CHARACTERIZING MULTIPLE INTERACTIONS IN DYNAMIC ATTRIBUTED NETWORKS BASED ON SOCIAL CONCEPTS

THIAGO H. P. SILVA

CHARACTERIZING MULTIPLE INTERACTIONS IN DYNAMIC ATTRIBUTED NETWORKS BASED ON SOCIAL CONCEPTS

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

ORIENTADOR: ALBERTO H. F. LAENDER
COORIENTADOR: PEDRO O. S. VAZ DE MELO

Belo Horizonte - MG
Dezembro de 2020

THIAGO H. P. SILVA

CHARACTERIZING MULTIPLE INTERACTIONS IN DYNAMIC ATTRIBUTED NETWORKS BASED ON SOCIAL CONCEPTS

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

ADVISOR: ALBERTO H. F. LAENDER
CO-ADVISOR: PEDRO O. S. VAZ DE MELO

Belo Horizonte - MG
December 2020

Silva, Thiago Henrique Pereira.

S586c Characterizing multiple interactions in dynamic attributed networks based on social concept [manuscrito] / Thiago Henrique Pereira Silva.- 2020.
xxvii, 102 f. il.

Orientador: Alberto Henrique Frade Laender.
Coorientador: Pedro Olmo Stancioli Vaz de Melo.
Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f.93-102
.
1. Computação – Teses. 2. Reds sociais – Teses. 3. Redes de relações sociais – Teses. 4. Redes complexas – Teses. 5. Computação social – Teses. I. Laender, Alberto Henrique Frade. II. Melo, Pedro Olmo Stancioli Vaz de.III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV.Título.

CDU 519.6*75(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

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

FOLHA DE APROVAÇÃO

Characterizing Multiple Interactions in Dynamic Attributed Networks based on Social Concepts

THIAGO HENRIQUE PEREIRA SILVA

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Alberto Henrique Frade Laender - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Pedro Olmo Stancioli Vaz de Melo - Coorientador
Departamento de Ciência da Computação - UFMG

Jussara Marques de Almeida Gonçalves

Profa. Jussara Marques de Almeida Gonçalves
Departamento de Ciência da Computação - UFMG

Daniel Ratton Figueiredo

Prof. Daniel Ratton Figueiredo
Programa de Engenharia de Sistemas e Computação - UFRJ

Artur Ziviani

Prof. Artur Ziviani
Ciência da Computação - LNCC

Wagner meira junior

Prof. Wagner Meira Júnior
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 17 de Dezembro de 2020.

To Anita, Cláudia, Andréa and Renata.

Acknowledgments

My warm thanks to everyone who contributed in some way to the development of this work, particularly my family. Also, I would like to thank my advisors for their confidence in our studies and contributions to ideas and reviews. Additionally, I thank my colleagues of LBD and DCC-UFMG, as well as the excellent staff of our Computer Science Department. Finally, I am very grateful for the support of professors Alberto H. F. Laender, Jussara M. Almeida and Wagner Meira Jr for sharing personal stories and encouraging me to continue my studies.

"Run, rabbit run
Dig that hole, forget the sun
And when at last the work is done
Don't sit down, it's time to dig another one"
(David Gilmour / Richard Wright / Roger Waters)

Resumo

Caracterizar interações dinâmicas é uma questão importante ao analisar redes sociais complexas. Com base na autonomia estrutural que informa quando as pessoas estão estreitamente conectadas umas às outras com extensos laços que atuam como pontes além delas, reforçamos a importância de conceitos sociais como fundamental para a compreensão da complexidade que envolve os atores e suas relações. Nesse sentido, discutimos como modelar múltiplas interações em redes dinâmicas com atributos e propomos um método para classificar nós e arestas dinâmicas com base em relações nó-atributos. Como resultado, o método captura a força das interações sociais e como o conhecimento é transferido pela rede social. Em seguida, discutimos e ilustramos as diferenças de interações sociais em diferentes redes sociais acadêmicas e comunidades de perguntas e respostas. Com base no posicionamento estratégico de um determinado ator em uma estrutura social, validamos estatisticamente nossa estratégia proposta por meio de propriedades de rede. Além disso, realizamos uma análise de sensibilidade destacando-a em termos de sua robustez para lidar com aspectos de tempo, poder discriminativo dos atributos e cenários aleatórios. Por fim, propomos estratégias não-supervisionadas e supervisionadas que aplicam nosso método para identificar nós influentes em uma estrutura social, os quais superam as métricas de rede tradicionais e outros algoritmos baseados em conceitos sociais.

Palavras-chave: Redes sociais, Redes Diâmicas com Atributos, Computação Social.

Abstract

Characterizing dynamic interactions is currently an important issue when analyzing complex social networks. Based on the structural autonomy that informs when people are tightly connected to one another with extensive bridge ties beyond them, we reinforce the importance of the network theory paradigm as fundamental for understanding the complexity that involves actors and their relationships. In this regard, we discuss how to model multiple interactions in dynamic attributed networks and propose a classification method that classifies nodes and dynamic edges based on node-attribute relationships. As a result, it captures the strength of social interactions and how knowledge is transferred across the network. Then, we unveil and illustrate the differences of social interactions in different academic social networks and Q&A communities. Based on the strategic positioning of a particular actor in a social structure, we statistically validate our proposed strategy by means of network properties. Moreover, we perform a sensitivity analysis by stressing it in terms of its robustness to deal with aspects of time, discriminative power of attributes and random scenarios. Finally, we propose unsupervised and supervised strategies that apply our method to identify influential nodes in a social structure, which outperform traditional network metrics and other social-based algorithms.

Keywords: Social Networks, Dynamic Attributed Networks, Social Computing.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

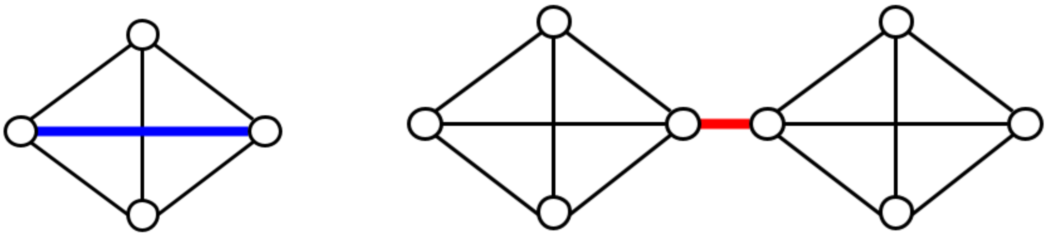
Characterizing multiple social interactions involving actors in social networks has become increasingly popular in recent years because of their importance to better understand the behavior and evolution of a social structure. Social networks capture the interactions among people, which represent the social network structure of the system. There are cases where the network is not explicitly defined, being inferred from the history of interactions among a set of actors, for instance, in situations such as mobile user encounters [Vaz de Melo et al., 2015], e-mails [Paruma-Pabón et al., 2016] and strategic alliances [Inkpen and Tsang, 2005]. However, when the interactions are not properly mapped (e.g., every interaction is simply represented as a link), the underlying social structure is not accurately represented in such networks [Vaz de Melo et al., 2015], i.e., the mapping function must choose the relevant features from a set of interactions that make a relationship exist or not in a network.

Graphs can be constructed from explicit relationships such as Facebook friendships [Adamic and Adar, 2003] and character relationships from novel summaries [Chaturvedi et al., 2017]. Graphs can also be constructed from implicit relationships, which are inferred from interactions among the actors. This is a more challenging task, because of random interactions or strong relationships that, for some reasons, were not represented in the data. Additionally, interactions may be of several types, which can or cannot be described in the data. When additional information is available about the interactions, one can use this information to construct the so called *attributed networks*, where nodes and edges have attributes, or features, associated with them. Attributes can be directly associated with nodes (e.g., age, gender, hometown, etc.) and with edges (e.g., parent-child, adviser-advisee, follower-followee, etc.). The idea is that these attributes can provide more information about the system and help to accurately characterize the actors and their relationships.

Furthermore, we also have the case where the attributes associated with actors and edges change over time (i.e., the so called *dynamic attributed networks*), whether in terms of location (a new job or country), relationships with other people (childhood friends who no longer participate in their network) or new skills acquired. Although such networks provide more information about the social motivation involving each interaction [Orman et al., 2014; Rezaei et al., 2017; Yo and Sasahara, 2017], this adds an extra layer of complexity to the problem of characterizing actors and relationships from sets of interactions and their attributes [Aggarwal et al., 2016, 2017]. In this case, a single graph constructed from the interactions might not be enough to represent the social system, and a temporal graph might be necessary. To overcome this situation, we can explore the duration of the relationships among the nodes over time, as well as consider the persistence of attributes involved in each interaction.

Once there are models that incorporate all specificities of a particular social network, a challenge is to extract knowledge from the analysis of the history of interactions among a set of actors. Moreover, Barabási [2009] reinforces the importance of the network theory paradigm as fundamental for understanding the complexity that involves actors and their relationships. Traditionally, several works have investigated topological properties and patterns of social networks in order to define the behavior of their actors and measure the strength of their relationships [Huang et al., 2018; Leão et al., 2018; Levchuk et al., 2012; Newman, 2004]. Exploring the behavior and the dynamics of the actors in a social network is essential for a better understanding of its social structure, which is usually characterized by graphs that capture the social aspects involved [Medo et al., 2016; Silva et al., 2015a; Yang and Xie, 2016]. For instance, Newman [2004] measures the influence of the nodes in a network based on their proximity and the number of shortest paths between them. In fact, network science is so valuable because it only relies on the relationships among the actors, which means that it can be applied to practically any system.

Regarding social perspectives, some studies have explored the notion of *social capital* given by the strategic positioning of a particular actor in a social structure [Burt, 2005; Coleman, 1988, 1994; Granovetter, 1973; Silva et al., 2014; Valverde-Rebaza et al., 2018]. Based on the premise that actors can make a network stronger by integrating different parts, Granovetter [1973] defines the concept of *weak ties* as being those important relationships that make a network more cohesive by means of the creation of bridges between communities. In addition, Coleman [1988] emphasizes that a social structure is formed with high degree of trustworthiness among members of a group. In another seminal work, Burt [2005] describes a *structural hole* as the gap formed by individuals who have complementary knowledge, and then defines as *brokers* those individuals that hold certain positional advantages due to their good location in the social structure. Indeed, as discussed by Aral [2016], the most influential sociological theories of networks explore bridging ties (e.g., connecting different parts



(a) Closure

(b) Brokerage

Figure 1.1: Example of closure and brokerage in a graph.

of a network) and cohesive ones (e.g., building a trust circle), which provide more advantage when accessing information passing through a network.

This thesis contributes to the aforementioned discussion related to relationships in dynamic attributed networks in terms of the social tie of individuals and their associated attributes. We also consider the benefit derived from individuals who occupy specific places in a social structure, particularly assuming *closure* and *bridging* as forms of social capital [Burt, 2005; Coleman, 1994]. In other words, this powerful analysis determines how different nodes can play structural distinct roles in a social network, such as joining in a tightly-knit group or connecting such groups [Easley and Kleinberg, 2010]. For achieving this, we rely on Burt's definition of two social concepts, *closure* and *brokerage* [Burt, 2005]:

- **Closure.** It occurs when there are several strong connections between individuals, which can be interpreted as a dense network. Thus, this social structure can be understood as the ability of aggregating individuals with similar social patterns.

- **Brokerage.** It captures the behaviour of individuals when acting as an intermediaries between two or more closed groups to connect different parts of a network. Thus, this social structure can be understood as the ability of creating bridges with diversified social patterns.

Figure 1.1 illustrates both concepts. A closure occurs when individuals have dense ties involving members of a same group. As discussed by Coleman [1988], a closure creates trustworthiness in a social structure, since it proliferates with obligations and expectations within of a closed group. Indeed, individuals are more likely to transfer knowledge with someone they trust, therefore improving the network cohesion [Levin and Cross, 2004]. On the other hand, a brokerage occurs when individuals act as a *broker* or as a *hub*. As discussed by Granovetter [1973], it is more likely that people get jobs by accessing information from acquaintances (*weak ties*, i.e., bridges that represent distant and infrequent interactions) than from groups of closest friends (*strong ties*). Indeed, individuals already have access to information propagated within their social circle. In this way, new information tends to be acquired from other social groups, thus improving the network strength.

As discussed by Easley and Kleinberg [2010], studies on the *connectedness* of modern society have become a growing public fascination. Indeed, new perspectives and techniques from different areas (e.g., applied mathematics and economy) have contributed to understand how highly connected complex networks work, ranging from graph theory to property rights. Thus, based on how actors play structural roles in networks, the relevance of our study is to contribute with a social perspective in order to better analyze the complexity involving these individuals. For example, we would like to bring more information for decision making by highlighting what are the long-term links, what are the most profitable paths, which actors can have more influence, at what moment there is maturity in a community, etc. For this, as previously discussed, we aim to associate the social concepts of *brokerage* and *closure* to nodes and edges in order to inform how they are positioned in a social structure.

Next, we present our problem statement. Then, we list our research goals and describe the thesis organization.

1.1 Problem Statement

In this thesis, we address the problem of characterizing actors and their relationships in dynamic social networks based on social concepts. As an additional complexity factor, its scenario encompasses multiple interactions over time, as well as attributes associated with each interaction. With such an information, we indent to provide a novel characterization about how real world complex systems work.

More formally, given a set of actors and a sequence of interactions with them along the time, consider a temporal multigraph $G = (V, E)$, where $V = \{\mathscr{V}_1, ..., \mathscr{V}_n\}$ and $E = \{\mathscr{E}_1, ..., \mathscr{E}_n\}$ represent the set of nodes and edges, respectively. Thus, the multigraph $\mathscr{G}_k = (\mathscr{V}_k, \mathscr{E}_k)$ denotes a set of nodes and edges created at a discrete interval k. In addition, each edge $(u, v) \in \bigcup_{i=1}^{n} \mathscr{E}_i$ may be associated with a set of attributes. As a result, it ensures a social concept for all nodes and edges at all-time intervals, thus capturing how they are positioned in a social structure. That is, it associates them with a social class such as *closure* or *brokerage*. We reinforce that our proposal is an unsupervised method to classify the social roles of nodes and measure the importance of their relationships, since there are no social labels established in advance.

To approach this characterization problem, we consider the following steps, as depicted in Figure 1.2:

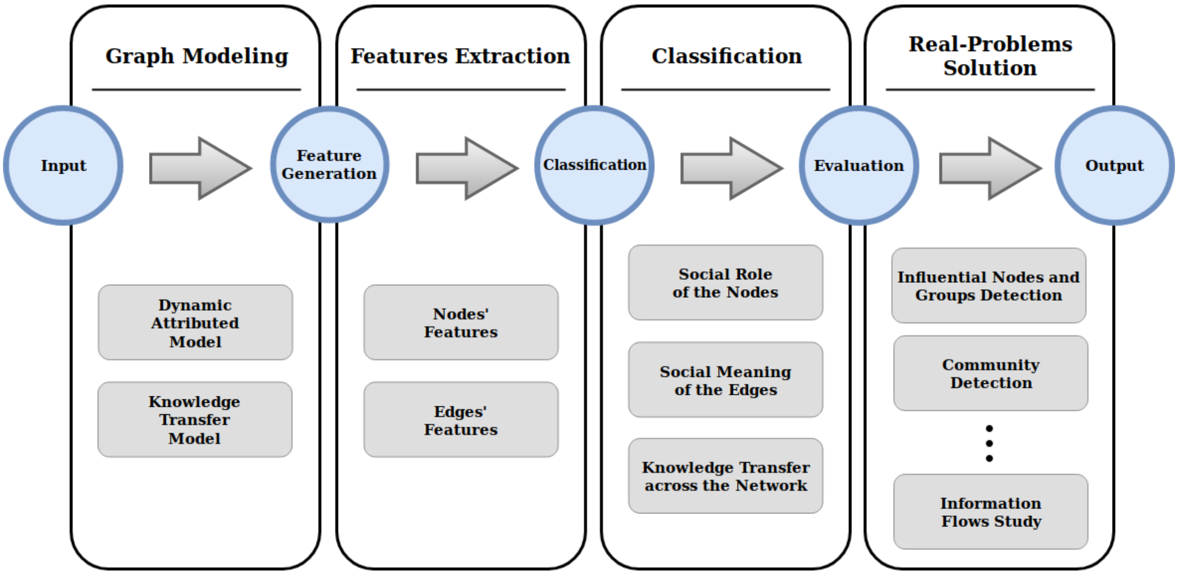


Figure 1.2: Overview of the steps for characterizing nodes and edges.

- *Step 1: Graph Modeling.* This step consists in constructing a temporal attributed graph from the actors and their interactions, considering two model perspectives:

  (i) **Dynamic Attributed Model.** This model enables us to explore the interaction history of the actors and unveil the persistence of their relationships along the time;

  (ii) **Knowledge Transfer Model.** This model enables us to analyze the dynamics involving the actors in terms of how their knowledge is transferred across the network.

- *Step 2: Feature Extraction.* This step consists in identifying social features and network properties associated with the actors. For this, we explore network metrics in order to establish both the strength and the meaning of the social ties. Moreover, from a social perspective, such as the advantage created by a good location in the social structure (i.e., the notion of social capital), we analyze real entities that act as bridges with well-defined closed groups [Burt, 2005].

- *Step 3: Classification.* This step consists in assigning social values to the nodes and edges in order to better understand the behavior and evolution of a social structure. For this, we use features based on network properties and attributes in order to propose functions to characterize nodes and their relationships as follows:

  (i) **Social Role of the Nodes.** From a social perspective, the social role of a node is defined according to its position in a specific group (e.g., father/mother/child, advisor/advisee, and predator/prey). Thus, our goal is to identify the social role of the nodes.

EXAMPLE 1. In coauthorship networks, we want to know the social behavior of researchers. In this way, we can construct an attributed coauthorship network in which nodes represent researchers, edges represent that two researchers have at least one coauthorship together and node attributes represent their specific competence (databases, networks, etc.). In this context, we can infer that a researcher that has a long-lasting association with attributes like "relational model", "data definition" and "query language" is likely to have an authority over them. Thus, we can classify this researcher as a "hub", since it has the potential to spread knowledge from the databases domain.

(ii) **Social Meaning of the Edges.** As interactions between nodes are far from arbitrary in social networks [Bhagat et al., 2011], we consider that edges carry some information. Thus, we aim to understand the social meaning of the edges in order to measure the strength of the interactions.

EXAMPLE 2. In a friendship social network, we want to know the strength of the relationships between people. For this, we construct a network in which nodes represent people, edges represent relationships between them (e.g., a partnership or an acquaintance) and attributes represent the persistence of these relationship along the time (e.g., the period during which each kind of relationship lasts). Then, we can classify the strengths of the edges as "strong", indicating a social tie between relatives, whereas a "weak" one could represent a social tie with a co-worker.

(iii) **Knowledge Transfer through the Edges.** In order to characterize how knowledge is transferred across the network, we need to observe how the flow of attributes occur, as well as to identify its origins and terminations.

EXAMPLE 3. Consider a Questions and Answers social network in which nodes represent users and edges link who posted a question to be answered/commented by whom. In this context, let a question posted by a lay user (i.e., non-expert or novice). Thus, an edge can be considered with high potential of knowledge transfer when it is answered by a user with expertise in the area of that specific question.

- *Step 4: Application.* This final step consists in applying our model and classification methods to different social scenarios (i.e., showing that our approach is general and can be applied to any system) in order to provide a novel understanding about real world complex systems, for instance, in situations such as measuring influential individuals and groups, detecting communities, and identifying information dissemination strategies.

1.2 Research Goals

Given the problem of characterizing nodes and edges, our overall goal is to provide a novel understanding about how real world complex systems work, thus building models and strategies for processing dynamic interactions with attributes. As a result, we expect to create an environment that promotes a new characterization perspective based on social concepts. To achieve this goal, we state our associated research goals (RGs) as follows:

RG1 - ***Modeling Dynamic Attributed Interactions.*** Our first research goal focuses on characterizing and modeling dynamic interactions in attributed social networks. Specifically, we deal with social scenarios that have multiple social facets. For this, a feasible approach is to model several aspects by associating attributes to each interaction over time. To address this goal, we discuss how to model social interactions over time based on their associated attributes. Assuming that a node that has a long-lasting association with specific attributes tends to consider them important, we abstract such attributes by representing them as new nodes in order to investigate the strength of node-attribute relationships. We also explore the roles of the nodes in a social structure, thus representing different social concepts to explain the evolution of a social network (e.g., explaining how networks become stronger by the concepts of brokers [Burt, 2005], weak ties [Granovetter, 1973] and newcomers on the networks [Guimera et al., 2005]). Although characterizing the behavior of actors is very broad and depends on the context of the social network to be modeled, an additional issue consists in proposing a general model that can be applied to any system.

