p. 80–93, 1962. Citation on page 89.

SOARES, P. R. da S.; PRUDêNCIO, R. B. C. Time series based link prediction. In: **The 2012 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2012. p. 1–7. Citation on page 67.

SOCIEVOLE, A.; RANGO, F. D.; MARANO, S. Link prediction in human contact networks using online social ties. In: **2013 International Conference on Cloud and Green Computing**. [S.l.: s.n.], 2013. p. 305–312. Citations on pages 28 and 76.

SORENSEN, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation. **Danish commons. Biol. Skr. 5**, n. 4, 1948. Citation on page 72.

SOUNDARAJAN, S.; HOPCROFT, J. Using community information to improve the precision of link prediction methods. In: **Proceedings of the 21st international conference companion on World Wide Web**. [S.l.: s.n.], 2012. (Proceedings of WWW '12 Companion), p. 607–608. ISBN 978-1-4503-1230-1. Citations on pages 28, 77, 78, 108, and 111.

SRINIVAS, K.; REDDY, L. K. K.; GOVARDHAN, A. A theoretical approach to link mining for personalization. **International Journal of Computer Science Issues (IJCSI)**, v. 7, n. 9, p. 41–44, 2010. Citation on page 61.

SRINIVAS, V.; MITRA, P. Applications of link prediction. In: ____. **Link Prediction in Social Networks: Role of Power Law Distribution**. Cham: Springer International Publishing, 2016. p. 57–61. ISBN 978-3-319-28922-9. Citations on pages 26, 60, 64, and 86.

STATE, B.; WEBER, I.; ZAGHENI, E. Studying inter-national mobility through ip geolocation. In: **Proceedings of the Sixth ACM International Conference on Web Search and Data Mining**. [S.l.]: ACM, 2013. (WSDM '13), p. 265–274. Citation on page 89.

STEURER, M.; TRATTNER, C. Acquaintance or partner predicting partnership in online and location-based social networks. In: **IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining**. [S.l.]: IEEE, 2013. (ASONAM 2013), p. 372–379. Citations on pages 98 and 163.

STEURER, M.; TRATTNER, C.; HELIC, D. Predicting social interactions from different sources of location-based knowledge. In: **Proceedings of the Third International Conference on Social Eco-Informatics**. Lisbon, Portugal: IARIA, 2013. (SOTICS 2013), p. 8–13. ISBN 978-1-61208-312-4. Citations on pages 98 and 163.

STEYVERS, M.; TENENBAUM, J. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. **Cognitive science**, v. 29, n. 1, p. 41–79, 2005. Citation on page 58.

STROGATZ, S. Exploring complex networks. **Nature**, v. 410, p. 269–276, 2001. Citation on page 32.

SUN, Y.; BARBER, R.; GUPTA, M.; AGGARWAL, C. C.; HAN, J. Co-author relationship prediction in heterogeneous bibliographic networks. In: **Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining**. [S.l.]: IEEE Computer Society, 2011. (ASONAM '11), p. 121–128. ISBN 978-0-7695-4375-8. Citation on page 58.

SUN, Y.; HAN, J. Ranking methods for networks. In: ____. **Encyclopedia of Social Network Analysis and Mining**. [S.l.]: Springer New York, 2014. p. 1488–1497. Citation on page 62.

SURANA, A.; KUMARA, S.; GREAVES, M.; RAGHAVAN, U. N. Supply-chain networks: a complex adaptive systems perspective. **International Journal of Production Research**, v. 43, n. 20, p. 4235–4265, 2005. Citation on page 58.

SYMEONIDIS, P.; NTEMPOS, D.; MANOLOPOULOS, Y. Recommender systems for location-based social network. In: ____. [S.l.]: Springer, 2014. chap. Location-based social networks, p. 35–48. Citations on pages 87, 88, and 93.

SZKLARCZYK, D.; FRANCESCHINI, A.; KUHN, M.; SIMONOVIC, M.; ROTH, A.; MINGUEZ, P.; DOERKS, T.; STARK, M.; MULLER, J.; BORK, P.; JENSEN, L.; MERING, C. von. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. **Nucleic Acids Res.**, p. D561–D568, 2011. Citation on page 56.

TAHRAT, S.; ROCHE, M.; TESSEIRE, M. Extraction of Geospatial Information from Documents. In: **Geographic Information Retrieval Tutorial - Panel discussion. AGILE Worskshop**. [S.l.: s.n.], 2012. p. 2. Citation on page 89.