RG2 - ***Classifying nodes and edges.*** As our second goal, we aim to classify the dynamic interactions in attributed social networks by exploring social concepts involving actors and their associated attributes. To achieve this goal, we consider the persistence of the attributes in the nodes' interaction history. Then, we classify the social role of the nodes and measure the strength of the relationships among them. In addition, by analyzing the dynamics involving the actors, we also discuss how knowledge transfers between them. In this regard, we explore the dynamics involving individuals in terms of how their attributes are transferred across the network. Then, we classify the edges and determine its direction by considering the flow of attributes across the network.

As a result, we expect to characterize different scenarios and comparing with other algorithms, thus portraying how good the proposed method succeeds in extracting the peculiarities underlying such social contexts.

RG3 - ***Application*****:** Besides providing a new perspective to discover knowledge from the node-attribute relationships, another issue involved is the applicability of our proposed method. We focus on this matter by mapping the importance of nodes and their relationships according to the classes assigned to them, thus revealing how better they are positioned in a social structure. In this way, we can unveil new social facets to be combined in order to improve other studies.

1.3 Contributions

Based on the aforementioned research goals, the main contributions of this thesis are:

1. A node-attribute graph model that captures the social tie of individuals and their associated attributes. This dynamic attributed model enables us to mine multiple interactions over time. As a result, we can better define the social roles of individuals in a social structure, as well as to determine the strength of their relationships.
2. A new method to classify nodes and their relationships based on temporal node-attributes that considers:
    a) the social role of the nodes;
    b) the social meaning of the edges;
    c) how knowledge is transferred across the network.
3. A social-based method that relies on the proposed model and classifiers that measures the importance of researchers in social academic networks.

1.4 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 discusses related work on social network analysis, focusing on how to model dynamic interactions and how to explore social concepts. Chapter 3 introduces the graph model proposed to mine multiple interactions over time, whereas Chapter 4 proposes strategies to classify nodes and edges based on social concepts. Chapter 5 presents the experimental methodology underlying the social networks considered and how to evaluate the proposed characterization. Chapter 6 analyzes and discusses the results of applying our classification method in different social contexts. Chapter 7 summarizes the results of our experimental validation and Chapter 8 demonstrates the applicability of our method for ranking nodes. Finally, Chapter 9 concludes this thesis by summarizing our findings and discussing some directions for further investigation.

Chapter 2

Basic Definitions and Related Work

In this chapter, we present some basic definitions and discuss related work covering three specific topics: (i) analysis of dynamic attributed networks, (ii) node and edge classification tasks, and (iii) social concepts.

2.1 Dynamic Attributed Networks

First, let us define social network features to better describe the complexity of the interactions that involves the actors in a social structure. As a result, such feature-rich networks provide more information and help us to accurately characterize the actors and their relationships. Next, we define dynamic attributed networks and show/discuss how to represent their dynamics over time.

2.1.1 Attributed Networks

Graphs are constructed considering a set of entities (nodes) and their relationships (edges). As these relationships evolve to other kinds of interactions (e.g., encounters, phone calls, messages exchanged, etc.), they become more complex, thus capturing different social meanings. In this way, a more general approach is required to model these specificities by using edge or node attributes [Aggarwal et al., 2016, 2017; Rezaei et al., 2017; Shah et al., 2016]. Thus, let be a graph $G = (V, E)$, where V denotes the set of nodes and E denotes the set of edges. We also represent by $\mathscr{A}_{nodes}$ and $\mathscr{A}_{edges}$ the sets of all possible attributes associated with nodes and edges, respectively. Then, we associate a subset of attributes to each node or edge.

- **Node attributes.** Let Γ: $n \in V \rightarrow A$ be a function that performs the mapping of each node n to a subset of attributes A contained in the set $\mathscr{A}_{nodes}$.

- **Edge attributes.** Let Φ: $e \in E \rightarrow A$ be a function that performs the mapping of each edge e to a subset of attributes A contained in the set $\mathscr{A}_{edges}$.

Such attributed networks promote rich information about connections between nodes, in which the presence of attributes associated with nodes and edges indicates the social value associated with each interaction. This allows us to know the characteristics of the nodes as well as the types of relations between them. For example, let be a network constructed from the characters and their interactions in a novel. Their characteristics could be drawn by the attributes of the nodes to designate age, gender, hometown, profession, etc. In a similar way, we can inform the kinship between two entities by means of attributes from the edge that binds them, thus characterizing the relationships as brother-sister, grandchild-grandparent, etc.

In this regard, some previous works have used attributes to better understand the social motivation involving each interaction [Orman et al., 2014; Rezaei et al., 2017; Yo and Sasahara, 2017]. For instance, Rezaei et al. [2017] characterize political subgroups based on the differences of their associated attributes (e.g., health and taxation), thus contrasting Democrats and Republicans in the US Congress by means of their national political interests. By considering online reviews on Amazon as textual attributes of edges, Jindal and Liu [2008] analyze linguistic indicators in order to detect spam activities. Yet, Yo and Sasahara [2017] studied personal attributes based on social data for the attribute prediction task, whereas Orman et al. [2014] proposed a method for characterizing communities in dynamic attributed networks by exploring topological properties and nodal features (e.g., publication venues as attributes).

2.1.2 Static and Temporal Networks

A static network is one that consists of a single instance of a graph that represents a complex system. For example, we can represent molecular structures by a static network, where atoms and chemical bonds are represented by nodes and edges, respectively. However, static representations do not properly show how such chemical compounds are formed, i.e., there is a lack of information on how they are held together by a variety of forces and reactions. For a more robust analysis to understand the overall evolution of such complex networks, we can dissect them in several temporal instances. Hence, a time-varying graph is represented by adding a time denotation, where it can indicate a time interval or determine the beginning and the end of each interaction. According to Holme and Saramäki [2012], there are three basic representations of such so-called temporal networks:

1. *Contact sequences.*: In this case, the network can be represented as a set of triples (i, j, t), where i and $j \in V$ are the nodes, and $t \in T = \{t_1, t_2, ..., t_n\}$ denotes when an interaction occurred. This type of representation is indicated when there is a strong tendency that interactions occur at different discrete intervals and, if aggregated, we would miss many of the unique relationships.

2. *Interval graphs.* Likewise the previous representation, there is also a triple, but now it defines the beginning and end of an interaction, i.e., $t \in T = \{(t_1, t'_1), (t_2, t'_2), ..., (t_n, t'_n)\}$. For example, we can map phone calls between individuals indicating that it started at t and ended at t', thus lasting $t' - t + 1$. This representation considers a set of intervals over which each edge is active, thus we can use it when the duration of the interactions is non-negligible.

3. *Snapshots.* The final case is when the networks are represented as a series of static networks, one for each time interval. Thus, the sets of nodes and edges become, respectively, $(\mathscr{V}_1, \mathscr{V}_2, ..., \mathscr{V}_n)$ and $(\mathscr{E}_1, \mathscr{E}_2, ..., \mathscr{E}_n)$. This case is used when it is possible to aggregate sufficient information in each one of the subgraphs $(\mathscr{V}_k, \mathscr{E}_k)$ without loss of generality or in order to simplify an analysis.

Although there are temporal representations that may describe best very specific complex networks, many analyses consider aggregating temporal graphs, either for convenience, lack of data or to reduce the complexity of the study [Barrat et al., 2013]. This is nothing more than to dismiss the temporal issues, thus joining all the nodes and edges in a final static graph comprising all the information from the beginning until a final time t. That is, the sets of nodes and edges become, respectively, $V = \bigcup_{i=1}^{t} \mathscr{V}_i$ and $E = \bigcup_{i=1}^{t} \mathscr{E}_i$.

On the other hand, distinct studies focus on analyzing the temporal evolution of the interactions and, for this, they inspect each subgraph step by step at specific time intervals. In other words, the interest is to unravel particular changes, as well as to examine precisely the overall dynamics of a specific complex system [Braha and Bar-Yam, 2009]. In this case, an alternative consists in analyzing social networks by means of snapshots as previously defined. For instance, Silva et al. [2015a] characterize the moving properties and the behavioral profile of how researchers move around publication venues stratified in terms of their quality, whereas Brandão et al. [2017] address how social roles change over time. Yet, Beutel et al. [2013] find fraudulent behavior by analyzing the social interaction between users and pages on the Facebook with temporal edge attributes (i.e., stating the times when *likes* occurred). In another context, Medo et al. [2016] statistically model the importance of those individuals that have effectively discovered items on e-commerce networks that later become quite popular, thus emphasizing a deeper understanding of their behavior and roles.

In this thesis, we focus not only on dynamic networks, but we also consider node-attribute associations. Thus, we expand the above considerations by defining attribute mapping functions for all time intervals, as well as allowing multiple edges at the same instant. Moreover, we abstract the attributes as entities (i.e., artificial nodes) in order to explore the dynamics between them over time. This enables us to understand the evolution of social structures, in which the persistence of attributes over time indicates the social value associated with each interaction. For instance, by using this approach, an academic social network can be studied by means of the dynamics of the attributes associated with its members in order to identify research trends (e.g., new research subjects or hot topics).

2.2 Node and Edge Classification

Classification of edges and nodes can be seen as an association of labels with each of its instances. For example, in a friendship network, we can classify nodes by labeling them as either *adult* or *child*. Similarly, we can classify the social bond between them by labeling their edges as being *acquaintant* or *relative*. This classification process can be seen as to predict a label or class for unlabeled entities. Then, we can perform it by a supervised approach that requires a set of properly labeled entities (i.e., a training set) to automatically indicate the labels, but also by an unsupervised method when the definition of the classes (and perhaps the number of them) are not known in advance [Zaki et al., 2014]. Formally, based on the model presented by Aggarwal et al. [2016], we can define the edge and node classification tasks as follows:

- *Edge Classification.* Given a graph $G = (V,E)$ and a set $E_l \subseteq E$ of labeled edges. Let X be the set of possible labels, and $X_l = \{x_1, x_2, ..., x_l\}$ be the associated labels on edges in the set E_l. The classification task is to infer labels X on all edges of the graph.

- *Node Classification.* Given a graph $G = (V,E)$ and a set $V_l \subseteq V$ of labeled nodes. Let Y be the set of possible labels, and $Y_l = \{y_1, y_2, ..., y_l\}$ be the associated labels on nodes in the set V_l. The classification task is to infer labels Y on all nodes of the graph.

In addition, we can denote the subsets of unlabeled edges and nodes, respectively, as $E_u = E \setminus E_l$ and $V_u = V \setminus V_l$.

In general, node and edge classifications learn similar features from annotated entities based on network topology (i.e., the sets E_l and V_l) [Aggarwal et al., 2017; Dai et al., 2016; Henderson et al., 2012; Gilpin et al., 2013; Henderson et al., 2011]. For instance, in attributed networks built from Enron (keywords in e-mails as attributes) and Twitter (tags in the messages as attributes) datasets, Aggarwal et al. [2017] classified the edges by learning the structure and content of labeled ones. The identification of such characteristics provides insights to supervised-learning methods infer the labels of the edges and nodes in the respective sets E_u and V_u.

As discussed by Bhagat et al. [2011] in their book, the node classification problem brings a new understanding to social network analysis by exploring the specificities of the nodes such as their activities, connections, opinions, thoughts, etc. For instance, we can classify the nodes in order to suggest new connections to individuals based on similar interests [Kajdanowicz et al., 2010], to recommend products from the interests of other individuals with overlapping characteristics [Asiri and Miri, 2016], and to identify spammers based on the social aspects of the community structure [Bhat and Abulaish, 2013]. By characterizing expert behavior of users in Questions & Answers communities, Yang et al. [2014] assigned two labels to them: (i) *sparrows*, as being those who answered alone the vast majority of the questions, and (ii) *owls*, as being those experts in the discussed topic. They concluded that such behaviors improve the knowledge creation and the community participation, thus guaranteeing responsive and useful answers.

Regarding the task of characterizing relationships, Trevithick and Clippinger [2008] filed a patent of a method to characterize edges based on the pattern and purposes of communications between members of social networks. For this, their model explores the structure of the messages, group performance, communication patterns and message-routing processes. In another context, Leskovec et al. [2010] combined signed networks properties (e.g., social balance) with machine-learning techniques to perform a prediction task of assigning positive (e.g., friendship) or negative (e.g., antagonism) to the relationships in different online social networks.

On the other hand, our proposal is a self-organization method to classify the social roles of nodes and measure the importance of their relationships, since there are no social labels established in advance (i.e., $V_l = \emptyset$ and $E_l = \emptyset$). Besides exploring network properties, our method allows attributes associated with each interaction to become artificial nodes, thus providing a new characterization perspective. In addition, the sets of possible labels are based entirely on structural roles that they represent in the social structure.

2.3 Social Concepts

As pointed out by Aral [2016] and Easley and Kleinberg [2010], social concepts promote powerful analyses of how different nodes can play structurally different roles in a social network. For example, the basic principle of *triadic closure* [Easley and Kleinberg, 2010] establishes that in the case of a strong tie between a node A with B and C, then there is a great likelihood that B and C will become connected (i.e., B and C will close the third side of the *triangle*). Based on this, we can observe, for instance, how integrated a social structure is in terms of its proportion of existing triangular structures (e.g., *clustering coefficient*) [Newman, 2003; Watts and Strogatz, 1998]. This kind of analysis brings new perspectives to understand the complexity that involves the actors and their relationships, as well as the evolution of the structures underlying these social networks [Barabási, 2009; Easley and Kleinberg, 2010; Kossinets and Watts, 2006; Newman et al., 2006; Sun et al., 2013]. For instance, Barabási et al. [2002] captured the social tie importance by observing the topology and the internal behavior of the nodes in coauthorship networks, whereas Sun et al. [2013] proposed a model based on social interactions among individuals to analyze the social dynamics of science in terms of scientific disciplines. By analyzing a dynamic social network formed by students, faculty and staff considering their affiliations and shared activities, Kossinets and Watts [2006] describe the evolution of such a network as a combined effect of its topology and organizational structure.

Indeed, social perspectives can capture the benefits deriving from a good location in a social structure, which can be formally specified by the notion of *social capital* [Burt, 2005, 2009; Coleman, 1994; Granovetter, 1973; Silva et al., 2017]. Based on the premise that actors can make a network stronger by integrating different parts, Granovetter [1973] defines the concept of *weak ties* as being those important relationships that make a network more cohesive by means of the creation of bridges. Yet, Coleman [1988, 1994] defines the concept of *network closure*, in which the number of relationships of each individual is close to the maximal number of individuals (i.e., a dense network), as a form of social capital because it ensures early access to information and facilitates people to trust one another. In other words, it emphasizes that without high degree of trustworthiness among members of a group a social structure does not exist. In another seminal work, Burt [2009] describes a *structural hole* as the gap formed by individuals who have complementary knowledge. Then, he defines as *brokers* those nodes that hold certain positional advantages due to their good location in the social structure. Moreover, he defined *social capital* as a tension between *closure* and *brokerage*, where the former corresponds to the Coleman's definition of social capital and the latter refers to the ability of acting between different groups across structural holes. Therefore, a good social strategy is to position where people are tightly connected to one another (i.e., building a trust circle) with extensive bridge ties beyond them (i.e., connecting different parts of a network).

Based on such social concepts, Feng et al. [2018] used structural holes to identify the most central and bridging group of individuals in a network. Likewise, Zhang et al. [2019] explore the roles of key nodes as bridges that promote knowledge transfer (e.g., patent citations), thus ranking the importance of them according to their collaborative relationships. In another context, Inkpen and Tsang [2005] studied how social aspects can facilitate knowledge transfer between individuals. They identified structural, cognitive and relational dimensions on three types of network (corporate networks, strategic alliances and industrial districts), concluding that strategies for facilitating knowledge transfer varies across networks, since the effectiveness and efficiency of this process depends on how individuals build social capital. Yet, Sanz-Cruzado and Castells [2018] explore strong ties (links within communities) and weak ties (links between communities), thus showing that bridges work as enhancers of the structural diversity in the Twitter social network. By exploring specific indicators (e.g., centrality metrics and publication count) from coauthorship networks comprising 137 information systems scholars, Li et al. [2013] investigated several strategies for leveraging social capital. As a result, they concluded that by improving the betweenness centrality of a scholar can significantly increase her research impact (e.g., citation count). More recently, Chen and Liu [2019] investigated how a group of actors takes advantage in a social structure by acting as brokers to cover an entire network. They showed that groups formed by actors with different influencing power tend to achieve an advantageous position.

Several other studies analyze social networks based on particular social concepts such as tie strengths [Brandão and Moro, 2015; Brandão et al., 2017; Lü et al., 2016; Vaz de Melo et al., 2015], homophily [Liao et al., 2018; Silva et al., 2014], friendship granularities [Adamic and Adar, 2003; Shi et al., 2007; Valverde-Rebaza et al., 2018] and social influence [Gupte et al., 2011; Jiang et al., 2017; Tang, 2017], as well as on structures such as triadic closure and social balance [Easley and Kleinberg, 2010; Huang et al., 2018]. For instance, Silva et al. [2015b] explore the concept of building bridge to reveal the social ties of individuals with their communities in order to measure their degree of social influence. To do so, they apply social capital concepts to measure the potential of knowledge acquired and the strength of sharing information. Yet, Levchuk et al. [2012] propose an approach to learn and detect network patterns such as repetitive groups of people involved in coordinated activities. Based on the information shared between nodes, Adamic and Adar [2003] measure the strength of relationships by analyzing the similarity between messages exchanged between individuals, whereas Leão et al. [2018] analyze the role of random interactions in the structure of communities.

With respect to the characterization of nodes and edges, traditional network metrics have been employed to identify the most important nodes within a graph [Newman, 2004, 2010]. For instance, Newman [2004] uses centrality metrics based on shortest paths (e.g., closeness and betweenness [Easley and Kleinberg, 2010]) for determining the best positioned nodes in academic social networks. Considering a random walk-based model, Lü et al. [2011] propose a new approach, called LeaderRank, to identify the most influential nodes, which outperforms the well-known PageRank algorithm [Page et al., 1999]. Likewise, by emphasizing the social roles of the nodes, Xu et al. [2017] perform node classification by learning multiple social roles, whereas Liao et al. [2018] address the problem by learning their structural properties and the similarity of their attributes.

Characterization studies also consider the strength of relationships, for example, by exploring dynamic relationships in order to classify social ties as *strong* or *weak* [Brandão and Moro, 2017; Brandão et al., 2017; Leão et al., 2018; Vaz de Melo et al., 2015]. Additionally, one can characterize dynamic relationships by considering historical interactions [Brandão et al., 2017, 2018; Vaz de Melo et al., 2015]. For instance, Vaz de Melo et al. [2015] proposed the RECAST (*Random rElationship ClASsifier sTrategy*) algorithm to identify random and social interactions based on network properties. RECAST explores topological and temporal aspects to measure the strength of the nodes' relations, where such strength is derived from the neighborhood overlap and the persistence of interactions. As a result, it classifies the edges of a network by assigning them to one of the following social classes: *friend*, *acquaintant*, *bridge* and *random*. Then, Brandão et al. [2018] extended RECAST by adding co-authorship count as a feature related to the tie strength. They concluded that strong ties and bridges tend to persist over the years more than random ties. By exploring social aspects, Gilbert and Karahalios [2009] modeled tie strength as a linear combination of seven dimensions: *intensity* (e.g., words exchanged), *intimacy* (e.g., relationship status), *duration* (days), *reciprocal services* (e.g., links shared), *structural* (e.g., common groups), *emotional support* (e.g., positive words) and *social distance* (e.g., political differences). In another context, Srivastava et al. [2016] exploited narrative regularities with text-based features (linguistic, semantic and discourse) for characterizing relationships between people in movie summaries (e.g., denoting familial ties between characters).

2.4 Summary

In this chapter, we presented some approaches to analyze dynamic attributed networks and how to characterize node and edges from a social perspective. We also exemplify some representations of temporal networks. Finally, we discussed the importance of social concepts for bringing new viewpoints for a better understanding of complex networks.

In this thesis, we take all aforementioned topics one step forward. Specifically, we propose to put everything together: actors, interactions, time, attributes and social concepts. Our strategy consists in modeling nodes by associating them with their attributes in order to extract persistent features over time. Then, based on social concepts, we classify nodes and their relationships dynamically.

As we shall see next in the following chapters, our approach provides a new characterization perspective to better understanding the participation of actors and the dynamic interactions between them in terms of their positions in the social network structure.

Chapter 3

Modeling Dynamic Interactions

In this chapter, we discuss how to model actors and their relationships in dynamic social networks along the time, thus providing a novel understanding for characterizing complex systems. Specifically, we address our first research goal, which regards the graph model to mine multiple interactions over time. For this, we explore dynamic multigraph representations in order to extract knowledge from the associations between nodes and their attributes. In addition, we inspect the dynamics of the nodes to reflect the degree of cooperation between the actors in terms of the flow of attributes across the network.

3.1 Modeling Dynamic Interactions with Attributes

We can model a set of actors as nodes and their relationships as edges (a link between a pair of nodes), thus structuring their social interactions by means of a graph. In order to provide more information about this graph, such social interactions can be enriched with additional information associated with their nodes (e.g., age, gender, hometown, etc.) and edges (e.g., parent-child, adviser-advisee, follower-followee, etc.). Furthermore, this so-called attributed graphs can change over time in terms of their attributes (e.g., new skills acquired) or structurally (e.g., no more interactions with specific actors). Therefore, it needs to be modelled as a temporal graph that can change over time.

In this way, we model dynamic interactions as a graph G by considering a set of entities (nodes) and their relationships (edges), where the dynamics can be shown as a series of subgraphs by adding a time denotation. As discussed in Chapter 2, we examined three representations of such so-called temporal networks (*contact sequences*, *interval graphs* and *snapshots*) [Holme and Saramäki, 2012], which provide some features such as observing interactions at different discrete intervals or using one of them when the interaction is not negligible. As we are interested in identifying structural changes at each specific moment or even aggregating graphs from a given time interval, we opted to model the entire network for each observation of the social structure over time (i.e., *snapshots*).

Formally, given a set of actors and a sequence of interactions involving them at a discrete time k, we defined a temporal multigraph $\mathscr{G}_k = (\mathscr{V}_k, \mathscr{E}_k)$, where $\mathscr{V}_k$ and $\mathscr{E}_k$ represent, respectively, the set of nodes (actors) and the set of edges (relationships). In other words, this graph allows the existence of multiple edges between two nodes at the same time k, thus providing a view of the entire social structure at each moment. We can also join such subgraphs within a specific time interval, in such a way that $G = \bigcup_{i=0}^{t} \mathscr{G}_i$ represents the temporal aggregated graph that comprises the set of all nodes and their interactions within a time interval $[0,t]$.

Additionally, our scenario encompasses attributes that are associated values belonging to nodes (e.g., age, gender and hometown) or edges (e.g., parent-child, adviser-advisee and follower-followee). Given the sets $\mathscr{A}_{nodes}$ and $\mathscr{A}_{edges}$, which represent all possible attributes for nodes and edges, respectively, we associate a subset of attributes to each node and edge as follows:

- **Node attributes.** Let Γ: $n \in V \to A$ be a function that performs the mapping of each node n to a specific subset of attributes A contained in the set $\mathscr{A}_{nodes}$.
- **Edge attributes.** Let Φ: $e \in E \to A$ be a function that performs the mapping of each edge e to a specific subset of attributes A contained in the set $\mathscr{A}_{edges}$.

As we are dealing with time-varying graphs, we can denote the set of attributes for a node or an edge by adding a time notation t. For example, $\Gamma_k(u)$ returns the set of attributes associated with node u at time k, while $\bigcup_{i=6}^{8} \Phi_i(e)$ returns all attributes associated with the edge e during the time interval $[6,8]$.

In addition, we are interested in quantifying how strong the bond of the nodes is with their associated attributes, as well as what information is exchanged between them (also described as attributes). A trivial strategy is to weight the attributes as, for example, calculating their frequency. Instead, we can obtain more information if we model them as entities that participate in the interactions between individuals. In this way, likewise in relationships between nodes, we can also extract knowledge from social structures linked to attributes. Hence, we define the attribute graph as H to model the node-attribute dynamics. The idea is to consider dynamic temporal attributes, thus defining a heterogeneous graph formed by two types of node: actors (e.g., researchers) and attributes (e.g., expertise). Likewise the graph G, we can also describe the graph H as a multi-edge temporal graph and denote it as $H = (\mathscr{H}_0, ..., \mathscr{H}_t)$. We construct each subgraph $\mathscr{H}_k = (\mathscr{V}'_k, \mathscr{E}'_k)$ from the mapping function Φ, i.e., we create an edge to link each node $n \in \mathscr{V}_k$ for each one of its attributes a

in $\Phi_k(u)$. Formally, the sets of nodes and edges are, respectively, $\{V_k \cup \bigcup_{u \in V_k} \Phi(u)\}$ and $\{(u,a) | u \in V_k \wedge a \in \Phi(u)\}$. In other words, this strategy is an abstraction that transforms the attributes of each edge into additional nodes, allowing an original actor node to be directly connected to these new attribute nodes.

In order to illustrate this, we recall the compartmental models SIR and SIS [Hethcote, 2000], since they bring an interesting problem related to dynamic interactions. Then, we exemplify how to apply the aforementioned modeling to this network structure.

The SIR Epidemic Model. It models how a disease spreads in a population. For this, an individual can be in one of three stages during the course of the epidemic, namely:

- **Susceptible (S).** The individual can be infected by an infected one.
- **Infectious (I).** The individual is infected and has some probability of infecting susceptible individuals.
- **Removed (R).** The individual is recovered and no longer poses a threat of future infection.

The SIS Epidemic Model. Based on the SIR model, it establishes that an individual can be reinfected multiple times. In other words, an individual alternates between the stages susceptible (S) and infectious (I).

As the mechanisms of these two models are quite similar, we exemplify the dynamics of the SIS model as follows. This epidemic structure can be seen as a *contact network*, where we have an initial setup and its changes at each step of transmission. In this way, we have as input the set of all individuals and the set of relationships (contacts) between them, as well as the indication of all those individuals who were initially infected. We can then model this scenario by defining a temporal graph $G = (\mathscr{G}_0, ..., \mathscr{G}_t)$, where each $\mathscr{G}_k$ informs the step k of the epidemic spreads. In our perspective, the nodes in the *infectious state* are those have the attribute *infectious* assigned to them.

Figure 3.1 (a) shows three individuals, where initially (i.e., time $k = 0$) $\mathscr{V}_0 = \{A, B, C\}$ and $\mathscr{E}_0 = \{(A,B), (A,C), (B,C)\}$. Note that the individual B is the only one initially infected (shaded), i.e., $\Gamma(B) = \{infectious\}$. The SIS model establishes that an infected individual has a probability p of passing the disease to each of its susceptible neighbors. In this way, the node B can infect the nodes A and C, which happened, as observed at the next time instance ($k = 1$). Note that B recovered at $k = 1$ and became infected again at $k = 3$. The node C, which was infected at $k = 1$, recovered shortly and remained in the *susceptible* state.

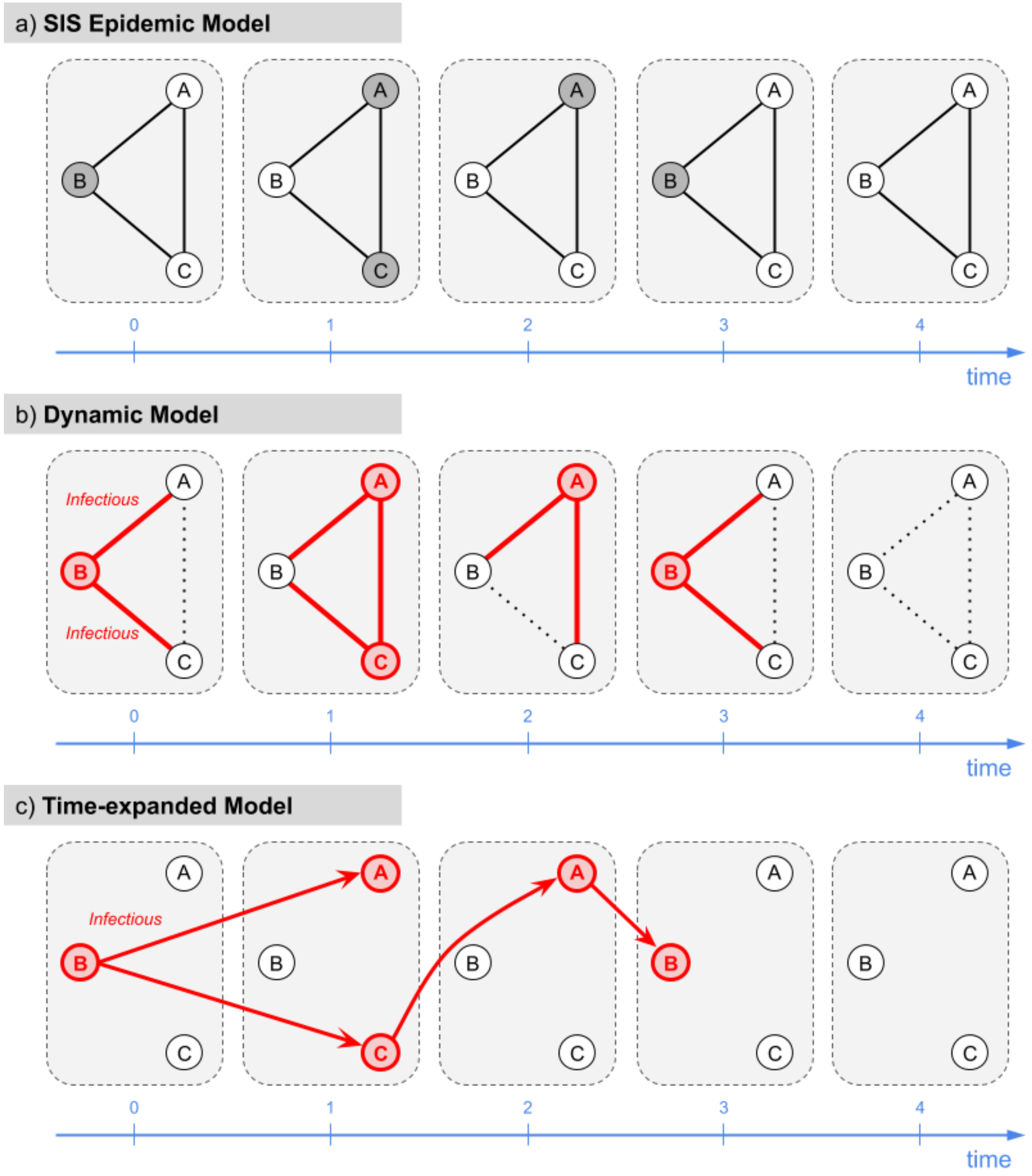


Figure 3.1: Different strategies to model an SIS epidemic. In (a), a contact network for each time step, where the nodes in the infected state are shaded. In (b), an attributed network as the nodes and edges in boldface depicting an association with the *infectious* attribute, respectively, those who are in the infected state and the potential transmission paths. In (c), a time-expanded network from (a) in order to designate the transmission sources.

In the last observation, the disease was controlled (i.e., there is no longer any *infectious* individuals).

In the previous example, we can clearly see how the epidemic spread, but in complex systems there are several ways for entities to interact with each other. For instance, specifically in epidemics, there are diseases that spread through droplets and others that are sexually transmitted. In this regard, two SIS models are reacquired to analyze such scenarios separately, since the types of social contact (by air and by sex) are not incorporated in these epidemiological model. However, we can map together the type of contact between individuals, thus informing by means of an edge attribute whether the interaction between them was, for example, *respiratory* or *sexual*. For example, we can describe that the contact between A and B at time $k = 5$ has associated attributes, such as the risk of spreading the disease, i.e., $\Phi_5((A,B)) = \{\textit{respiratory infection}, \textit{sexual infection}\}$.

In this regard, Figure 3.1 (b) shows the same aforementioned dynamics in Figure 3.1 (a), but now with the use of attributes at the edges. Note that this model brings more information that allows us to better visualize the potential paths (in bold) by means of which the disease can spread. Furthermore, from an informative point of view, the dotted edges illustrate that there is no relevant information being exchanged between them (e.g., B and C at $k = 2$).

Finally, in addition to incorporating information in the network structure, we can observe the origins of the spread of diseases (i.e., who transferred each attribute and when). We can model such propagation as a directed graph by observing the transition between the stages $k = t$ and $k = t + 1$. Figure 3.1 (c) depicts the flow of the disease along the time. We can see that B is the origin of the disease that was transmitted to A and C. Then, A infected B again, which recovered in the next period.

As mentioned earlier, what determines the contagion between two individuals in the SIR and SIS models are the associated probabilities in each contact. In these more basic versions, a node is equally contagious with a probability p. However, more elaborate extensions of the modeling process can be done by representing it in several states of infection (e.g., higher viral load in early or late periods of infection) or according to the predisposition of the individual (e.g., COVID-19 is more severe with those who have comorbidities).

From our perspective that observes the relationship of entities with attributes over time, we need to determine how to measure the strength of such associations (i.e., the probability of *acquiring* an attribute), as well as to understand how attributes flow (transmit) across the network. Next, we present how to deal with these two issues.

3.2 Extracting Relevant Attributes

The next step in our approach is to determine the set of relevant attributes for each node at each time interval. We define as relevant attributes those that are closely connected to the nodes, i.e., persistent in their histories. The idea is to identify, for each actor, all attributes and evaluate them according to their stability along the time (i.e., to identify the set of attributes most strongly statistically associated with the actor nodes).

For this, we analyze the nodes' interaction history in order to extract knowledge from the node-attribute relationships. We apply the concept of persistence of an edge along the time, which provides the notion of the importance of the relationship between two nodes in terms of their associated attributes. The *persistence* metric of an edge is defined as $pers_t(u,a) = \frac{1}{t}\sum_{k=1}^{t} \mathbb{1}_{\mathscr{E}_k^{'}}((u,a))$, where the indicator function is defined as

$$\mathbb{1}_{\mathscr{E}_k^{'}}((u,a)) = \begin{cases} 1, \text{if } (u,a) \in \mathscr{E}_k^{'}, \\ 0, \text{otherwise.} \end{cases} \tag{3.1}$$

Note that this operation is performed on each attributed graph at discrete intervals and not on the aggregated graph. In other words, it captures the dynamics by observing the persistence in each temporal subgraph within the time interval $[1,t]$.

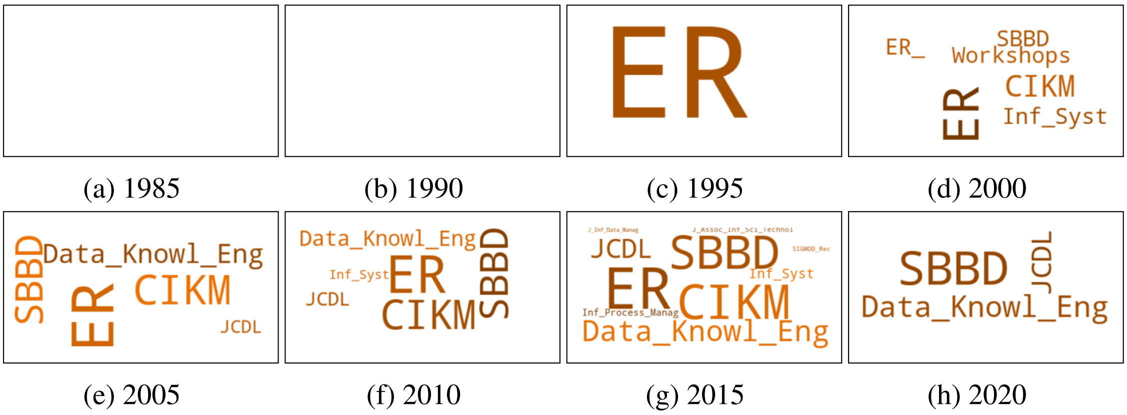
More precisely, Algorithm 1 details the process of extracting relevant attributes. It receives as input the aggregated graph $H = \{\mathscr{H}_1, ..., \mathscr{H}_t\}$ and the final time interval t. In summary, the algorithm inspects, for each actor, all attributes and evaluates them according to their persistence along the time by means of percentiles (function *percentile* on lines 9 and 11), thus identifying the set of attributes most strongly statistically associated with the actor's nodes. The idea is to filter such attributes that are exaggeratedly linked to a node in a specific period in comparison to the others, i.e., identifying the abnormal presence of certain attributes at each time point. In order to choose the appropriate statistical method to select the most significant attributes, we first check whether the values of the edge persistence metric follows a normal distribution. Then, we extract the relevant attributes based on the definition of an outlier given by the interquartile range (IQR). Another approach is to use the modified z-score for the same purpose [Iglewicz and Hoaglin, 1993]. Since the experimental results were similar for IQR and for the modified z-score, we chose IQR due to the possibility of applying different percentages by means of percentiles (i.e., adapting the constraints according to specifics problems).

As a result, this strategy builds a set comprising all attributes statistically relevant for each node $u \in G$ at a time interval k (i.e., for each subgraph), referenced as $\Gamma_k(u)$. Note that the sets $(\Gamma_1(u), \Gamma_2(u), ..., \Gamma_t(u))$ are dynamically built according to the degree of persistence, i.e., different instants k may contain completely distinct sets of attributes. In the worst case, when all edges exist at all intervals with all attributes, the time complexity of Algorithm 1 is $O(t|V|(|E| + |\mathscr{A}_{nodes}|))$. In practice, since this process is performed without taking into account the nodes' neighborhood (i.e., attributes are defined by the node itself), the relevant attributes for each node are computed in parallel.

In order to illustrate this strategy, we extracted statistically relevant attributes from the

Algorithm 1 Extracting Relevant Attributes

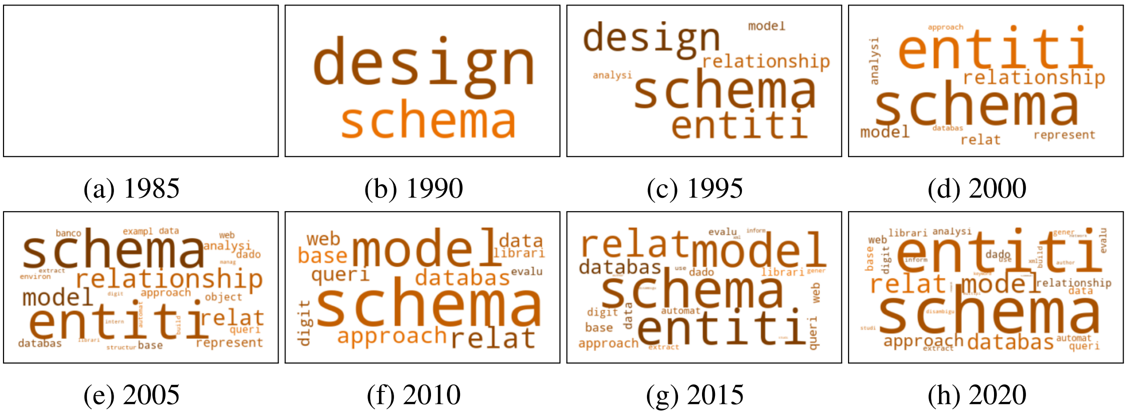
```
Require: $H, t$
Ensure: $\Gamma_k(u), \forall u \in \bigcup_{k=0}^{t} V_k$
for all $u \in V_t$ do
    $\mathscr{A}_{temp} \leftarrow \{\}$
    for all $k \in [1,t]$ do
        $\Gamma_k(u) \leftarrow \{\}$
        $\mathscr{A}_{temp} \leftarrow \mathscr{A}_{temp} \cup \{a | (u,a) \in \mathscr{E}_k'\}$
        $vector \leftarrow \{\}$
        for all $a \in \mathscr{A}_{temp}$ do
            $vector.add(pers_k(u,a))$
        $IQR \leftarrow percentile(vector, 75) - percentile(vector, 25)$
        for all $a \in \mathscr{A}_{temp}$ do
            if $pers_k(u,a) > percentile(vector, 75) + IQR * 1.5$ then
                $\Gamma_k(u) \leftarrow \Gamma_k(u) \cup \{a\}$
```



(a) 1985 (b) 1990 (c) 1995 (d) 2000

(e) 2005 (f) 2010 (g) 2015 (h) 2020

Figure 3.2: Attribute clouds that represent statistically relevant associations between *Alberto H. F. Laender* and its publication venues.



(a) 1985 (b) 1990 (c) 1995 (d) 2000

(e) 2005 (f) 2010 (g) 2015 (h) 2020

Figure 3.3: Attribute clouds that represent statistically relevant associations between *Alberto H. F. Laender* and terms extracted from the titles of his articles.

researcher *Alberto H. F. Laender*'s articles between the period 1984 to 2020[1]. Specifically, we analyzed his strong bond with publication venues (Figure 3.2) and with terms extracted from the titles of his publications (Figure 3.3). For example, Figure 3.2(a) informs that no publication venue is statistically relevant for him in 1985 (Γ_{1985}(*'Alberto H. F. Laender'*) $=$ $\{\}$), while in Figure 3.3(b) it indicates that the terms *design* and *schema* are relevant in 1990 (Γ_{1990}(*'Alberto H. F. Laender'*) $= \{$ *design, schema* $\}$).

Considering his publication venues (Figure 3.2), the first community to stand out as relevant is ER in 1995, i.e., at the beginning of his career the trend was not to focus on specific targets. Along the time, there are relevant attributes in and out, but there is clearly a central set that remains. Indeed, the dynamics of his research targets is moderated, emphasizing a well-defined and perennial research pattern with the ER, CIKM, Data & Knowledge Engineering, JCDL and SBBD communities. Regarding terms extracted from his publications (Figure 3.3), initially we observe a pattern of very specific research terms (*design, schema* and *entiti*) that expands from 2000. In 2005, there are the presence of new terms (e.g., *web*, *extract*, *structur*, *digit* and *librari*), while in other periods the pattern tends to be maintained. Analyzing the two attribute clouds, we can see a conservative pattern of the researcher, where the subjects covered and the targets maintain a well-defined core over time.