TAN, F.; XIA, Y.; ZHU, B. Link prediction in complex networks: A mutual information perspective. **PLOS ONE**, Public Library of Science, v. 9, n. 9, p. 1–8, 09 2014. Citation on page 73.

TANAKA, R. Scale-rich metabolic networks. **Physical Review Letters**, v. 94, n. 16, p. 168101, 2005. Citation on page 56.

TANG, L.-A.; ZHENG, Y.; XIE, X.; YUAN, J.; YU, X.; HAN, J. Retrieving k-nearest neighboring trajectories by a set of point locations. In: **Proceedings of the 12th International Conference on Advances in Spatial and Temporal Databases**. [S.l.]: Springer-Verlag, 2011. (SSTD'11), p. 223–241. Citation on page 88.

TANTA-NGAI, H.; MILIOS, E. E.; KESELJ, V. Self-organizing peer-to-peer networks for collaborative document tracking. In: **Proceedings of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management**. New York, NY, USA: ACM, 2009. (CNIKM '09), p. 59–66. ISBN 978-1-60558-807-0. Citation on page 57.

TASHIRO, H.; MORI, J.; FUJII, N.; MATSUSHIMA, K. Email network analysis for organizational management. In: **2010 IEEE International Conference on Management of Innovation Technology**. [S.l.: s.n.], 2010. p. 958–963. Citation on page 58.

TELESFORD, Q.; JOYCE, K.; HAYASAKA, S.; BURDETTE, J.; LAURIENTI, P. The ubiquity of small-world networks. **Brain Connectivity**, n. 1, p. 367–375, 2011. Citation on page 48.

THOMAS, S.; NAIR, J. J. A survey on extracting frequent subgraphs. In: **2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)**. [S.l.: s.n.], 2016. p. 2290–2295. Citation on page 62.

TONG, H.; FALOUTSOS, C.; PAN, J.-Y. Fast random walk with restart and its applications. In: **Proceedings of the Sixth International Conference on Data Mining**. [S.l.]: IEEE Computer Society, 2006. (ICDM '06), p. 613–622. ISBN 0-7695-2701-9. Citation on page 74.

TRAUD, A. L.; MUCHA, P. J.; PORTER, M. A. Social structure of facebook networks. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 16, p. 4165 – 4180, 2012. Citations on pages 58 and 124.

TRUYEN, T. T.; PHUNG, D. Q.; VENKATESH, S. Preference networks: Probabilistic models for recommendation systems. In: **Proceedings of the Sixth Australasian Conference on Data Mining and Analytics - Volume 70**. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2007. (AusDM '07), p. 195–202. ISBN 978-1-920682-51-4. Citation on page 58.

TSUGAWA, S.; KITO, K. Retweets as a predictor of relationships among users on social media. **PLOS ONE**, Public Library of Science, v. 12, n. 1, p. 1–19, 2017. Citation on page 86.

VALEJO, A.; DRURY, B.; VALVERDE-REBAZA, J.; LOPES, A. Identification of related brazilian portuguese verb groups using overlapping community detection. In: **Computational Processing of the Portuguese Language**. [S.l.]: Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8775). p. 292–297. Citation on page 178.

VALEJO, A.; VALVERDE-REBAZA, J.; DRURY, B.; LOPES, A. Multilevel refinement based on neighborhood similarity. In: **Proceedings of the 18th International Database Engineering & Applications Symposium**. [S.l.]: ACM, 2014. (IDEAS '14), p. 67–76. Citations on pages 86, 105, and 178.

VALEJO, A.; VALVERDE-REBAZA, J.; LOPES, A. A multilevel approach for overlapping community detection. In: **Proceedings of 2014 Brazilian Conference on Intelligent Systems**. [S.l.]: IEEE, 2014. (BRACIS 2014), p. 390–395. Citations on pages 62 and 178.

VALVERDE-REBAZA, J.; LOPES, A. Link prediction in complex networks based on cluster information. In: **Advances in Artificial Intelligence, SBIA 2012, 21th Brazilian Symposium on Artificial Intelligence**. [S.l.]: Springer, 2012. (Lecture Notes in Computer Science, v. 7589), p. 92–101. Citations on pages 28, 77, 78, 108, 111, 112, 113, 114, 122, and 136.