3.3 Modeling Knowledge-Transfer

Finally, considering that attribute nodes carry information that can spread throughout the graph, we model their dynamics by defining a directed graph D in which each edge reflects the degree of cooperation between the actor nodes in terms of their attributes. We can define such a graph as $D = (V_d, E_d)$, where $V_d \subseteq V$ is a set that contains only those actor nodes that have transferred knowledge along the time and E_d represents the set of such directed edges. Note that we can create such a graph from the aggregated graph $G = (V, E)$ or observing its dynamics over time through each instance $\mathscr{G}_k = (\mathscr{V}_k, \mathscr{E}_k)$. For the first case, as a result, it models only the final state of the interaction among the nodes (single-edge) and, therefore, it does not regard multiple interactions over time.

Regarding the distinct dynamics behind the knowledge transfer involving the nodes, we can redefine this graph as a directed multigraph $D = (V_d, (\mathscr{F}_1,\ \mathscr{F}_{1+k}, ..., \mathscr{F}_t))$ to map knowledge transfer dynamically over time, where $\mathscr{F}_k$ is similar to $\mathscr{E}_k$ in comprising multi-edges in time k. More specifically, there will be an edge in $\mathscr{F}_k$ originating from a node u to a node v whenever there is a knowledge transfer from u to v at k.

[1] Data extracted from his DBLP page in November 2020.

Algorithm 2 Directed Knowledge-Transfer Graph

```
Require: $G$, $t$ and $\Gamma$
Ensure: $D = (V_d, (\mathscr{F}_1, \mathscr{F}_2, ..., \mathscr{F}_t))$
 1: $V_d = \{\}$
 2: for all $k \in [1,t]$ do
 3:     $\mathscr{F}_k = \{\}$
 4:     for all $(u,v) \in \mathscr{E}_k$ do
 5:         if $\exists a \in \Gamma_k(u) | a \notin \Gamma_k(v)$ then
 6:             $\mathscr{F}_k = \mathscr{F}_k \cup \{(u,v)\}$
 7:             $\mathscr{V}_d = \mathscr{V}_d \cup \{u,v\}$
 8:         if $\exists a \in \Gamma_k(v) | a \notin \Gamma_k(u)$ then
 9:             $\mathscr{F}_k = \mathscr{F}_k \cup \{(v,u)\}$
10:             $\mathscr{V}_d = \mathscr{V}_d \cup \{u,v\}$
```

Algorithm 2 describes how to build the knowledge-transfer graph D by inspecting all edges from the graph G. For each edge (u,v), it verifies the existence of some knowledge transferred between u and v in terms of their attributes (lines 5-10). If at a time instant k a node u has at least one relevant attribute that does not belong to node v, then a directed edge from node u to node v is created in $\mathscr{F}_k$. Note that not all nodes and multiedges in G will be in D, but only those that represent some knowledge transfer.

3.4 Summary

In this chapter, we introduced the graph model proposed to deal with dynamic interactions over time, thus addressing our first research goal. Specifically, this chapter presented how to model social interactions over time based on their associated attributes. For this, we defined three graph representations to deal with dynamic interactions, their associated attributes and how the knowledge-transfer flows across the network. We also exemplify how to apply our strategy to an epidemiological problem and illustrate the statistically relevant attributes of a researcher.

As we shall see next in Chapter 4, our general-purpose model provides a framework for exploring the social role of nodes, extracting the meaning of their interactions and understanding their dynamics in terms of attributes associated with them over time.

Chapter 4

Social-based Classification

In this chapter, we emphasize the social behavior of nodes and the meaning of their relationships. Here, we address our second research goal, which regards the social-based classification of the dynamic interactions. Specifically, our classification scheme reinforces the importance of social concepts as a relevant factor for better understanding the complexity that involves actors and their relationships in dynamic attributed networks. In summary, we classify nodes and edges based on the following two social perspectives:

- **Strength of the Social Structure.** This classification captures the nodes' social behavior, as well as it assigns a social meaning to edges in order to measure the strength of the interactions. For example, in an academic social network, a node that has a long-lasting association with attributes like *relational model*, *data definition* and *query language* is likely to have an authority over them. Then, this node can be seen as a *hub*, since it has the potential to spread knowledge from the *databases* domain. In this context, a *strong* edge may indicate a social tie between an advisor and an advisee, whereas a *weak* one may represent a multidisciplinary social tie with researchers from other scientific disciplines.

- **Knowledge Transfer Dynamics.** This classification captures the ability of the nodes to transfer and spread knowledge across a network, as well as to characterize the transfer potentials involved in their relationships. For example, collaborations between researchers from different fields of study may result in the sharing of new information across the edges. Likewise, a node that acts as both a receiver and as a transmitter play a more crucial role when spreading knowledge on a social structure.

Next, we present the algorithms proposed to classify nodes and edges based on social concepts.

4.1 The Strength of the Social Structure

Based on the social structure that models the dynamic interactions along the time, we might measure the structural strength of social ties by means of relevance degree of the attributes associated with each node by considering its past interactions. For this, our strategy determines the dynamic state of a node at each time interval as representing a *strong*, *weak* or *non-relevant* association with a specific attribute (i.e., knowledge). In this context, a *strong* state represents the importance of a node in terms of its expertise within a closed group, whereas a *weak* one captures its potential for connecting different parts of a network.

The next step consists in mapping these dynamic states in order to determine the social classes of the edges. In a preliminary version of our work [Silva et al., 2018], we defined a comprehensive classification scheme by quantifying levels of specific social concepts, namely: *very strong*, *strong*, *strong bridge*, *regular bridge*, *weak bridge*, *ordinary* and *sporadic*. However, our experiments showed that the properties of such classes were not very discriminatory. In particular, the edges classified as weak and regular bridges were not statistically different in terms of the betweenness centrality metric.

Thus, in order to provide more representative results, here we define the edge classes following Burt's social theory that considers *closure* and *brokerage* as forms of social capital [Burt, 2005]. Thus, we define three social classes for edges: *closure*, *brokerage* and *innocuous*. Such classes capture the benefits derived from individuals that occupy specific places in a social structure. More specifically, they emphasize the strength of relationships as strong ties (*closure*), weak ties (*brokerage*) and non-relevant information passing through the edge (*innocuous*).

Formally, Algorithm 3 describes our process for classifying multiple edges. Note that the nodes' dynamic states are assigned independently at each iteration of the algorithm and considering each instant k in which an edge is inspected. A node is assigned a state *strong* when there is a strong temporal link with its attributes at the exact moment of the interaction (lines 4 and 5, and 9 and 10). However, if these attributes do not apply to the inspected edge, then the state *weak* is assigned to it (lines 6 and 11). If there are no relevant attributes and the node is active in more than one time interval, then the state *non-relevant* is assigned to it (lines 7 and 12). Once the dynamic states have been assigned to nodes u and v, the class of the corresponding edge e is assigned according to them (lines 13-17). More specifically, the *brokerage* class can be seen as a social tie of nodes from distinct domains (lines 13 and 14), whereas the *closure* one establishes a social role by demonstrating a high tightness between a node and its attributes (line 15 and 16). Finally, an *innocuous* class means that there is no knowledge being disseminated through the inspected relationship (line 17). In the worst case, when all edges exist at all time intervals, the time complexity of Algorithm 3 is $O(t|E|)$.

Algorithm 3 Classifying Edges

```
Require: $G, t, \Phi$ e $\Gamma$
Ensure: $\Delta((u,v)), \forall (u,v) \in \bigcup_{k=1}^{t} \mathscr{E}_k$
for all $k \in [1,t]$ do
    for all $(u,v) \in \mathscr{E}_k$ do
        if $|\Gamma_k(u)| \neq 0$ then
            if $|\Gamma_k(u) \cap \Phi((u,v))| \neq 0$
                then $u_{state} \leftarrow strong$
            else $u_{state} \leftarrow weak$
        else $u_{state} \leftarrow non\text{-}relevant$
        if $|\Gamma_k(v)| \neq 0$ then
            if $|\Gamma_k(v) \cap \Phi((u,v))| \neq 0$ then
                $v_{state} \leftarrow strong$
            else $v_{state} \leftarrow weak$
        else $v_{state} \leftarrow non\text{-}relevant$
        if $u_{state} = strong$ or $v_{state} = strong$ then
            $\Delta((u,v)) \leftarrow closure$
        else if $u_{state} = weak$ or $v_{state} = weak$ then
            $\Delta((u,v)) \leftarrow brokerage$
        else $\Delta((u,v)) \leftarrow innocuous$
```

For classifying the nodes, the same classes are assigned to them, in which case we mean by *closure* a node that has authority on certain attributes, by *brokerage* a node that has a weak association with its attributes and by *innocuous* a node that has an occasional presence in the network. The function Ω for this node classification is given by

$$\Omega(u) = \begin{cases} closure, & \text{if } |\Gamma_t(u)| \neq 0 \\ brokerage, & \text{else if } \sum_{k=1}^{t} \mathbb{1}_{\mathscr{V}_k}(u) > 1 \\ innocuous, & \text{otherwise,} \end{cases} \tag{4.1}$$

where the indicator function is defined as

$$\mathbb{1}_{\mathscr{V}_k}(u) = \begin{cases} 1, \text{if } u \in \mathscr{V}_k, \\ 0, \text{otherwise.} \end{cases} \tag{4.2}$$

This function can be implemented by simply adding flags to Algorithm 3.

In summary, the aforementioned social classes reinforce a sociological perspective based on their positioning in a social structure [Burt, 2005; Granovetter, 1973; Guimera et al., 2005], i.e., by applying social concepts to better understand the strength of the *node-attribute* associations. More precisely, based on Burt's definition of social capital [Burt, 2005], a *closure* edge means a high tightness between nodes by means of their relevant attributes, whereas a *brokerage* edge can be seen as a social tie of nodes from distinct relevant attributes. Likewise, when classifying a node, the *closure* class is assigned to it when there is a strong tie with some knowledge under its set of relevant attributes and the *brokerage* class when it represents a potential to acquire knowledge from attributes outside its own set of relevant attributes. Indeed, strong ties with certain attributes show an authority on them, whereas weak ties express a great potential to diffuse knowledge from its domain. Finally, the *innocuous* class assigned to a node or edge represents no skill acquired by an individual or non-relevant information passing through a relationship, respectively.

4.2 Knowledge Transfer Dynamics

Now, we discuss knowledge transfer dynamics in a scenario composed by actors and a set of historical interactions among them. Each actor has a set of attributes associated with them, which can change over time. We also assume that all attributes are related to a skill (or knowledge), which can be transferred from (or taught by) an actor, who has this skill, to another one, who does not have it.

Thus, let us consider the directed multigraph D (Section 3.3) that models how knowledge is dynamically transferred between two actors over time. Then, considering D, we characterize the dynamics behind the knowledge transfer among nodes by using four classes of relationship:

(i) *Closure*. A closure relationship occurs when two nodes are teaching to and learning from each other. In other words, there is a closure relationship between A and B if there is at least one relevant attribute in A that is not in B and also at least one relevant attribute in B that is not relevant to A. Thus, this relationship represents new knowledge being disseminated by both sides.

(ii) *Brokerage.* A brokerage relationship characterizes an one directional knowledge transfer of an attribute between an expert on that attribute and an inexpert on it. That is, there is a brokerage relationship between A and B if there is at least one relevant attribute in A that is not in B, and there is no relevant attribute in B that is new to A. Thus, in this case the knowledge transfer occurs in only one direction.

(iii) *Dependent.* A dependent relationship is similar to the *brokerage* one, except that the destination node is not yet an expert in any attribute. Thus, this relationship class establishes a total dependence relation, i.e., a situation where the destination node has only in-edges.

Algorithm 4 Edge Classification

```
Require: $G = (V, (\mathscr{E}_1, \mathscr{E}_2, ..., \mathscr{E}_t))$,
         $D = (V_d, (\mathscr{F}_1, \mathscr{F}_2, ..., \mathscr{F}_t))$ and $t$
Ensure: $\Delta_{KT}((u,v)), \forall (u,v) \in \bigcup_{k=1}^{t} \mathscr{E}_k$
 1: for all $(u,v) \in \bigcup_{k=1}^{t} \mathscr{E}_k$ do
 2:     if $(u,v) \notin \mathscr{F}_k$ and $(v,u) \notin \mathscr{F}_k$ then
 3:         $\Delta_{KT}((u,v)) \leftarrow innocuous$
 4:     else if $(u,v) \in \mathscr{F}_i$ and $(v,u) \in \mathscr{F}_k$ then
 5:         $\Delta_{KT}((u,v)) \leftarrow closure$
 6:     else
 7:         if $(u,v) \in \mathscr{F}_k$ then
 8:             if $\forall w \in V | (v,w) \notin \mathscr{F}_k$ then
 9:                 $\Delta_{KT}((u,v)) \leftarrow dependent$
10:             else
11:                 $\Delta_{KT}((u,v)) \leftarrow brokerage$
12:         else
13:             if $\forall w \in V | (u,w) \notin \mathscr{F}_k$ then
14:                 $\Delta_{KT}((u,v)) \leftarrow dependent$
15:             else
16:                 $\Delta_{KT}((u,v)) \leftarrow brokerage$
```

(iv) *Innocuous.* An innocuous relationship characterizes pairs of actor nodes who do not share any relevant attributes, i.e., there is no knowledge passing through that edge.

Algorithm 4 uses D to classify the set of dynamic edges in G according to the four classes of knowledge transfer. Following the previous definitions, in lines 2 and 3, an edge (u,v) in G is considered *innocuous* when there is no directed edges (u,v) and (v,u) in the knowledge-transfer graph D, whereas it represents a *closure* relationship when there is a bidirectional edge between them. Otherwise, there is some knowledge transfer from one node to the other. In this case, if the destination node is totally dependent (lines 8-9 and 13-14), then the assigned relationship class is *dependent*, i.e., the destination node has no out-edges. On the other hand, if the destination node has out-edges, then the assigned relationship class is *brokerage* (lines 11 and 16), i.e., it carries some potential to transfer knowledge to its neighbor nodes.

Finally, we expand the aforementioned social classification to also consider four classes of nodes regarding their knowledge transfer capabilities:

(i) *Closure.* A closure node behaves as a knowledge transmitter and receiver in the network. That is, it is capable of teaching and learning.

(ii) *Brokerage.* A brokerage node works as a knowledge transmitter. In other words, this kind of node only disseminates knowledge from its own domain and thus does not spread content learned from others.

(iii) *Dependent.* Unlike a brokerage node, a dependent one is characterized by only receiving knowledge from the network.

(iv) *Innocuous.* An innocuous node is one that transfers or receives no knowledge.

Formally, Algorithm 5 defines the classes of nodes in G by inspecting their behaviors as transmitters (sources) and receivers (targets) of relevant knowledge in the directed temporal graph D. Based on the previous considerations, lines 1 to 5 define the sets of transmitter and receiver nodes in terms of, respectively, out-edges and in-edges in D. The *closure* class is assigned to a node that behaves as both transmitter and receiver (lines 7 to 9), thus establishing a strong relationship with other nodes. In case of acting only as a transmitter (line 11), the *brokerage* class is assigned to the node, thus indicating its behavior as a knowledge disseminator. However, if that node acts only as a receiver (line 12), then the *dependent* class is assigned to it in order to indicate the behavior of accumulating knowledge from others. Finally, the *innocuous* class is assigned to a node when no transmitter or receiver behavior is identified (line 15).

Algorithm 5 Node Classification

```
Require: $G = (V, (\mathscr{E}_1, \mathscr{E}_2, ..., \mathscr{E}_t))$,
         $D = (V_d, (\mathscr{F}_1, \mathscr{F}_2, ..., \mathscr{F}_t))$ and $t$
Ensure: $\Delta_{KT}(u), \forall u \in V$
 1: Transmitters $\leftarrow \emptyset$
 2: Receivers $\leftarrow \emptyset$
 3: for all $(u', v') \in \bigcup_{k=1}^{t} \mathscr{F}_k$ do
 4:     Transmitters $\leftarrow$ Transmitters $\cup \{u'\}$
 5:     Receivers $\leftarrow$ Receivers $\cup \{v'\}$
 6: for all $u \in V$ do
 7:     if $u \in Transmitters$ then
 8:         if $u \in Receivers$ then
 9:             $\Delta_{KT}(u) \leftarrow closure$
10:         else
11:             $\Delta_{KT}(u) \leftarrow brokerage$
12:     else if $u \in Receivers$ then
13:         $\Delta_{KT}(u) \leftarrow dependent$
14:     else
15:         $\Delta_{KT}(u) \leftarrow innocuous$
```

In summary, the proposed knowledge-transfer classification allows us to better understand the diffusion of relevant information and the existence of social ties without an apparent flow or even a dependent behavior. With such an information, we can not only assess more accurately the social role of the nodes, but also scale the degree of influence based on the disseminated knowledge.

4.3 Discussing Examples

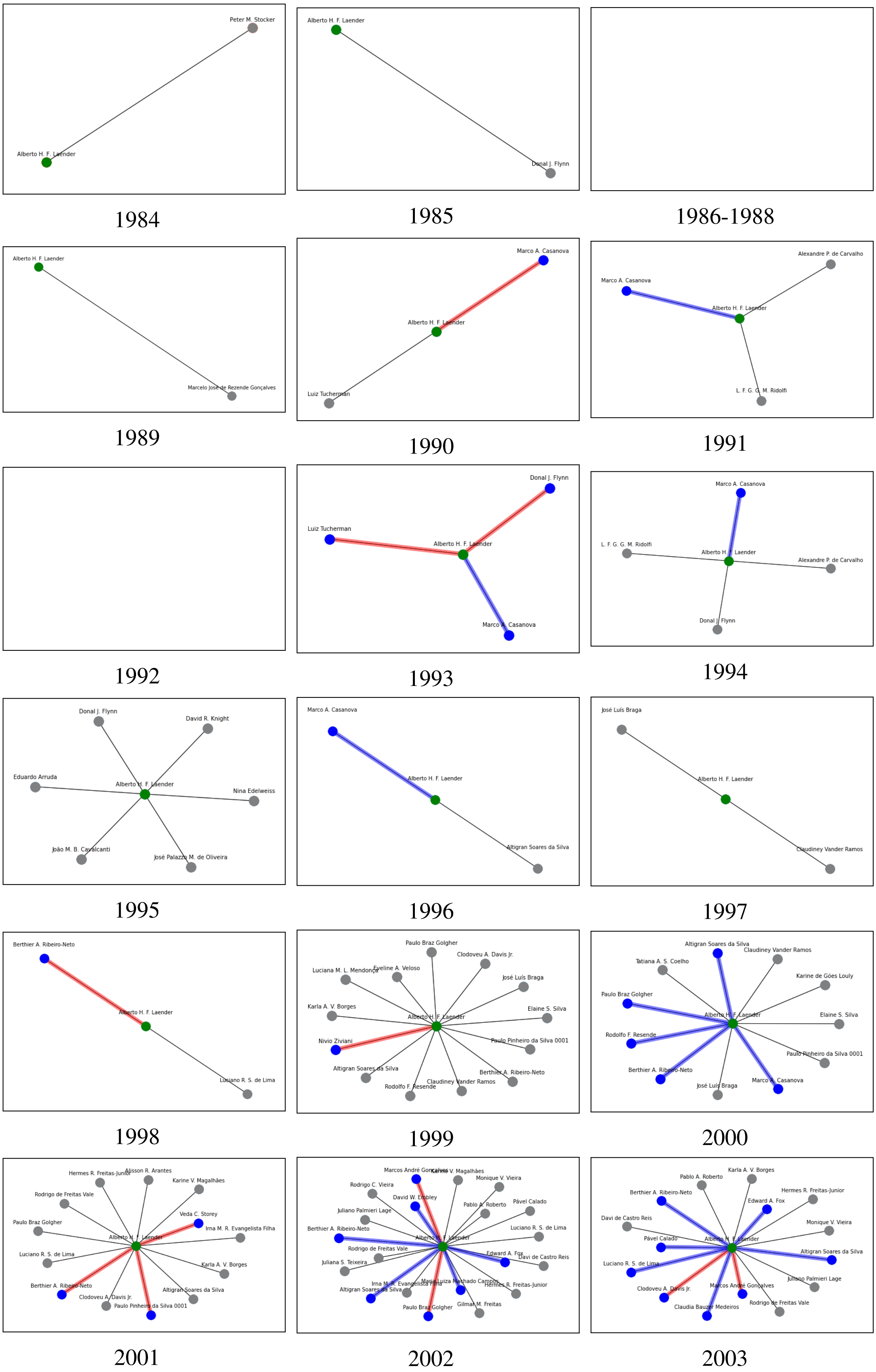
To exemplify our method, we will discuss two examples. First, a broader scenario visualizing the relationship types over time. Then, a more specific scenario considering a closed group of researchers related to a specific scientific conference.