____. Structural Link Prediction Using Community Information on Twitter. In: **Fourth International Conference on Computational Aspects of Social Networks**. [S.l.]: IEEE, 2012. (CASoN 2012), p. 132–137. ISBN 978-1-4673-4792-1. Citation on page 114.

____. Exploiting behaviors of communities of Twitter users for link prediction. **Social Network Analysis and Mining**, Springer Vienna, v. 3, n. 4, p. 1063–1074, 2013. ISSN 1869-5450. Citations on pages 111 and 177.

____. Link prediction in online social networks using group information. In: **Computational Science and Its Applications - ICCSA 2014 - 14th International Conference, Proceedings, Part VI**. [S.l.]: Springer, 2014. (Lecture Notes in Computer Science, v. 8584), p. 31–45. Citations on pages 67 and 177.

VALVERDE-REBAZA, J.; ROCHE, M.; PONCELET, P.; LOPES, A. Exploiting social and mobility patterns for friendship prediction in location-based social networks. In: **23rd International Conference on Pattern Recognition**. [S.l.]: IEEE, 2016. (ICPR 2016), p. 2526–2531. Citations on pages 64, 65, and 178.

VALVERDE-REBAZA, J.; SORIANO, A.; BERTON, L.; OLIVEIRA, M. C. F. de; LOPES, A. Music genre classification using traditional and relational approaches. In: **Proceedings of 2014 Brazilian Conference on Intelligent Systems**. [S.l.]: IEEE, 2014. (BRACIS 2014), p. 259–264. Citations on pages 61, 105, and 178.

VALVERDE-REBAZA, J.; VALEJO, A.; BERTON, L.; FALEIROS, T.; LOPES, A. A naïve bayes model based on overlapping groups for link prediction in online social networks. In: **Proceedings of the 30th Annual ACM Symposium On Applied Computing**. [S.l.]: ACM, 2015. (SAC' 15), p. 1136–1141. Citation on page 177.

VANUNU, O.; SHARAN, R. A propagation-based algorithm for inferring Gene-Disease associations. In: **German Conference on Bioinformatics**. [S.l.: s.n.], 2008. (GCB '08). Citation on page 75.

VICSEK, T. Complexity: The bigger picture. **Nature**, Nature Publishing Group, v. 418, p. 131, 2002. Citation on page 32.

VIRINCHI, S.; MITRA, P. Similarity measures for link prediction using power law degree distribution. In: ____. **Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part II**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 257–264. Citation on page 71.

WALDROP, M. **Complexity: The Emerging Science at the Edge of Order and Chaos**. [S.l.]: Simon & Schuster, 1993. (A Touchstone Book). Citation on page 32.

WANG, C.; SATULURI, V.; PARTHASARATHY, S. Local probabilistic models for link prediction. In: **Seventh IEEE International Conference on Data Mining (ICDM 2007)**. [S.l.: s.n.], 2007. p. 322–331. Citations on pages 81 and 82.

WANG, D.; CAO, W.; XU, M.; LI, J. Etcps: An effective and scalable traffic condition prediction system. In: **Proceedings, Part II, of the 21st International Conference on Database Systems for Advanced Applications - Volume 9643**. [S.l.]: Springer-Verlag, 2016. (DASFAA 2016), p. 419–436. Citation on page 88.

WANG, D.; PEDRESCHI, D.; SONG, C.; GIANNOTTI, F.; BARABASI, A.-L. Human mobility, social ties, and link prediction. In: **ACM KDD**. [S.l.: s.n.], 2011. p. 1100–1108. ISBN 978-1-4503-0813-7. Citations on pages 66, 91, 94, 153, and 164.

WANG, H.; HUANG, Z.; ZHONG, N.; HUANG, J. Semantically modeling mobile phone data for urban computing. In: **Proceedings of the 9th International Conference on Active Media Technology - Volume 8210**. [S.l.]: Springer-Verlag, 2013. (AMT 2013), p. 203–210. ISBN 978-3-319-02749-4. Citation on page 89.

WANG, H.; WELLMAN, B. Social connectivity in america: Changes in adult friendship network size from 2002 to 2007. **American Behavioral Scientist**, v. 53, n. 8, p. 1148–1169, 2010. Citation on page 58.