Example 1. Let us look at the strength of the relationships of the researcher Alberto H. F. Laender[1] with his collaborators regarding their social bonds with academic communities, here represented by his publication venues. Figures 4.1 and 4.2 depict the classifications of each instance for the time interval $[1984, 2020]$, where each plot represents the ties of this researcher with his collaborators. At each graph instance, blue edges represent closure relationships and red edges brokerage ones, whereas gray edges correspond to innocuous edges. Note that we omit the parallel edges and the relationships among collaborators in such graphs for better visualization. For example, Figure 4.1(a) consists of the graph $\mathscr{G}_{1984} = (\mathscr{V}_{1984}, \mathscr{E}_{1984})$, where there is only one edge (*Alberto H. F. Laender*, *Peter M. Stocker*) classified as *innocuous* (in gray), thus informing that there is no strong link between those two nodes regarding that community. On the other hand, in 1990 there is a bridge established from Marco A. Casanova to at least one community, i.e., from that year onwards, other parts of the network could be accessed from this strong relationship.

We clearly notice that the classification is dynamic and can have edges with different social concepts over time. For example, relationships with Berthier A. Ribeiro-Neto are either of type closure (e.g., 2000 and 2003) or brokerage (e.g., 1998 and 2001). Even so, the strong bond from common communities tends to generate more cohesion (closure) between the collaborators, thus reinforcing the formation of a group in specific research areas. Nevertheless, there are cases in which a collaboration may become weaker, as in the example of the collaboration with Thiago H. P. Silva, which in 2013 and 2020 are seen as *innocuous*, although it was considered strong in 2015. In fact, in the period of $[2014, 2015]$, Silva was a master's student with several collaborations in areas of mutual interest (e.g., digital libraries).

Likewise, we can analyze particular cases such as that of Rodrygo L. T. Santos, in which his relationship is *innocuous* in 2006, *closure* in 2008 and *brokerage* in 2010. This

[1] Data extracted from his DBLP page in November 2020.



1984

1985

1986-1988

1989

1990

1991

1992

1993

1994

1995

1996

1997

1998

1999

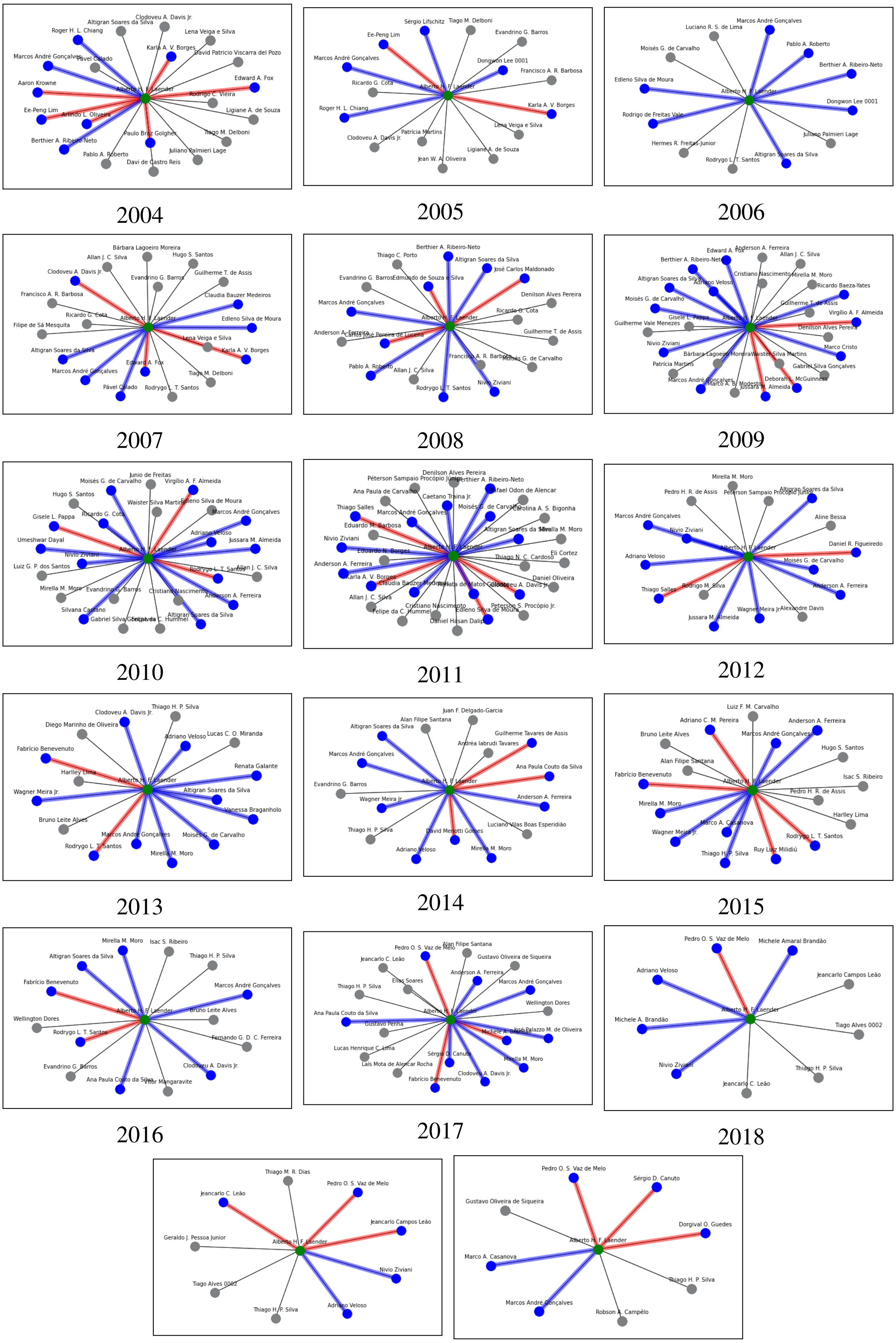
2000

2001

2002

2003

Figure 4.1: Social-based classification of Alberto H. F. Laender's edges. Part 1: [1984,2003] (the graphs can be best viewed by zooming their labels).



2004 2005 2006

2007 2008 2009

2010 2011 2012

2013 2014 2015

2016 2017 2018

2019 2020

Figure 4.2: Social-based classification of Alberto H. F. Laender's. Part 2: [2004,2020] (the graphs can be best viewed by zooming their labels).

.

example portrays that the then master's student had a strong bond for being a member from the same laboratory as the researcher Alberto H. F. Laender, but he became a bridge by diversifying his studies to other subjects during his doctorate. Also, we can look at the past history involving other researchers such as Edward A. Fox, with whom he started a strong link in 2002 within the CIKM community, which became a bridge to the digital library community in 2003 and 2004. In fact, the main conference in the area (JCDL) was considered relevant to Alberto H. F. Laender in 2005, as exemplified in his cloud of relevant venues (see Figure 3.2).

Example 2. Let be the coauthorship network of the researchers *Edward A. Fox, Marcos A. Gonçalves, Alberto H. F. Laender* and *Berthier A. Ribeiro-Neto*, where the edges represent a joint cooperation in a same paper. We would like to analyze their social behavior in the *ACM/IEEE Joint Conference on Digital Library* (JCDL) community during the period from 2002 to 2004. In other words, we intend to discover those researchers that have a very strong temporal relationship within this conference (*closure*), as well as those that act as bridges connecting other parts of the graph (*brokerage*). Also, we are interested in determining the kind of social interactions between them.

To analyze such social interactions in the period $[2002, 2005]$ for the JCDL community[2], we inspected the history of these specific researchers in the entire network from 1980 to 2001. As a result, a possible answer to this example is depicted in Figure 4.3. In summary, this specific scenario shows the participation of a young researcher (*Marcos A. Gonçalves*) and three experienced ones in the initial formation of the JCDL community (started in 2001). We can observe both *Marcos A. Gonçalves*' maturity and how he structurally acts by establishing a bridge between the other researchers, thus reinforcing the social nature of the *triadic closure* (high probability that mutual friends become friends). Also, we can identify which ones have brought contributions from outside of this community.

More specifically, in 2002, *Marcos A. Gonçalves* is classified as an *innocuous* node (the white one) because there is no strong link between him and any publication venue from his past interaction on that network $[1998, 2002)$. The other nodes are classified as bridges to other parts of the graph (the red ones) since they represent researchers that have a regular presence in specific communities along the time. Notably, *Edward A. Fox, Alberto H. F. Laender* and *Berthier A. Ribeiro-Neto* are considered bridges, since the bring knowledge from the JASIST, CIKM and SIGIR communities, respectively. Note that the only two edges (dotted lines) inform that there is no relevant information passing through them and, therefore, are classified as (*innocuous*), since none of the respective researchers regarded the JCDL community as one of their main target in 2002. Then, in 2003, *Marcos A. Gonçalves*

[2] Data extracted from their DBLP pages in November 2020.

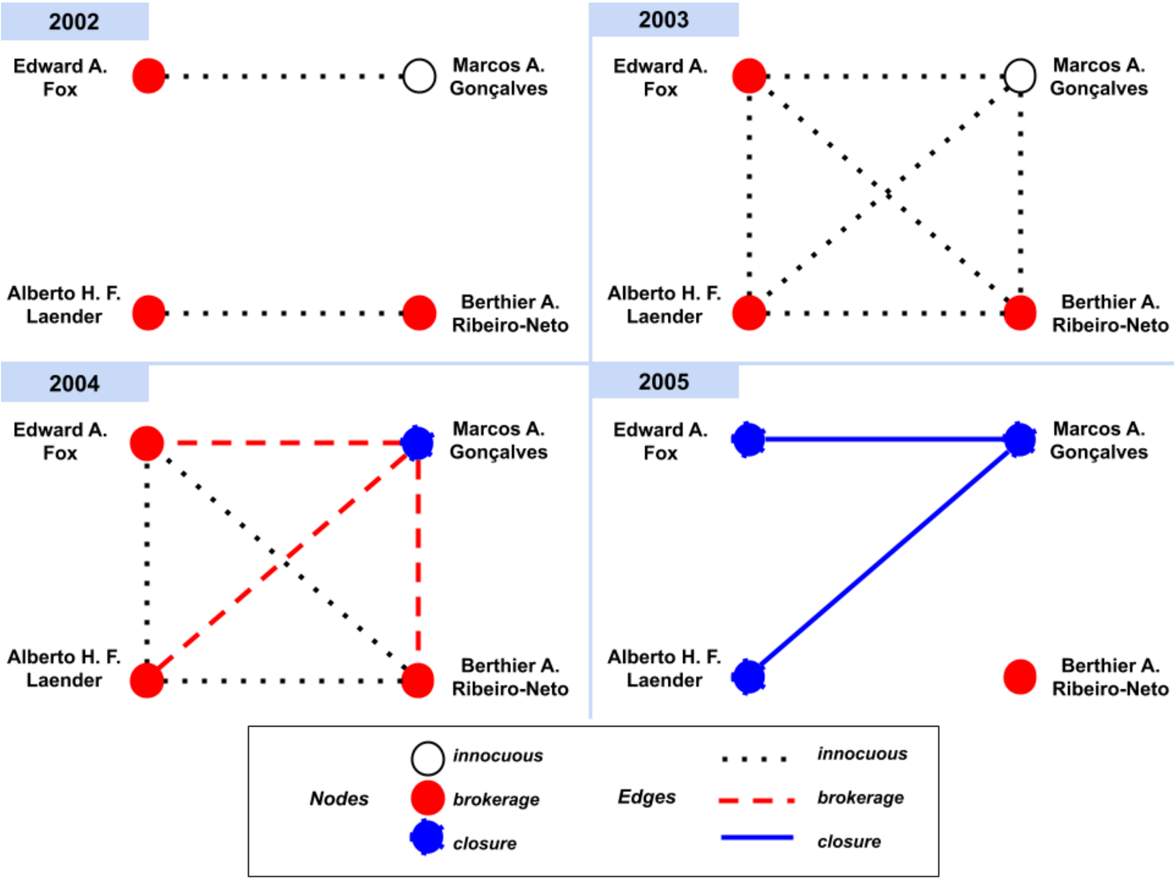


Figure 4.3: Social classification of nodes and edges of a coauthorship network involving four researchers in the JCDL community at four different moments.

introduced *Edward A. Fox* to Alberto H. F. Laender and Berthier A. Ribeiro Neto, thus creating a closed network; however, none of the edges are regarded as important, since they still do not have a strong bond within this community.

Surprisingly, in 2004 *Marcos A. Gonçalves*, who until then was classified as an *innocuous* node, is now seen as a *closure* one. We can understand this strong bond with JCDL because he obtained his PhD at that same year in the area of *Digital Libraries*, whose main conference is JCDL. In fact, his presence in that community is persistent, while the other researchers diversified their contributions to other ones. In this way, there are two types of edge, *innocuous* ones between those who do not have a strong relationship within that community, and *brokerage* ones like those established through *Marcos A. Gonçalves*. Finally, in 2005, *Edward A. Fox, Marcos A. Gonçalves* and *Alberto H. F. Laender* demonstrate a strong bond with the JCDL community and also a strong interaction between themselves, while *Berthier A. Ribeiro-Neto* does not play any specific role in that community. This excerpt extracted from the researchers' complete graph provides insights about social structures as facilitators for integrating individuals according to their social motivations (e.g., preferences for discussing subjects related to digital libraries), resulting not only in a highly connected network, but also in a more informative one. Accordingly, Burt [2005] argues that the maximum advantages occur with high *closure* (e.g., trust) and *brokerage* (e.g., cooperation) values, whereas minimum ones occur when these values are lower (e.g., distrust and indifference, respectively). That is, a good social strategy is to position where people are tightly connected to one another with extensive bridge ties beyond them.

4.4 Summary

In this chapter, we first reinforced the importance of the network theory paradigm for understanding the complexity that involves real world actors and their relationships [Barabási, 2009]. Based on the *structural autonomy* that informs when people are tightly connected to one another with extensive bridge ties beyond them [Burt, 2005], we emphasize the concept of *closure* as representing the importance of a node in terms of its expertise according to their associated attributes (strong ties), whereas the the concept of *brokerage* captures the potential of a node for transferring its attributes (weak ties). Then, we presented a characterization method in order to mine multiple interactions in dynamic attributed networks based on such social concepts. Overall, our proposed social-based classification reveals the social role of the nodes and the strength of the social meaning of their multiple interactions. Finally, we presented how to capture the knowledge transfer between the nodes by considering the dynamics involving individuals in terms of how their attributes are transferred across the network.

As we shall see, the proposed classification method reflects the social characteristics in different social scenarios. Moreover, the classes assigned to nodes and edges are associated with their strategic positioning in a social structure given by network properties. Next, Chapter 5 introduces the datasets considered and discusses the experimental methodology adopted for evaluating our classification methods.

Chapter 5

Experimental Methodology

In this chapter, we present the experimental methodology we adopt in order to assess in several social contexts the classification model that we have proposed. Indeed, classifying social interactions in a social network is a challenging task due to the difficulty to analyze real networks that clearly define the social role of its nodes and the meaning of its edges. To overcome this situation and evaluate the robustness of new methods for classifying nodes and edges in social networks, an experimental methodology should characterize the impact of a proposed approach in different scenarios [Alves et al., 2013; Silva et al., 2012; Vaz de Melo et al., 2015].

For this, we begin by introducing the datasets considered, which provide different social scenarios for a more robust experimental evaluation (Section 5.1). Then, we discuss the methodology adopted for evaluating the social characterization of nodes and edges, as provided by our proposed classification method (Section 5.2).

5.1 Datasets

To assess our proposed classification method, we consider two different social contexts derived from coauthorship networks and Questions and Answers (Q&A) communities, which are described next.

5.1.1 Social Academic Networks

We chose to analyze academic social networks that are, undoubtedly, well known in the Computer Science community, thus enabling us a more accurate discussion of their behavioral dynamics. Specifically, we considered several networks derived from data collected from DBLP[1] in June 2018. DBLP is a high-quality repository of Computer Science publications and has been widely used in several mining studies [Freire and Figueiredo, 2011; Leão et al., 2018; Li et al., 2018; Moreira et al., 2015; Rezaei et al., 2017; Silva et al., 2012; Yang and Leskovec, 2015; Wang et al., 2017]. In order to investigate the effects of our classification method in different social scenarios, we built the following networks constructed from our DBLP dataset:

- **ACM SIGs:** 24 coauthorship networks derived from all papers published in the proceedings of the flagship conferences of 24 ACM Special Interest Groups[2];
- **Full Network:** An integrated coauthorship network comprising all 24 ACM SIGs;
- **DBLP$_J$:** A coauthorship network derived from all journal articles indexed by DBLP;
- **DBLP$_C$:** A coauthorship network derived from all conference papers indexed by DBLP;
- **DBLP:** An integrated coauthorship network comprising the union of DBLP$_J$ and DBLP$_C$.

Table 5.1 presents some statistics of the aforementioned networks. Overall, the networks present distinct characteristics, which allow us to contrast the effect of our classification method on different social scenarios. For instance, the networks derived from the ACM SIGs are important because they represent well-known Computer Science communities and have already been addressed in other works [Alves et al., 2013; Benevenuto et al., 2015; Silva et al., 2015b], thus enabling us a more accurate discussion. We also have considered journal and conference publications separately (DBLP$_J$ and DBLP$_C$), since their communities have distinct behavioral dynamics, such as publishing more papers and collaborating with more coauthors in conferences than in journals [Kim, 2019; Laender et al., 2008]. Moreover Kim [2019] concluded that coauthors and paper titles of authors across conferences and journals tend not to overlap much, therefore we aim to verify the effect of our classification method also in these two specific scenarios. For building all aforementioned social academic networks, we considered only publication venues at least 10 years old and, as we are interested in analyzing interactions between individuals, publications with more than one author.

Given the above networks, our next step was to define the relationship attributes expressed by the academic coauthorships. Considering that scientific publication titles carry some specific meaning [Silva et al., 2012], we used tokens taken from them as such attributes. For this, we removed meaningless words (stop words) and reduced inflected words

[1] DBLP: `https://dblp.uni-trier.de/`

[2] Association for Computing Machinery: `http://www.acm.org/sigs`

Table 5.1: Statistics of the academic coauthorship networks.

Networks	**Period**	**#nodes**	**#edges**	**#multiple edges**
SAC	1992-2017	10,804	18,066	19,712
DAC	1964-2017	10,272	27,800	31,972
CHI	1989-2017	8,959	27,587	32,154
CIKM	1993-2017	7,342	16,347	18,822
MMSys	1992-2017	7,124	18,728	22,783
SIGCSE	1986-2018	6,247	15,252	18,232
KDD	1995-2017	4,998	13,614	15,150
SIGIR	1971-2017	4,905	11,247	13,595
SIGMOD	1975-2017	4,869	16,042	18,090
CCS	1993-2016	2,854	6,851	7,612
SIGCOMM	1981-2017	2,844	8,653	9,715
ICSE	1976-2017	2,829	4,977	5,354
SIGUCCS	1975-2017	2,517	2,349	2,734
STOC	1969-2017	2,500	5,568	6,608
SIGMETRICS	1976-2016	2,440	4,500	4,906
SIGGRAPH[a]	1974-2003	2,439	4,568	4,935
ISCA	1973-2017	2,257	8,748	9,231
MOBICOM	1995-2017	2,074	5,056	5,732
PODC	1982-2017	1,972	3,573	4,353
POPL	1973-2017	1,858	3,129	3,495
SIGDOC	1982-2017	1,570	1,847	2,048
MICRO	1972-2017	1,321	2,907	3,108
ISSAC	1988-2017	1,253	1,705	2,154
HSCC	1998-2017	361	546	572
	Average	4,025.4	9.569.2	10,961.1
	Median	2,673.0	6,209.0	7,110.0
	Std. Dev.	2,959.3	7,929.9	9,224.4
Full Network (24 SIGs)	1964-2018	79,684	221,541	263,067
DBLP$_J$	1954-2018	298,660	772,281	878,427
DBLP$_C$	1959-2018	439,048	1,381,421	1,830,436
DBLP	1954-2018	617,833	2,022,727	2,687,403

[a] Proceedings discontinued after 2003 and replaced by the ACM Transactions on Graphics.

to their roots. Another feature used as attribute is the relationship between the researchers and their communities, in our case their publication venues. In our experiments we use these two types of attribute separately.

In each network, we modeled authors as nodes and coauthorships in each paper as edges. Each temporal multigraph takes into account the year k of each publication, such that all pairs of coauthors form edges (u,v) in the graph $\mathscr{G}_k$. Note that our model allows the existence of multiple edges between the nodes at the same time k, as discussed in Section 3.1. In the case that we use a title as an attribute, each token taken from it is a member of the set of attributes. Likewise, considering a community as an attribute, such community is seen as a member of the set of attributes.

5.1.2 Questions and Answers Communities

As our second social network, we chose the Stack Exchange[3] network that consists of 173 Q&A communities divided into six categories (Technology, Culture/Recreation, Life/Arts, Science, Professional and Business). Due to its large number of communities, we first randomly selected five communities from each category (except for the Business one that has only three). The data was collected in September 2018.

Given these communities, we considered nodes as representing community members and edges as representing answers to questions, comments to questions and comments to answers as described by Paranjape et al. [2017]. In addition, we considered each time interval as lasting one minute, and tokens taken from the questions and answers as attributes. The data preparation process included removing stop-words and reducing inflected words to their roots (i.e., stemming).