WANG, J.; TAN, R.; ZHANG, R.-P.; YOU, F. A recommender system research based on location-based social networks. In: ____. **Social Computing and Social Media: 8th International Conference, SCSM 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17–22, 2016. Proceedings**. Cham: Springer International Publishing, 2016. p. 81–90. ISBN 978-3-319-39910-2. Citation on page 103.

WANG, L.; ZHENG, Y.; XIE, X.; MA, W.-Y. A flexible spatio-temporal indexing scheme for large-scale gps track retrieval. In: **Proceedings of the The Ninth International Conference on Mobile Data Management**. [S.l.]: IEEE Computer Society, 2008. (MDM '08), p. 1–8. Citation on page 88.

WANG, P.; XU, B.; WU, Y.; ZHOU, X. Link prediction in social networks: the state-of-the-art. **Science China Information Sciences**, v. 58, n. 1, p. 1–38, 2014. ISSN 1869-1919. Citations on pages 60, 62, 64, 67, 68, 70, 73, 74, 76, and 78.

WANG, Q.; FLEURY, E. Overlapping community structure and modular overlaps in complex networks. In: ____. **Mining Social Networks and Security Informatics**. [S.l.]: Springer Netherlands, 2013. p. 15–40. Citation on page 51.

WANG, Z.; ZHANG, D.; ZHOU, X.; YANG, D.; YU, Z.; YU, Z. Discovering and profiling overlapping communities in location-based social networks. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, v. 44, n. 4, p. 499–509, 2014. Citation on page 97.

WATTS, D. **Small Worlds: The Dynamics of Networks between Order and Randomness**. [S.l.]: Princeton University Press, 2003. (Princeton Studies in Complexity). Citation on page 48.

WATTS, D.; STROGATZ, S. Collective dynamics of small-world networks. **Nature 393**, n. 6684, p. 440–442, 1998. Citations on pages 26, 46, 48, 51, and 57.

WEI, D.; DENG, X.; ZHANG, X.; DENG, Y.; MAHADEVAN, S. Identifying influential nodes in weighted networks based on evidence theory. **Physica A: Statistical Mechanics and its Applications**, v. 392, n. 10, p. 2564–2575, 2013. Citations on pages 27 and 60.

WEI, L.-Y.; ZHENG, Y.; PENG, W.-C. Constructing popular routes from uncertain trajectories. In: **Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2012. (KDD '12), p. 195–203. ISBN 978-1-4503-1462-6. Citation on page 103.

WHANG, J. J.; RAI, P.; DHILLON, I. S. Stochastic blockmodel with cluster overlap, relevance selection, and similarity-based smoothing. In: **2013 IEEE 13th International Conference on Data Mining**. [S.l.]: IEEE, 2013. (ICDM 2013), p. 817–826. Citation on page 81.

WHITE, H. C.; BOORMAN, S. A.; BREIGER, R. L. Social structure from multiple networks. i. blockmodels of roles and positions. **American Journal of Sociology**, University of Chicago Press, v. 81, n. 4, p. 730–780, 1976. Citation on page 80.

WHITE, H. D.; WELLMAN, B.; NAZER, N. Does citation reflect social structure?: Longitudinal evidence from the "globenet" interdisciplinary research group. **Journal of the American Society for Information Science and Technology**, Wiley Subscription Services, Inc., A Wiley Company, v. 55, n. 2, p. 111–126, 2004. ISSN 1532-2890. Citation on page 57.

WILLIAMS, M. J.; MUSOLESI, M. Spatio-temporal networks: reachability, centrality and robustness. **Royal Society Open Science**, The Royal Society, v. 3, n. 6, 2016. Citation on page 59.

WU, F.; HUBERMAN, B. Finding communities in linear time: a physics approach. **Eur. Phys. J. B**, v. 38, n. 2, p. 331–338, 2004. Citations on pages 110 and 111.

WU, J.; HONG, Z.; PAN, S.; ZHU, X.; CAI, Z.; ZHANG, C. Multi-graph-view subgraph mining for graph classification. **Knowledge and Information Systems**, v. 48, n. 1, p. 29–54, 2016. Citation on page 62.

WU, Z.; CHEN, Y. Link prediction using matrix factorization with bagging. In: **2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)**. [S.l.: s.n.], 2016. p. 1–6. Citation on page 84.

WU, Z.; HUANG, I.; ZHENG, X.; BARTEL, J.; VITKUS, A.; DEWAN, P. A test-bed for facebook friend-list recommendations. In: **Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems**. [S.l.]: ACM, 2015. (EICS '15), p. 222–225. Citation on page 123.

WU, Z.; LIN, Y.; WANG, J.; GREGORY, S. Link prediction with node clustering coefficient. **Physica A: Statistical Mechanics and its Applications**, v. 452, p. 1 – 8, 2016. Citations on pages 28, 78, and 111.

WUCHTY, S. What is a social tie? **Proceedings of the National Academy of Sciences**, v. 106, n. 36, p. 15099–15100, 2009. Citation on page 76.

XIANG, R.; NEVILLE, J.; ROGATI, M. Modeling relationship strength in online social networks. In: **Proceedings of the 19th International Conference on World Wide Web**. [S.l.]: ACM, 2010. (WWW '10), p. 981–990. ISBN 978-1-60558-799-8. Citation on page 88.

XIAO, X.; ZHENG, Y.; LUO, Q.; XIE, X. Finding similar users using category-based location history. In: **Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems**. [S.l.]: ACM, 2010. (GIS '10), p. 442–445. ISBN 978-1-4503-0428-3. Citations on pages 28, 88, 97, and 154.

____. Inferring social ties between users with human location history. **Journal of Ambient Intelligence and Humanized Computing**, v. 5, n. 1, p. 3–19, 2014. ISSN 1868-5145. Citations on pages 28, 88, 97, and 154.

XIE, J.; KELLEY, S.; SZYMANSKI, B. K. Overlapping community detection in networks: The state-of-the-art and comparative study. **ACM Comput. Surv.**, ACM, v. 45, n. 4, p. 43:1–43:35, 2013. Citation on page 62.

XIE, J.; SZYMANSKI, B. K. Labelrank: A stabilized label propagation algorithm for community detection in networks. In: **2013 IEEE 2nd Network Science Workshop (NSW)**. [S.l.: s.n.], 2013. p. 138–143. Citations on pages 110 and 111.

XIE, W.; LI, C.; ZHU, F.; LIM, E.-P.; GONG, X. When a friend in twitter is a friend in life. In: **Proceedings of the 4th Annual ACM Web Science Conference**. New York, NY, USA: ACM, 2012. (WebSci '12), p. 344–347. ISBN 978-1-4503-1228-8. Citation on page 58.

XU-RUI, G.; LI, W.; WEI-LI, W. An algorithm for friendship prediction on location-based social networks. In: THAI, T. M.; NGUYEN, P. N.; SHEN, H. (Ed.). **Proceedings of Computational Social Networks: 4th International Conference, CSoNet 2015, Beijing, China**. Cham: Springer International Publishing, 2015. p. 193–204. Citation on page 101.

____. Using multi-features to recommend friends on location-based social networks. **Peer-to-Peer Networking and Applications**, p. 1–8, 2016. ISSN 1936-6450. Citations on pages 88 and 101.

XU, Z.; PU, C.; SHARAFAT, R. R.; LI, L.; YANG, J. Entropy-based link prediction in weighted networks. **Chinese Physics B**, v. 26, n. 1, p. 018902, 2017. Citations on pages 28 and 76.

YAN, E.; DING, Y. Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. **J. Am. Soc. Inf. Sci. Technol.**, John Wiley & Sons, Inc., v. 63, n. 7, p. 1313–1326, Jul. 2012. ISSN 1532-2882. Citations on pages 57 and 58.

YANG, B.; GUO, C.; JENSEN, C. S. Travel cost inference from sparse, spatio temporally correlated time series using markov models. **Proc. VLDB Endow.**, VLDB Endowment, v. 6, n. 9, p. 769–780, Jul. 2013. ISSN 2150-8097. Citation on page 88.

YANG, J.; LESKOVEC, J. Defining and evaluating network communities based on ground-truth. **Knowledge and Information Systems**, v. 42, n. 1, p. 181–213, 2015. Citations on pages 78 and 124.

YANG, Q.; DONG, E.; XIE, Z. Link prediction via nonnegative matrix factorization enhanced by blocks information. In: **2014 10th International Conference on Natural Computation (ICNC)**. [S.l.: s.n.], 2014. p. 823–827. Citation on page 85.

YANG, Y.; GUO, H.; TIAN, T.; LI, H. Link prediction in brain networks based on a hierarchical random graph model. **Tsinghua Science and Technology**, v. 20, n. 3, p. 306–315, 2015. Citation on page 80.