Table 5.2 lists all categories, their respective communities and some of their statistics. As discussed by Posnett et al. [2012] and Vasilescu et al. [2014], the trend is to have several users that are experts and enthusiasts on specific topics. Since that each category and community has specific characteristics, then such networks provide different scenarios for discussion and comparison.

5.2 Properties of the Networks

Finally, inspired by the methodology introduced by Newman [2004], which explores the closeness and betweenness centrality metrics for determining the best positioned nodes in an academic coauthorship network, we rely on network metrics to quantify the potential of our social-based classification method. We evaluate our approach from the perspective of three centrality metrics (*degree*, *closeness* and *betweenness*), which show those classes that tend to have well-defined social roles in a social structure, and the *clustering coefficient*, which measures the degree of cohesion of each class. Table 5.3 formally defines these network metrics, where V represents the set of all nodes. We also use the *PageRank* algorithm [Page

[3] Stack Exchange: `https://stackexchange.com/`

Table 5.2: Statistics of the Questions and Answers communities.

Category	**Community**	**#instants**	**#nodes**	**#multiple edges**
Business	Ask Patents	3,147,891	3,511	10,630
	Project Management	4,201,744	4,645	27,872
	Quantitative Finance	4,181,132	5,972	28,956
Culture/Recreation	Anime & Manga	3,009,735	6,302	28,273
	Board Games	4,287,591	6,321	35,687
	Buddhism	2,213,628	2,249	30,917
	Islam	3,261,927	5,529	26,158
	Vegetarianism	832,128	280	1,767
Life/Arts	Coffee	1,889,441	990	3,772
	Law	1,719,589	5,754	24,726
	Literature	851,488	713	3,900
	Parenting	4,129,958	6,432	34,739
	Pets	2,575,865	3,654	15,553
Professional	Aviation	2,475,645	6,681	49,365
	CS Educators	670,882	807	5,294
	Freelancing	2,777,681	1,769	7,381
	Open Source	1,677,603	1,575	6,257
	Writing	4,183,549	6,265	41,608
Science	AI	1,094,726	1,690	6,203
	Astronomy	2,596,235	3,831	18,002
	Biology	3,823,745	9,684	45,465
	Economics	1,991,423	3,563	16,660
	Theoretical CS	4,230,954	4,588	26,630
Technology	Android	4,903,585	38,412	113,989
	Comp. Graphics	1,616,526	1,017	4,449
	Internet of Things	911,373	748	2,871
	Robotics	3,080,148	3,091	12,193
	Windows Phone	3,340,411	2,450	8,755
	Average	2,670,213.7	4,798.6	22,111.5
	Median	2,596,235	3,563	16,660
	Std. Dev.	1,241,558.2	6,904.3	22,613.4

et al., 1999] to rank the nodes of a graph based on the structure of their ties, thus revealing their importance on the network.

In other words, we employ the notion of social capital given by the strategic positioning of actors in a social structure [Burt, 2005; Granovetter, 1973] in order to evaluate the assigned classes. In this way, we expect that the values of the aforementioned metrics associated with the nodes and edges are able to reveal their social importance given by our proposed classification method. For instance, we expect nodes and edges assigned to *closure*

Table 5.3: Network metrics.

Metric	Formula
Degree centrality of a node *i*	$d_i = \frac{\sum_{j \in V} a_{ij}}{\underset{x \in V}{\arg\max} \sum_{y \in V} a_{xy}}$, where $a_{ij} = \begin{cases} 1, & \text{if the edge } (i,j) \in V \\ 0, & \text{otherwise} \end{cases}$
Closeness centrality of a node *i*	$cl_i = \frac{\lvert V \rvert - 1}{\sum_{j \in V} d(i,j)}$, where $d(i,j)$ is the distance between nodes i and j
Betweenness centrality of a node *i*	$bc_i = \sum_{s,t \in V : s \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}, s \neq i, t \neq i$ where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(i)$ is the number of those paths that pass through the node i
Betweenness centrality of an edge *e*	$bc_e = \sum_{s,t \in V : s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$ where $\sigma_{st}(e)$ is the number of shortest paths from node s to node t that pass through the edge e
Clustering coefficient of a node *i*	$cc_i = \frac{e_i}{n_i(n_i - 1)}$, where e_i is the number of edges between neighbors of i and n_i is the number of neighbors of node i

and *brokerage* classes to have higher betweenness centrality values than those assigned to *innocuous* ones.

5.3 Summary

In this chapter, we presented the experimental methodology designed to assess our proposed social-based classification method. First, we introduced our target social networks derived from two distinct social contexts: academic coauthorship networks and Q&A communities. In general, they have distinct characteristics, which allow us to contrast the effect of our classification method in different social scenarios. Then, based on Newman [2004]'s approach, we presented the idea of applying network properties for determining the importance of nodes and edges by means of their positions in a social structure.

As we shall see in Chapter 6, our analysis covers such social scenarios, thus bringing a more accurate discussion of their behavioral dynamics (e.g., contrasting social behaviors).

Chapter 6

Analysis and Discussion

In this chapter, we characterize several social contexts based on our proposed classification method. In order to evaluate our method for classifying social networks based on social capital concepts, we divide our analysis into three parts. First, we analyze the overall results of our method when classifying the nodes' social behavior and the social meaning of the interactions (Section 6.1). Then, we analyze the effects of our method with respect to the knowledge transfer classification (Section 6.2). Finally, we contrast the results of our two classifications strategies (Section 6.3).

6.1 Social-based Classification

In this section, we present an overall analysis of our classification methodto determine the social role of the nodes and the social meaning of the edges in social networks. For this we apply it to two application domains: academic coauthorship networks and Q&A communities.

6.1.1 Academic Coauthorship Networks

As mentioned in the previous chapter, we investigate separately two scenarios by exploring the persistence of the researchers with respect to the relevant words from their articles' titles (tokens) and the social ties with their respective communities [Silva and Laender, 2018; Silva et al., 2018].

Relevant Tokens. Table 6.1 summarizes the distribution of the node and edge classes for the networks considered. Overall, the classification shows a significant presence of nodes of the class *innocuous*, whereas there is more balanced classification for the edges. First, we focus on the large networks and then we detail the SIG communities.

Table 6.1: Relevant tokens: Social classification of nodes and edges for the coauthorship networks.

Networks		Nodes			Edges		
		closure	*brokerage*	*innocuous*	*closure*	*brokerage*	*innocuous*
24 SIGs	*Average*	27.0%	15.2%	57.8%	29.4%	35.3%	35.2%
	Median	27.1%	14.9%	57.2%	30.4%	34.8%	32.3%
	Std. Dev.	8.5%	2.8%	10.8%	7.0%	10.4%	14.6%
Full Network (24 SIGs)		18.8%	12.2%	69.0%	31.5%	37.1%	31.4%
DBLP$_J$		16.7%	11.8%	71.5%	25.6%	31.8%	42.6%
DBLP$_C$		24.0%	12.0%	64.0%	39.1%	41.6%	19.3%
DBLP		24.2%	12.4%	63.4%	38.5%	41.3%	20.2%

Regarding the DBLP$_J$ network, although its node classes show similar percentages when compared with the *Full Network* (comprising all 24 SIGs), this network stands out with a well differentiated behavior for its edges. The high proportion of 42.6% of *innocuous* edges implies a slight reduction in this figure for the other edge classes. This fact can be seen as specific for this class of venues, which usually does not evidentiate a regular participation of the majority of its members. This can be illustrated by the case of students who tend to publish more regularly in conferences than in journals.

On the other hand, the networks DBLP$_C$ and DBLP have very similar characteristics. The main difference when they are compared with the *Full Network* and DBLP$_J$ networks is their significant percentage of *closure* nodes (24.0% and 24.2%, respectively), since those networks aggregate more information about the academic trajectory of the researchers with a more established career. With respect to the edges, the *brokerage* class is the most representative one with similar percentages. In contrast, there is a low percentage of *innocuous* edges. Overall, the DBLP network shows that 79.8% of its edges, i.e., excluding the *innocuous* ones, carry some relevant information according to the social concepts explored in our study.

Now, we analyze in detail each one of the 24 ACM SIG networks. Figure 6.1 presents the distribution of the node classes for the 24 ACM SIG communities and the *Full Network* that includes all these communities. Overall, the classification shows a significant presence of nodes of the class *innocuous* (average of 57.8%). Indeed, an academic coauthorship network usually has a strong presence of new nodes (e.g., students or sporadic collaborators). Despite that, there is also a strong presence of nodes of the class *closure* with percentages above 30% for more established communities such as CIKM, KDD, SIGIR, SIGMOD, STOC, SIGMETRICS, ISCA, PODC, POPL and MICRO. Particularly, most members from these communities tend to be coherent in the research topics addressed throughout their academic trajectories. In contrast, communities such as SAC, SIGUCCS, SIGGRAPH and

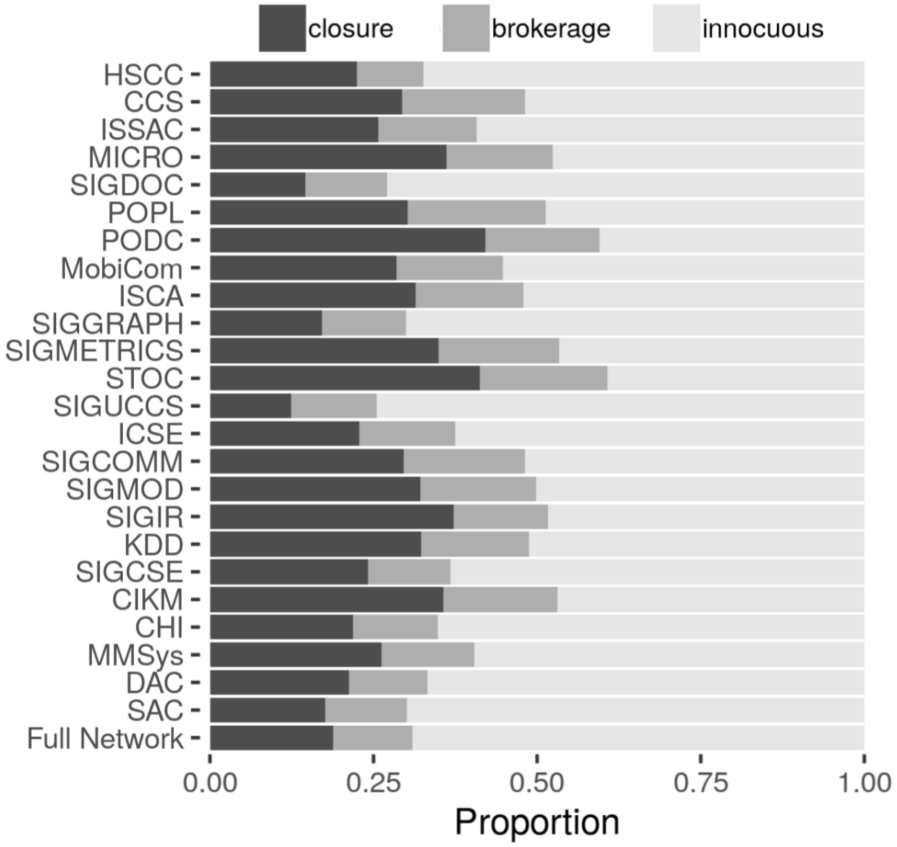


Figure 6.1: Social-based classification of nodes for the 24 ACM SIG communities and the Full Network.

SIGDOC show percentages below 18% for the class *closure*, which represents some lack of synergy among their members. Particularly, SIGUCCS (University and College Computing Services) and SIGDOC (Design of Communication) are two communities that address very specific topics. SIGGRAPH (Computer Graphics), although a well established scientific community, covers here only its editions up to 2003, since after that year their proceedings were discontinued and replaced by special issues of the ACM Transactions on Graphics.

Generally, such percentages can be seen as evidence of the characteristics of each community. For instance, members of the STOC (Theory of Computing) community have a tendency to show more competence in specific topics related to computation theory, thus the higher number of nodes of the class *closure* (41.3%). On the other hand, SAC (Applied Computing) is a community mainly focused on applied issues, thus covering a wide range of topics, which justifies the high number of *innocuous* nodes (69.9%).

Regarding the edge classification, Figure 6.2 presents the distribution of the edge classes for the 24 ACM SIG networks and the *Full Network*, which comprises the 24 SIG networks altogether. As we can see, most of these edge classes carry some kind of infor-

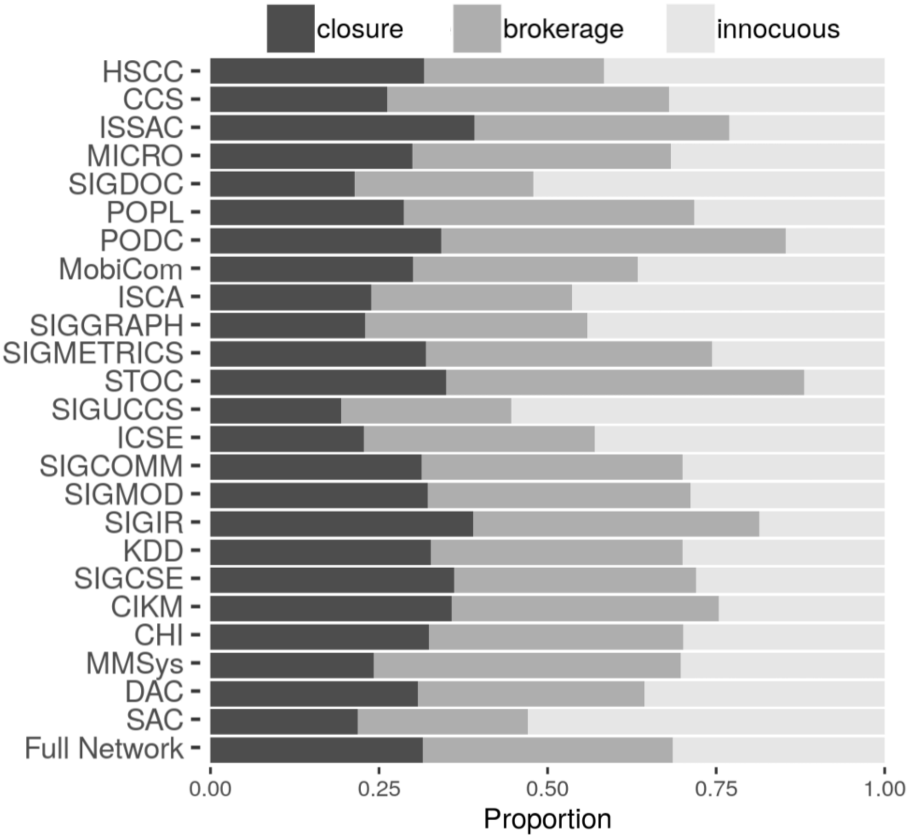


Figure 6.2: Social-based classification of edges for the 24 ACM SIG networks and the Full Network.

mation and have been characterized as *closure* or *brokerage* (on average, they sum 64.7%), thus demonstrating a strong social tie between the researchers and their relevant topics. On the other hand, edges without any social meaning (i.e., *innocuous*) tend to be less present in the networks. Again, specific communities show a singular behavior, such as ISSAC (Symbolic and Algebraic Computation) and SIGIR (Research and Development in Information Retrieval) with the highest presence of *closure* edges. SAC, SIGUCCS and SIGDOC also stand out for having an expressive number of *innocuous* edges (more than 50%), thus reinforcing the fact their members show no regularity with their research topics.

The *Full Network* shows a substantial drop from 27.0% to 18.8% in the number of *closure* nodes when compared with the average of all 24 ACM SIG conferences (Table 6.1 and Figure 6.1). Despite that, the number of *innocuous* nodes considerably increased from 57.8% to 69.0%. This was expected due to the fact that more active nodes (researchers) tend to participate in more than one community. Thus, with more subjects covered, the likelihood of having many relevant attributes decreases.

With respect to the edges (Table 6.1 and Figure 6.2), all classes tend to maintain their

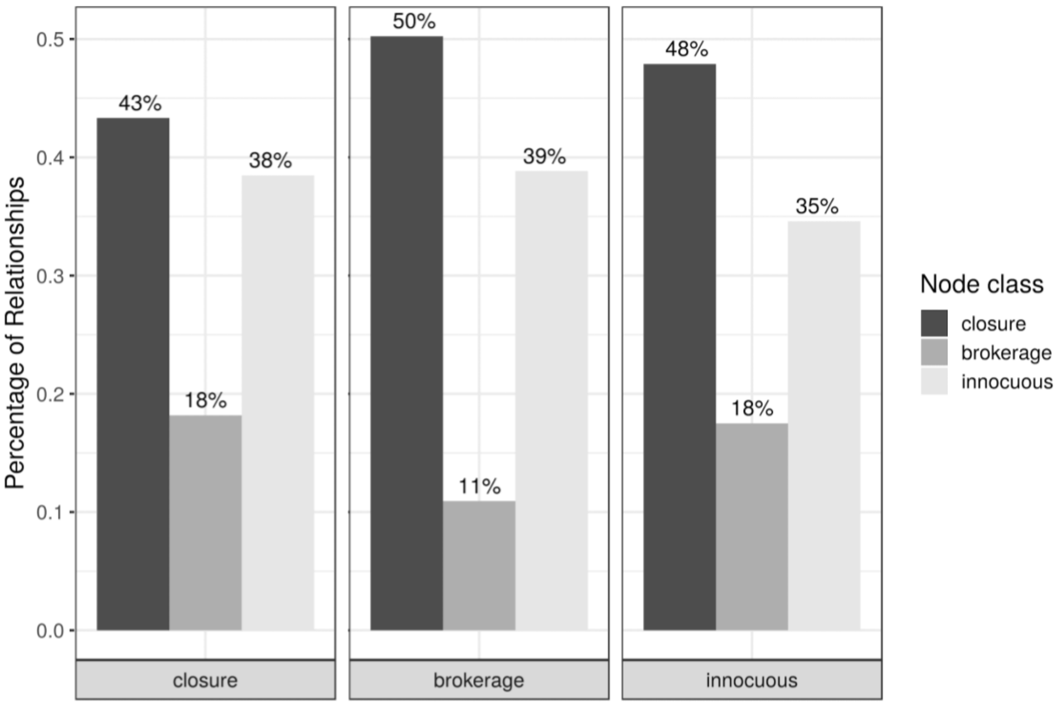


Figure 6.3: Proportion of relationships for each node class.

proportions when compared with the average of all 24 ACM SIG networks (maximum variation of 3.8%). This emphasizes the importance of characterizing the social meaning of the relationships because, although there are fewer *closure* nodes, there is still relevant information flowing through them.

Analyzing the relationships according to their nodes' classes, Figure 6.3 shows their percentages for each one. In proportion, all nodes tend to maintain strong relations with those of the class *closure*, followed by the *innocuous* ones. In all cases, relationships with *brokerage* nodes are less frequent, emphasizing a tendency to privilege relationships with the most important nodes in the network (*closure*) or to establish new social ties with specific ones (*innocuous*). We can see such results, especially the *innocuous-closure* relationships, as a social tie between expert and novice. We also can interpreted these collaborative patterns as a close-knit research group with newcomers and, to a lesser extent, bridging with other social groups. For example, a cohesive computer theory team of experienced members that includes new graduate students annually, as well as creating bridges when applying their results in other areas.