YANG, Y.; LICHTENWALTER, R.; CHAWLA, N. V. Evaluating link prediction methods. **Knowl. Inf. Syst.**, Springer-Verlag New York, Inc., v. 45, n. 3, p. 751–782, 2015. ISSN 0219-1377. Citations on pages 65, 66, 74, 83, and 162.

YAVEROGLU, O. N. **Graphlet Correlations for Network Comparison and Modelling: World Trade Network Example**. Phd Thesis (PhD Thesis) — Department of Computing, Imperial College London, UK, 2013. Citations on pages 44, 45, and 55.

YE, M.; YIN, P.; LEE, W.-C. Location recommendation for location-based social networks. In: **Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems**. [S.l.]: ACM, 2010. (GIS '10), p. 458–461. ISBN 978-1-4503-0428-3. Citation on page 163.

YE, Q.; WU, B.; WANG, B. Distance distribution and average shortest path length estimation in real-world networks. In: ____. **Proceedings of Advanced Data Mining and Applications: 6th International Conference, ADMA 2010, Part I**. [S.l.]: Springer Berlin Heidelberg, 2010. p. 322–333. Citation on page 33.

YING, J. J.-C.; LEE, W.-C.; YE, M.; CHEN, C.-Y.; TSENG, V. S. User association analysis of locales on location based social networks. In: **Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks**. [S.l.]: ACM, 2011. (LBSN '11), p. 69–76. ISBN 978-1-4503-1033-8. Citation on page 88.

YING, J. J.-C.; LU, E. H.-C.; LEE, W.-C.; WENG, T.-C.; TSENG, V. S. Mining user similarity from semantic trajectories. In: **Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks**. [S.l.]: ACM, 2010. (LBSN '10), p. 19–26. Citations on pages 88 and 97.

YOKOI, S.; KAJINO, H.; KASHIMA, H. Link prediction by incidence matrix factorization. In: **European Conference on Artificial Intelligence (ECAI 2016), Frontiers in Artificial Intelligence and Applications**. [S.l.]: IOS Press, 2016. v. 285, p. 1730–1731. Citation on page 84.

YOOK, S.-H.; JEONG, H.; BARABáSI, A.-L. Modeling the internet's large-scale topology. **Proceedings of the National Academy of Sciences**, v. 99, n. 21, p. 13382–13386, 2002. Citations on pages 56 and 57.

YOON, H.; ZHENG, Y.; XIE, X.; WOO, W. Smart itinerary recommendation based on user-generated gps trajectories. In: **Proceedings of the 7th International Conference on Ubiquitous Intelligence and Computing**. [S.l.]: Springer-Verlag, 2010. (UIC'10), p. 19–34. ISBN 3-642-16354-8, 978-3-642-16354-8. Citation on page 103.

YU, C.; LIU, Y.; YAO, D.; JIN, H.; LU, F.; CHEN, H.; DING, Q. Mining user check-in features for location classification in location-based social networks. In: **2015 IEEE Symposium on Computers and Communication (ISCC)**. [S.l.: s.n.], 2015. p. 385–390. Citation on page 97.

YU, X.; PAN, A.; TANG, L. A.; LI, Z.; HAN, J. Geo-friends recommendation in gps-based cyber-physical social network. In: **Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on**. [S.l.: s.n.], 2011. p. 361–368. Citations on pages 28, 97, and 154.

YUAN, G.; MURUKANNAIAH, P. K.; ZHANG, Z.; SINGH, M. P. Exploiting sentiment homophily for link prediction. In: **Proceedings of the 8th ACM Conference on Recommender Systems**. [S.l.]: ACM, 2014. (RecSys '14), p. 17–24. Citation on page 77.

YUAN, J.; ZHENG, Y.; XIE, X.; SUN, G. Driving with knowledge from the physical world. In: **Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2011. (KDD '11), p. 316–324. ISBN 978-1-4503-0813-7. Citation on page 88.

YUAN, Z.; JIANG, Y.; GIDóFALVI, G. Geographical and temporal similarity measurement in location-based social networks. In: **Proceedings of the Second ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems**. [S.l.]: ACM, 2013. (MobiGIS '13), p. 30–34. Citation on page 97.

ZAKI, M. J.; JR., W. M. **Data Mining and Analysis: Fundamental Concepts and Algorithms**. [S.l.]: Cambridge University Press, 2014. Citations on pages 25 and 60.

ZANIN, M. Can we neglect the multi-layer structure of functional networks? **Physica A: Statistical Mechanics and its Applications**, v. 430, p. 184–192, 2015. Citation on page 60.

ZANIN, M.; PAPO, D.; SOUSA, P.; MENASALVAS, E.; NICCHI, A.; KUBIK, E.; BOCCALETTI, S. Combining complex networks and data mining: Why and how. **Physics Reports**, v. 635, p. 1–44, 2016. Citations on pages 25 and 26.

ZELINSKY, W. The hypothesis of the mobility transition. **Ekistics**, v. 32, n. 192, p. 337–347, 1971. Citation on page 89.

ZENG, W.; SHANH, M.-S.; ZHANG, Q.-M.; LÜ, L.; ZHOU, T. Can dissimilar users contribute to accuracy and diversity of personalized recommendation? **International Journal of Modern Physics C**, v. 21, n. 10, p. 1217–1227, 2010. Citation on page 86.

ZHANG, F.; WILKIE, D.; ZHENG, Y.; XIE, X. Sensing the pulse of urban refueling behavior. In: **Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing**. [S.l.]: ACM, 2013. (UbiComp '13), p. 13–22. Citations on pages 88 and 89.

ZHANG, H.; KAN, M.-Y.; LIU, Y.; MA, S. Online social network profile linkage. In: ____. **Proc. of Information Retrieval Technology: 10th Asia Information Retrieval Societies Conference, AIRS 2014**. Cham: Springer International Publishing, 2014. p. 197–208. Citation on page 62.

ZHANG, J.; KONG, X.; YU, P. S. Transferring heterogeneous links across location-based social networks. In: **Proceedings of the 7th ACM International Conference on Web Search and Data Mining**. [S.l.]: ACM, 2014. (WSDM '14), p. 303–312. ISBN 978-1-4503-2351-2. Citations on pages 97, 103, and 163.

ZHANG, J.; ZHANG, Y.; YANG, H.; YANG, J. A link prediction algorithm based on socialized semi-local information. **Journal of Computational Information Systems**, v. 10, n. 10, p. 4459–4466, 2014. Citations on pages 73 and 76.

ZHANG, Q.-M.; SHANG, M.-S.; LÜ, L. **International Journal of Modern Physics C**, World Scientific Publishing Company, v. 21, n. 6, p. 813–824, 2010. Citation on page 86.

ZHANG, S.; WANG, L.; LIU, Z.; WANG, X. Evolution of international trade and investment networks. **Physica A: Statistical Mechanics and its Applications**, v. 462, p. 752 – 763, 2016. Citation on page 58.

ZHANG, X.; FENG, L.; ZHU, R.; STANLEY, H. E. Applying temporal network analysis to the venture capital market. **The European Physical Journal B**, v. 88, n. 10, p. 260, 2015. Citations on pages 58 and 59.

ZHANG, Y.; PANG, J. Distance and friendship: A distance-based model for link prediction in social networks. In: CHENG, R.; CUI, B.; ZHANG, Z.; CAI, R.; XU, J. (Ed.). **Proceedings of Web Technologies and Applications: 17th Asia-Pacific Web Conference, APWeb 2015, Guangzhou, China**. Cham: Springer International Publishing, 2015. p. 55–66. Citations on pages 28, 95, 96, 97, 101, and 154.

ZHELEVA, E.; GETOOR, L.; GOLBECK, J.; KUTER, U. Using friendship ties and family circles for link prediction. In: **Proceedings of the 2nd International Conference on Advances in Social Network Mining and Analysis**. [S.l.: s.n.], 2008. (SNAKDD'08), p. 97–113. Citations on pages 28, 77, 78, 108, and 111.

ZHENG, K.; ZHENG, Y.; XIE, X.; ZHOU, X. Reducing uncertainty of low-sampling-rate trajectories. In: IEEE. **Data Engineering (ICDE), 2012 IEEE 28th International Conference on**. [S.l.], 2012. p. 1144–1155. Citation on page 91.

ZHENG, Y.; LIU, L.; WANG, L.; XIE, X. Learning transportation mode from raw gps data for geographic applications on the web. In: **Proceedings of the 17th International Conference on World Wide Web**. [S.l.]: ACM, 2008. (WWW '08), p. 247–256. Citations on pages 88 and 89.