Finally, Figure 6.4 shows the *Full Network*, where the edges classified according to the concept of *closure* are shown in blue (*strong ties*), those classified according to the concept of *brokerage* are shown in red (*weak ties*) and those that express no social meaning are shown in black (*innocuous*). Note that the edges based on the *brokerage* and *closure* concepts

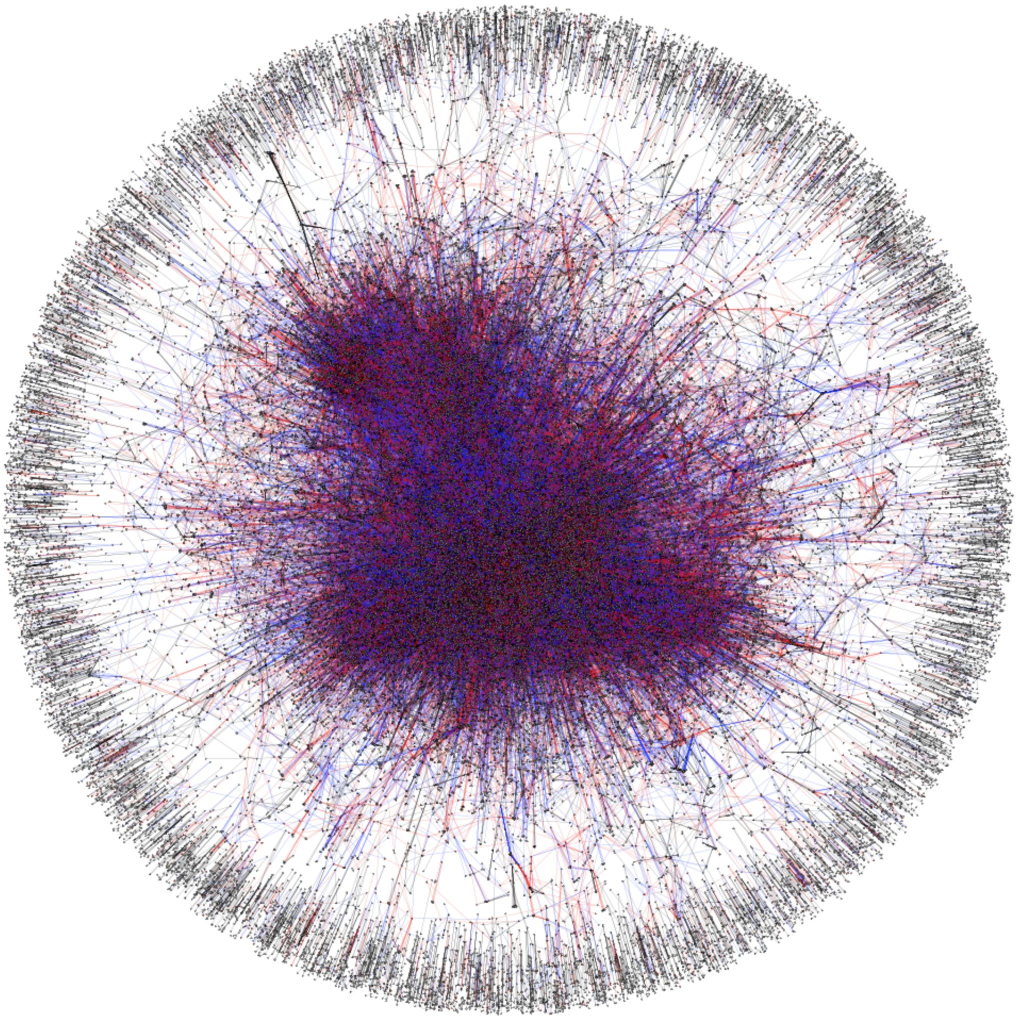


Figure 6.4: Social-based classification of the 24 SIGs communities. Blue edges emphasize the *closure* concept (strong ties) and red ones the *brokerage* concept (weak ties). Black edges correspond to those regarded as *innocuous* (i.e., edges that have no important information passing through them). **Best viewed in color.**

dominate the center of the graph, while the extremities tend to have a greater prominence of edges regarded as non-relevant (those in black). This means that edges strongly related to social concepts tend to be better positioned in a social structure (i.e., linked to central nodes), which provides early access to information passing through the network.

Relevant Community Ties. By exploring the social ties with communities, Table 6.2 shows a very low percentage of nodes classified as *closure*. In other words, these figures show that not all nodes have strong ties with certain communities. Nevertheless, as the networks grow, from the *Full Network* (all 24 SIGs) to the entire DBLP, there is an increase in the *closure* rate. Even so, note that the percentage of nodes acting as bridges for others is quite stable,

Table 6.2: Community ties: Social-based classification of nodes and edges for the coauthorship networks.

Networks	Nodes			Edges		
	closure	brokerage	innocuous	closure	brokerage	innocuous
Full Network (24 SIGs)	0.7%	29.7%	69.6%	1.6%	6.7%	91.7%
DBLP$_J$	2.5%	26.0%	71.5%	3.8%	17.8%	78.4%
DBLP$_C$	6.5%	29.5%	64.0%	9.5%	43.3%	47.2%
DBLP	7.3%	29.2%	63.5%	9.6%	46.7%	43.7%

corresponding to almost 30% of the total in all cases.

Considering the classification of edges, we can see a distinct scenario regarding the assigned classes. Despite that, the *closure* class also has the lowest percentages for all networks. As the *Full Network* includes few communities, very low values of relevant classes are expected (specifically, *closure* and *brokerage* sum up 8.3%). Although with more expressive values than the *Full Network* for important classes, the network DBLP$_J$ stands out with less than a quarter of edges passing important information through the edges. In contrast, the DBLP$_c$ and DBLP show similar values, highlighting more than half their edges with a strong bond within communities.

Note that the percentages of the *innocuous* class decrease as the networks become larger (e.g., 91.7% for the *Full Network* and 43.7% for the entire DBLP network). Considering the DBLP network, we observe that the social ties of its researchers with their communities tend to present a low percentage of *closure* nodes (few strong ties), but with a remarkable presence of edges of the classes *brokerage* (46.7%). Overall, considering *closure* and *brokerage* classes altogether, only 8.3% of the edges carry some social meaning in the *Full Network* and 21.6% in the DBLP$_J$ network. On the other hand, in the DBLP$_C$ and DBLP networks these same classes represent 52.8% and 56.3% of the edges.

6.1.2 Questions and Answers Communities

As we only consider frequent users in the Q&A communities (see Section 5.1), by definition there are no *innocuous* nodes in these networks [Silva et al., 2018]. With respect to the node classes, Figure 6.5 shows few variations in the percentages of *closure* and *brokerage* nodes across the communities (average values of 79.8% and 20.2%, respectively). More specifically, the *Vegetarianism* and *Buddhism* communities show the highest proportions for the *closure* class (87.5% and 85.3%, respectively), whereas *Anime & Manga* stands for 72.8%.

In contrast, we notice that the full academic coauthorship network had 18.8% of its nodes classified as *closure*, 12.2% as *brokerage* and 69.0% as *innocuous* (see Table 6.1). Indeed, there are few *closure* nodes (e.g., research leaders) in an academic network compared

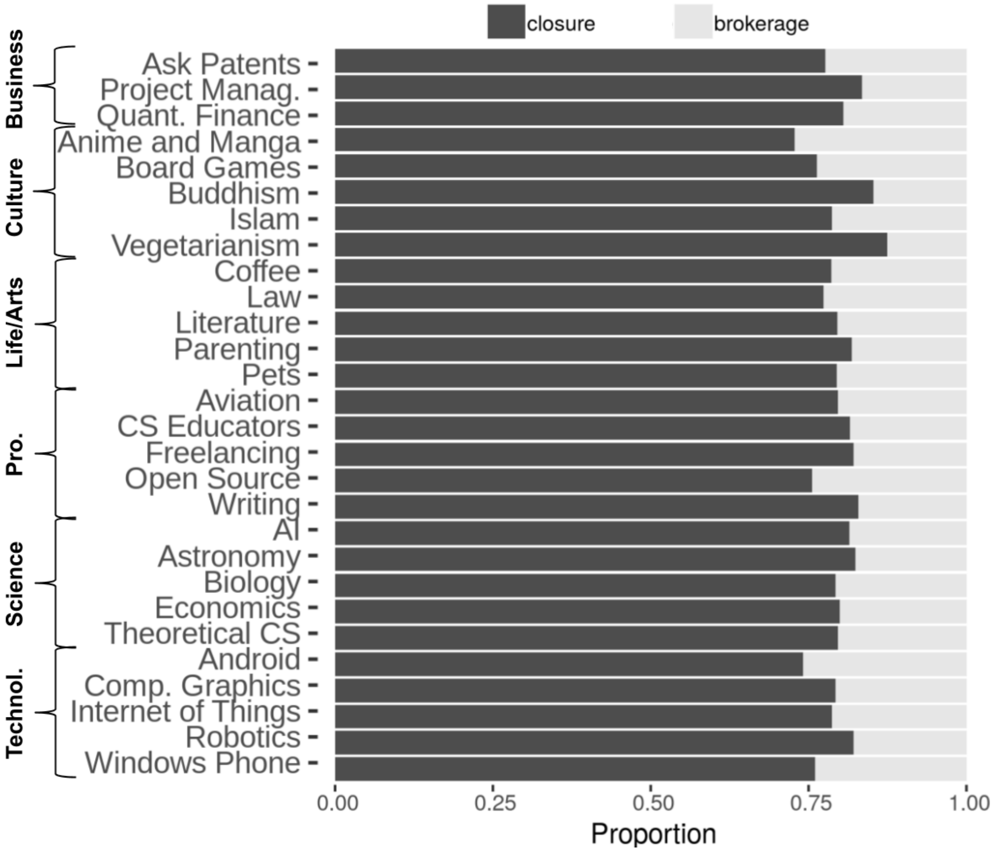


Figure 6.5: Social-based classification of nodes for the 28 Q&A communities.

with the other ones (e.g., new students). However, in the Q&A communities, users are in general experts and enthusiasts about specific topics, which gives them some authority [Posnett et al., 2012; Shah and Kitzie, 2012; Vasilescu et al., 2014].

Considering the social classification of the edges in Figure 6.6, the proportions by category and by community have significant oscillations, thus reinforcing a distinct behavior of our classification method on several topics. For example, the *Buddhism* community (85.3% of *closure* nodes) has 84.7% of *closure* edges, whereas the *AI* community (81.4% of *closure* nodes) has a much smaller proportion of 63.8% of edges of that same class. There are also notorious divergences between communities in the same category such as *Ask Patents* and *Quantitative Finance* from the *Business* category, *Aviation* and *Freelancing* from the *Professional* category, and *Literature* and *Parenting* from the *Life/Arts* category. As we only selected frequent users, it justifies the very low presence of *innocuous* edges.

By comparing the Q&A distribution by communities with the same figures from the academic ones (see Table 6.1), we observed that the entire DBLP academic coauthorship network had 38.5% of its edges classified as *closure*, 41.6% as *brokerage* and 20.2% as *innocuous*. That is, we note that both scenarios reveal very different proportions of assigned

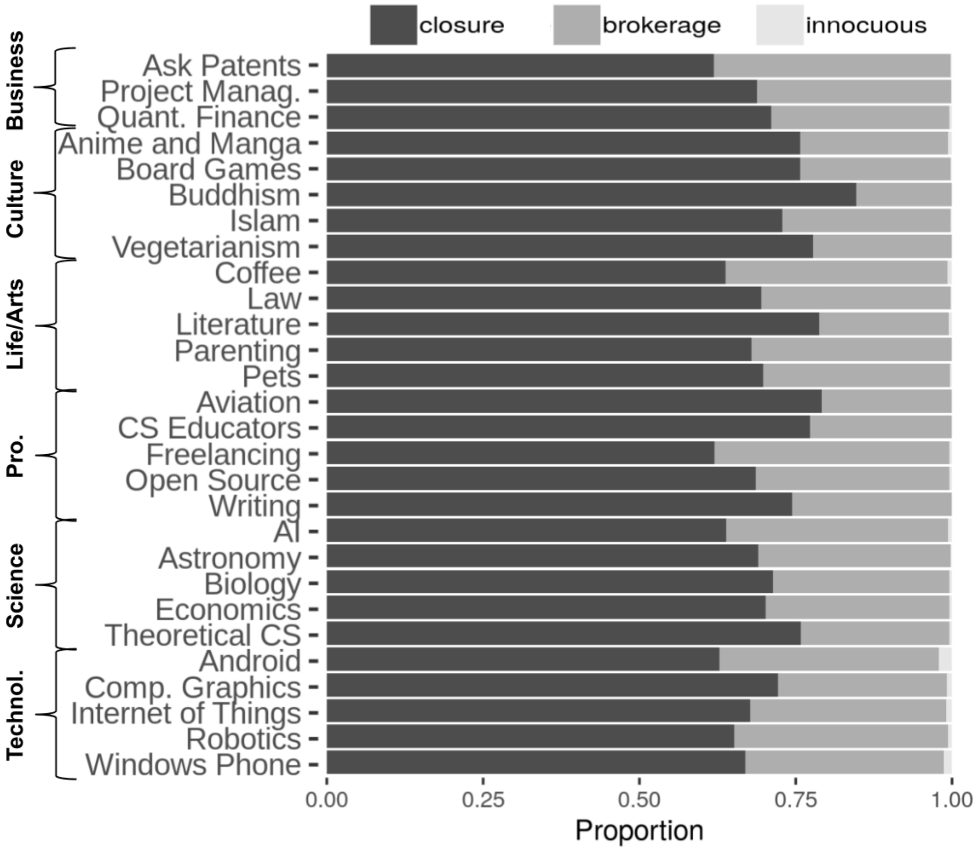


Figure 6.6: Social-based classification of edges for the 28 Q&A communities.

classes, particularly with a higher proportion of the *closure* class in the Q&A scenarios, whereas in the academic scenarios the most representative class tended to be *brokerage*.

6.2 Knowledge Transfer Dynamics

Now, we analyze the knowledge transfer dynamics by characterizing results regarding knowledge transfer across the edges and the social behavior of the nodes [Silva et al., 2019, 2020]. Next, we present characterization results of distinct social scenarios derived from the 24 SIG networks (Subsection 6.2.1) and the Question and Answers communities (Subsection 6.2.2).

6.2.1 Academic Coauthorship Networks

Although our discussion is about dynamics of multiple edges over time, let us first analyze a static scenario considering the final aggregated graph (i.e., containing all interactions as a single edge) [Silva et al., 2019]. In this context, we identify whether in the final stage of the interactions between two researchers there is a transfer of new attributes, thus neglecting

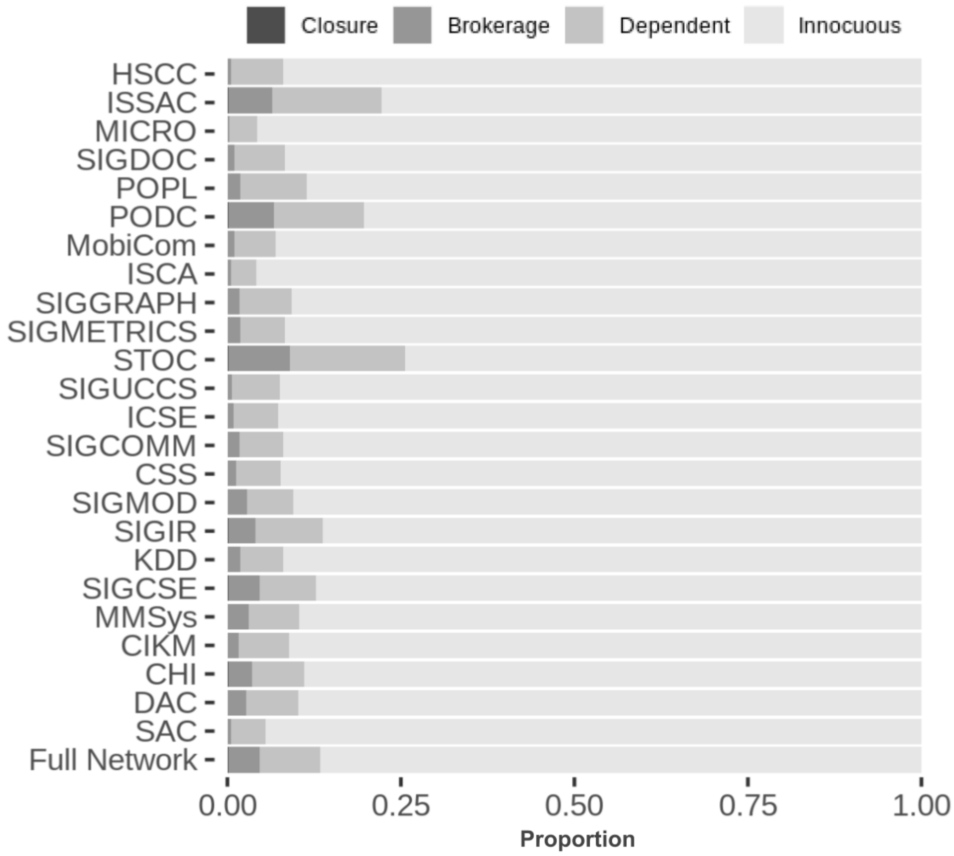


Figure 6.7: **Aggregated Graph.** Knowledge transfer classification of edges for the 24 ACM SIG communities and the Full Network. Networks sorted in ascending order according to their number of nodes.

the exchanging of information in previous periods. In this way, for instance, this allows us to conclude whether the current relationship between two individuals is still a source of knowledge (otherwise, both have reached a common maturity on a set of subjects).

Figure 6.7 shows the distribution of knowledge transfer classes according to the aggregated directed graph. First of all, it is clear that the *closure* class is not representative in any academic network (i.e., bidirectional edges are rare). On the other hand, relationships that do not present a well-defined knowledge transfer (i.e., *innocuous* edges) represent more than 74% of the edges in all networks. Although such a percentage certainly includes many newcomers (e.g., students and researchers from other areas), another explanation for this case is the fact that new mentions to specific associated attributes (i.e., new subjects) originated from a set of nodes (e.g., an article coauthored by several people) start becoming noticeable from that instant onwards. In fact, scientific knowledge is highly dynamic with new subjects constantly coming up.

Since such percentages can be seen as evidence of how knowledge is transferred within a network, we can now analyze specific aspects of each community. For instance, members

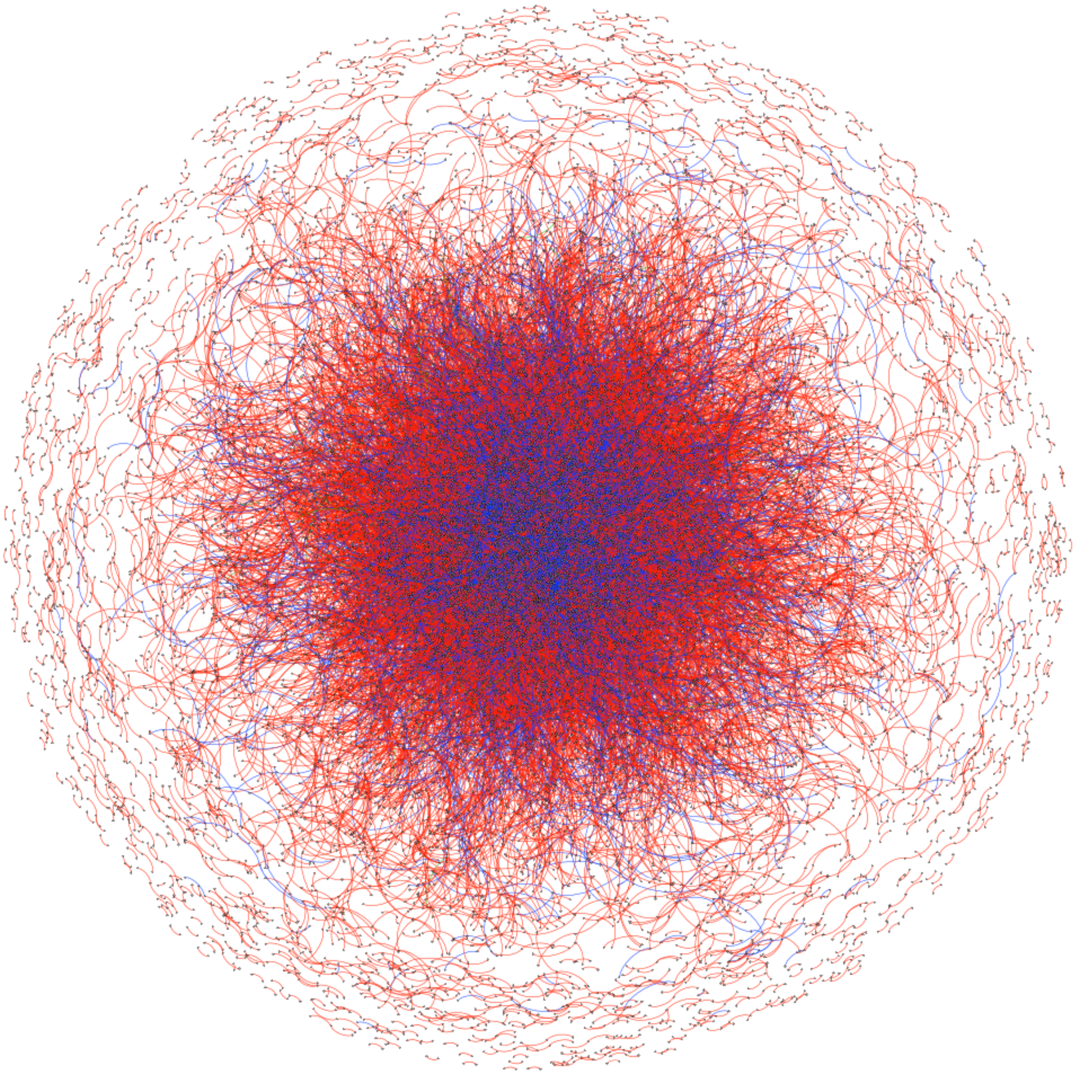


Figure 6.8: Knowledge transfer classification of the 24 SIG communities. Red edges represent brokerage relationships and blue edges dependent ones.

of the STOC (Theory of Computing) community have a tendency to show more competence in specific topics related to computing theory, which are usually more stable across time. Thus, the high number of edges with knowledge transfer detected (25.6%), which suggests a strong cohesion in knowledge dissemination within this community. On the other hand, SAC (Applied Computing) is a community mainly focused on applied computing, thus covering a wide range of topics (i.e., small groups focused on specific issues that do not collaborate with other internal groups), which justifies the low number of meaningful knowledge transfers (5.5%).