ZHENG, Y.; LIU, T.; WANG, Y.; ZHU, Y.; LIU, Y.; CHANG, E. Diagnosing new york city's noises with ubiquitous data. In: **Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing**. [S.l.]: ACM, 2014. (UbiComp '14), p. 715–725. Citations on pages 88 and 89.

ZHENG, Y.; LIU, Y.; YUAN, J.; XIE, X. Urban computing with taxicabs. In: **Proceedings of the 13th International Conference on Ubiquitous Computing**. [S.l.]: ACM, 2011. (UbiComp '11), p. 89–98. Citations on pages 88 and 89.

ZHENG, Y.; ZHANG, H.; YU, Y. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In: **Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems**. [S.l.]: ACM, 2015. (SIGSPATIAL '15), p. 2:1–2:10. Citation on page 88.

ZHENG, Y.; ZHANG, L.; MA, Z.; XIE, X.; MA, W.-Y. Recommending friends and locations based on individual location history. **ACM Trans. Web**, ACM, v. 5, n. 1, p. 5:1–5:44, Feb. 2011. ISSN 1559-1131. Citation on page 103.

ZHENG, Y.; ZHANG, L.; XIE, X.; MA, W.-Y. Mining correlation between locations using human location history. In: **Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems**. [S.l.]: ACM, 2009. (GIS '09), p. 472–475. Citation on page 88.

ZHENG, Y.; ZHOU, X. Computing with spatial trajectories. In: ____. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2011. chap. 8. ISBN 1461416280, 9781461416289. Citations on pages 27, 87, 88, 89, 90, 91, 93, 95, 96, 97, and 153.

ZHOU, C.; MENG, X. Sts: Complex spatio-temporal sequence mining in flickr. In: **Proceedings of the 16th International Conference on Database Systems for Advanced Applications - Volume Part I**. [S.l.]: Springer-Verlag, 2011. (DASFAA'11), p. 208–223. Citation on page 88.

ZHOU, K.; MARTIN, A.; PAN, Q. Semi-supervised evidential label propagation algorithm for graph data. In: ____. **Belief Functions: Theory and Applications: 4th International Conference, BELIEF 2016, Proceedings**. Cham: Springer International Publishing, 2016. p. 123–133. Citations on pages 110 and 111.

ZHOU, T.; LÜ, L.; ZHANG, Y.-C. Predicting missing links via local information. **The European Physical Journal B**, v. 71, n. 4, p. 623–630, 2009. Citation on page 71.

ZHOU, T.; REN, J.; MEDO, M. c. v.; ZHANG, Y.-C. Bipartite network projection and personal recommendation. **Phys. Rev. E**, American Physical Society, v. 76, p. 046115, 2007. Citation on page 86.

ZHU, B.; XIA, Y. An information-theoretic model for link prediction in complex networks. **Scientific Reports**, v. 5, n. 13707, 2015. Citation on page 73.

ZHU, L.; STEEG, G. V.; GALSTYAN, A. Scalable link prediction in dynamic networks via non-negative matrix factorization. **CoRR**, abs/1411.3675, 2014. Citation on page 67.

ZHU, W. Y.; WANG, Y. W.; CHEN, C. J.; PENG, W. C.; LEI, P. R. A bayesian-based approach for activity and mobility inference in location-based social networks. In: **17th IEEE International Conference on Mobile Data Management (MDM)**. [S.l.: s.n.], 2016. v. 1, p. 152–157. Citation on page 103.

ZHU, Y.-X.; Lü, L.; ZHANG, Q.-M.; ZHOU, T. Uncovering missing links with cold ends. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 22, p. 5769 – 5778, 2012. Citation on page 73.

ZOU, Z.; XIE, X.; SHA, C. Mining user behavior and similarity in location-based social networks. In: **2015 Seventh International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)**. [S.l.: s.n.], 2015. p. 167–171. Citations on pages 27 and 97.

ZWEIG, K. A. **On Local Behavior and Global Structures in the Evolution of Complex Networks**. Phd Thesis (PhD Thesis) — University of Tübingen, Germany, 2007. Citation on page 32.

ZWICK, U. Exact and approximate distances in graphs - a survey. In: **Proceedings of the 9th Annual European Symposium on Algorithms**. [S.l.]: Springer-Verlag, 2001. (ESA '01), p. 33–48. Citation on page 33.


ICMC
USP
SÃO CARLOS