Focusing on the *brokerage* and *dependent* classes, the ISSAC (Symbolic and Algebraic Computation), STOC (Theory of Computing) and PODC (Principles of Distributed

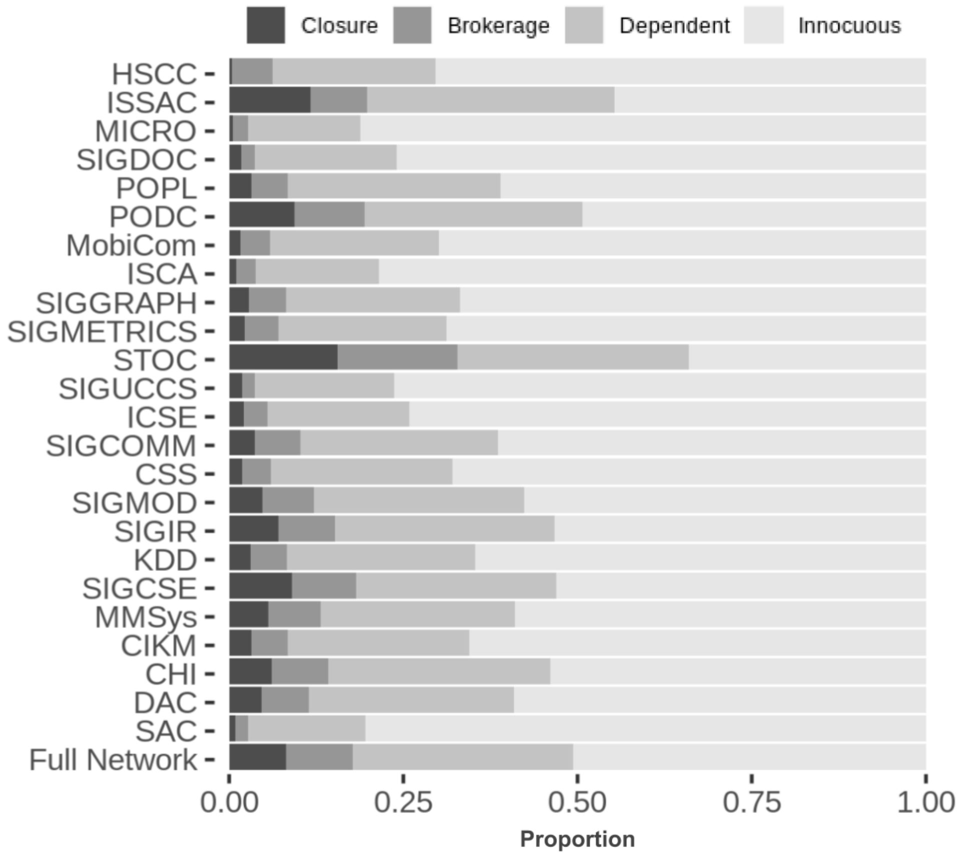


Figure 6.9: **Dynamic Multigraph.** Knowledge transfer classification of edges for the 24 ACM SIG communities and the Full Network. Networks sorted in ascending order according to their number of nodes.

Computing) communities stand out for having a *brokerage* class percentage of more than 6% and a *dependent* class representation of over 12%. Again, such communities have a strong theoretical background, favoring *dependent* relationships that involve the same group of collaborators. Such an evidence is in line with the tendency that theoretical subjects are more consolidated, whereas technological ones are more ephemeral.

Regarding the *Full Network* (last bar), which comprises all interactions across the 24 ACM communities, it shows that our previous observations are sustained. Specifically, Figure 6.8 illustrates the relationship classes of the type *brokerage* (in red) and *dependent* (in blue). *Closure* relationships (green edges) are rare and cannot be clearly visualized in this picture. As expected, there are more *dependent* edges (65.2%) than *brokerage* ones (33.9%). Note that the most central part of the graph tends to be blue, reinforcing the closure effect, i.e., there are important nodes linked to them. On the other hand, the extremities tend to have more edges classified as *brokerage*, reflecting a bridging aspect.

Regarding the dynamic directed multigraph (Section 3.3), which differs from the aggregated one by comprising in terms of parallel edges all multiple interactions between nodes

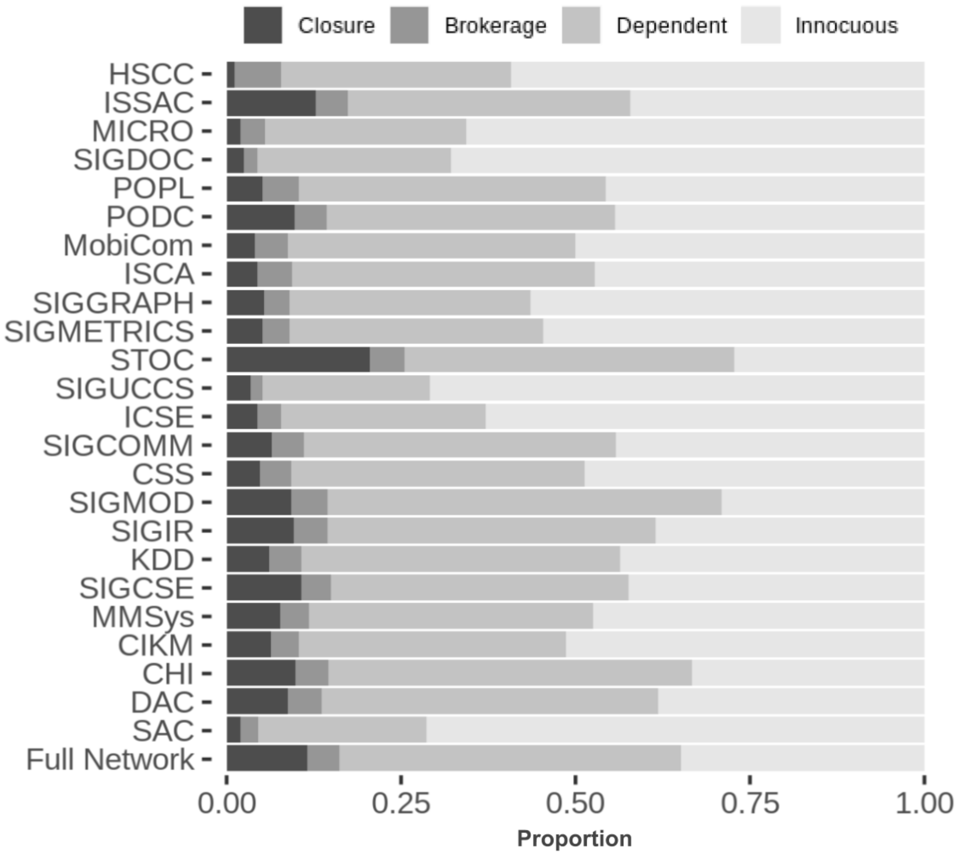


Figure 6.10: Knowledge transfer classification of nodes for the 24 ACM SIG communities and the Full Network.

over time, there is a greater dissemination of knowledge when inspecting the interaction history (Figure 6.9) [Silva et al., 2020]. Note that this new multiple interaction representation is fully expressed by distinct edges and not by a single one. Observe that this dynamic approach identifies greater proportion of edges with knowledge transfer (*closure* and *brokerage*), in spite of a predominance of the classes *dependent* and *innocuous*. This occurs because usually there is more knowledge transfer in the initial interactions than in the final ones, since nodes tend to learn similar knowledge along the time. Moreover, *brokerage* and *closure* edges are more likely to be associated with more experienced researchers, who in general collaborate many times along their careers. Again, the proportion of knowledge transfer edges tend to be larger in communities having a strong theoretical basis, such as ISSAC, PODC and STOC, whereas communities like SAC still present low knowledge dissemination. Overall, our discussion about the aggregated scenario also holds for the dynamic multigraph one.

With respect to node characterization, Figures 6.10 show the node's classification according to their behaviors when transferring knowledge. In particular, the nodes from the academic networks tend to achieve larger percentages in the classes *dependent* and *innocuous*. We can see it as a particular academic scenario where there are mentors (strongly bond to specific attributes) who have a timid collaboration with other members who are also very knowledgeable on certain attributes (i.e., establishing a two-way bridge), rather than establishing relationships with new students and young researchers. Nevertheless, there is the fact that few members are highly influential in their respective communities [Alves et al., 2013].

Comparing these results with the dynamic edge classification (Figure 6.9), in general, we observe that the percentages of *brokerage* classes are smaller, while the percentages of *dependent* ones tend to be larger. For the *closure* classes, there are slightly higher values. Particularly, the STOC network presents percentages of the *closure* class of 20.6% and of the *innocuous* one of 27.2%, which differs from its edge percentages of 15.6% and 34.0%, respectively. On the other hand, the tendency of members of the SAC network having a non disseminating behavior is confirmed.

6.2.2 Questions and Answers Communities

We now report the results of the social scenarios in the Q&A communities [Silva et al., 2019, 2020]. We omit the results from the static scenario (i.e., single-edge aggregated graph) because they are very similar to the dynamic one. One explanation for this similarity is that, unlike the scientific coauthorships, in the Q&A communities users do not tend to keep long interactions with the same group of people. Thus, we next discuss only the dynamic scenario and refer the reader to Silva et al. [2019] for more details about the aggregated knowledge transfer classification.

Regarding knowledge transferred across the network by the edges, Figure 6.11 clearly shows the *closure* class as the most representative for all communities, achieving more than 50.0% and reflecting the own nature of the social interactions in Q&A communities [Shah and Kitzie, 2012; Vasilescu et al., 2014]. Indeed, such online networks offer an environment of valuable information resources, thus enabling people to share their knowledge. This contrasts with the low presence of the *closure* class in academic social networks (see Figures 6.7 and 6.9), in which new knowledge tend to be originated from a set of nodes (e.g., coauthors of their first paper on a subject) instead of from single individuals. Similar to the academic context, the percentages of the *dependent* class are superior to those of the *brokerage* one.

Analyzing by category, in general, there are similar minimum and maximum percentages. However, the *dependent* class shows a great divergence. Specifically, communities from the Technology category present higher figures for the *closure* class, which is in line with the tendency that technological discussions are more ephemeral.

Overall, there are different percentages of classifications by communities. For exam-

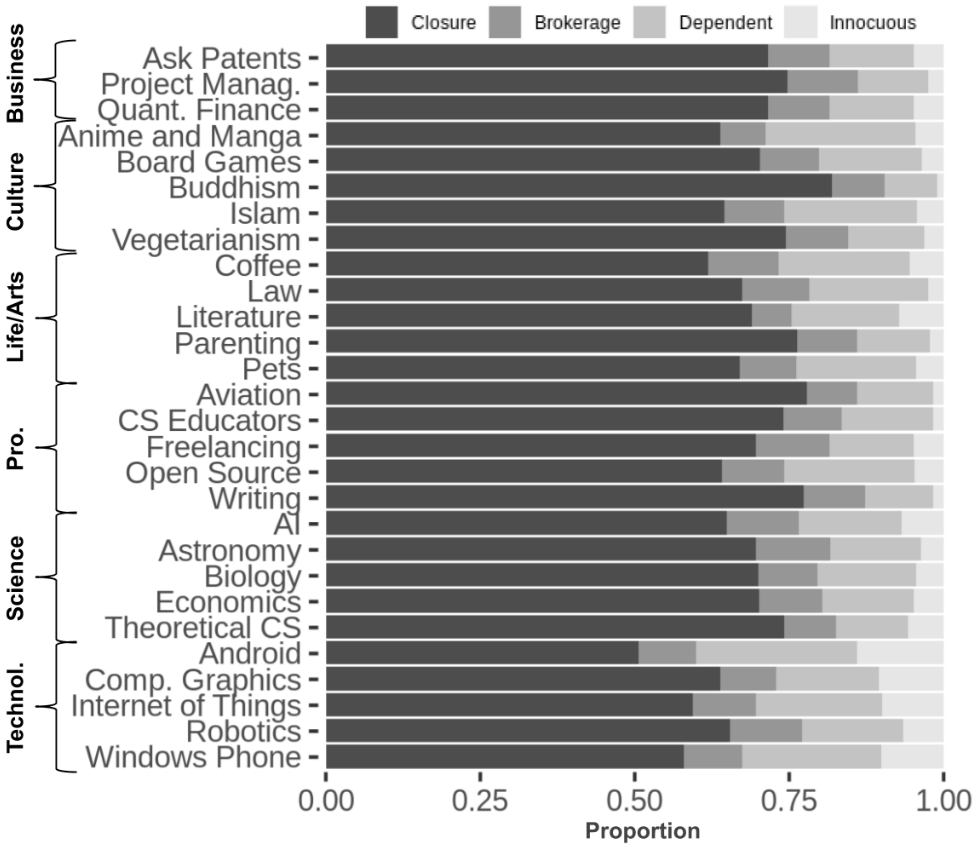


Figure 6.11: Knowledge transfer classification of edges for the Questions & Answers communities.

ple, the *Ask Patents* (Business) and *Android* (Technology) communities reveal very different percentages by class compared with the others. Focusing on communities from the same category, there are also distinct behaviors such as the case of *Buddhism* and *Islam* that, although discuss issues related to religious topics, their knowledge-sharing behaviors are discrepant. In fact, note that the *Buddhism* and *Writing* communities are the ones with the highest proportion of knowledge transfer edges, what may be a consequence of their more closed and coherent group of users.

With respect to node characterization, Figure 6.12 shows that the nodes from the Q&A communities tend to achieve larger percentages in the class *closure*, whereas the the classes *dependent* and *innocuous* are the most expressive in the academic networks. There is a sharp decrease in the percentages of the class *brokerage* and a slight one in the class *closure* when compared with the edges (see Figure 6.11). On the other hand, the *dependent* class achieves percentages of more than 20% for all communities, thus emphasizing a social behavior of just receiving knowledge from other users. Even so, the *closure* class is still the most representative of all communities with more than 40%, which can be interpreted as existing a clear

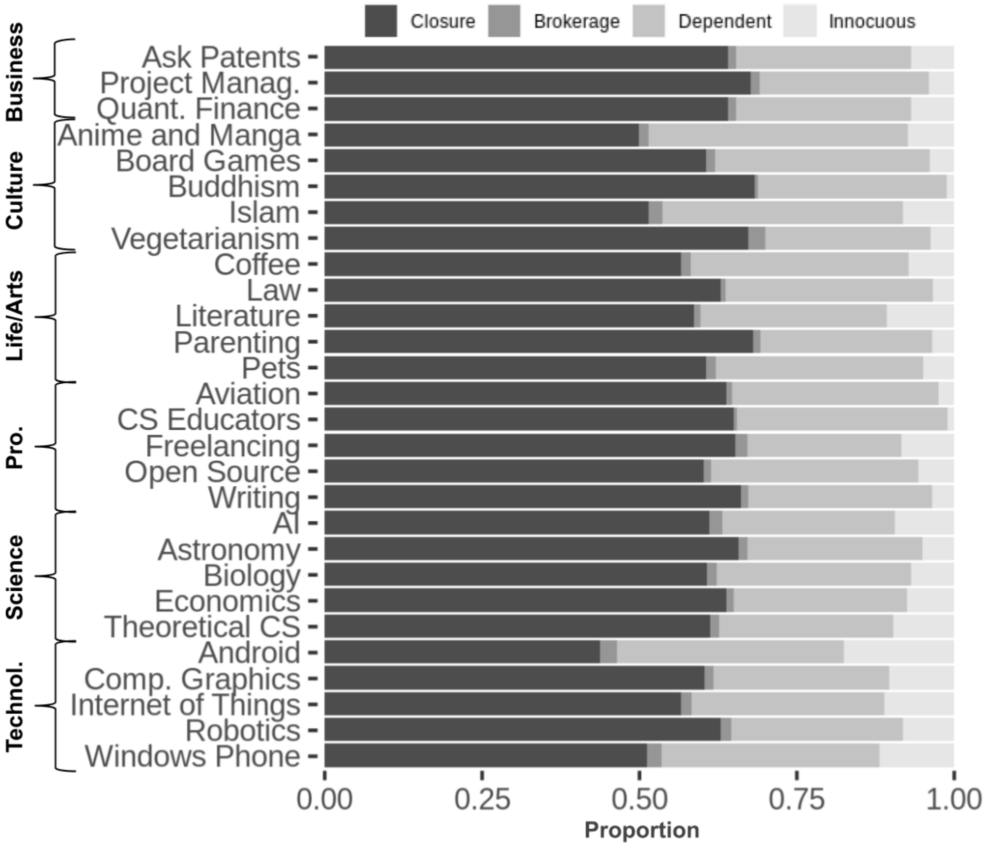


Figure 6.12: Knowledge transfer classification of nodes for the Questions & Answers communities.

social behavior of both teaching and learning when interacting with other users. Indeed, this reflects the own nature of the social interactions in the Q&A communities, which offers an environment of valuable information resources, thus allowing people to share their knowledge [Neshati et al., 2017]. Nevertheless, the proportions of edge and node classifications are consistent.

6.3 Comparative Analysis

We now contrast our two classification strategies for determining the strength of social structures (i.e., social-based classification), regarding the dynamics of knowledge transfer (Subsection 6.3.1). As we are interested in characterizing social interactions and in order to check how each method separates social interactions from non-relevant ones, we expanded our analysis to compare our strategies against the RECAST algorithm (*Random rElationship ClASsifier sTrategy*) [Vaz de Melo et al., 2015] (Subsections 6.3.2 and 6.3.3), which provides a strategy for identifying random and social interactions based on network properties.

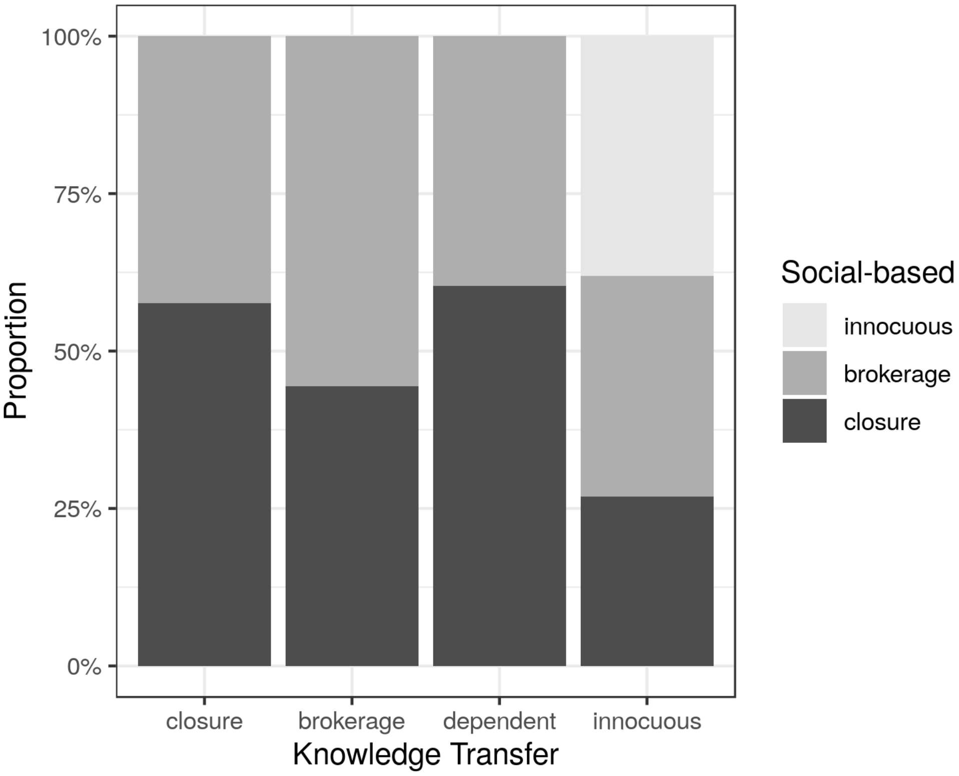


Figure 6.13: Social-based classification (edge class) versus knowledge transfer classification (bars) for the Full Network.

6.3.1 Social Structure versus Knowledge Transfer

Figure 6.13 depicts the strengths of social structures (Edge class) versus the transfer of knowledge (bars). First, we note a discrepancy between the two classifications, thus showing that they represent two distinct approaches based on different social concepts. For example, not all edges classified as *closure* or *brokerage* by the knowledge transfer classification (first bar) are respectively classified as *closure* or *brokerage* by the social-based classification (values inner the bars). Even so, there is a clear trend that reveals the classes associated with social concepts as those responsible for the great diversity in the social dynamics. On the other hand, analyzing the *innocuous* bar, we find that the highest percentage of edges in this knowledge-transfer class is associated with the social structural *brokerage*. However, we expected that no *brokerage* and *closure* instances would be assigned as *innocuous*, since the *innocuous* class represents relationships with no social meaning.

With respect to the nodes, Figure 6.14 contrasts the knowledge-transfer classes according to the social role of the nodes involved in each relationship. For example, *closure-brokerage* indicates that an edge was established by connecting a node of the *closure* class with a node of the *brokerage* class. As expected, it is more evident that the knowledge-

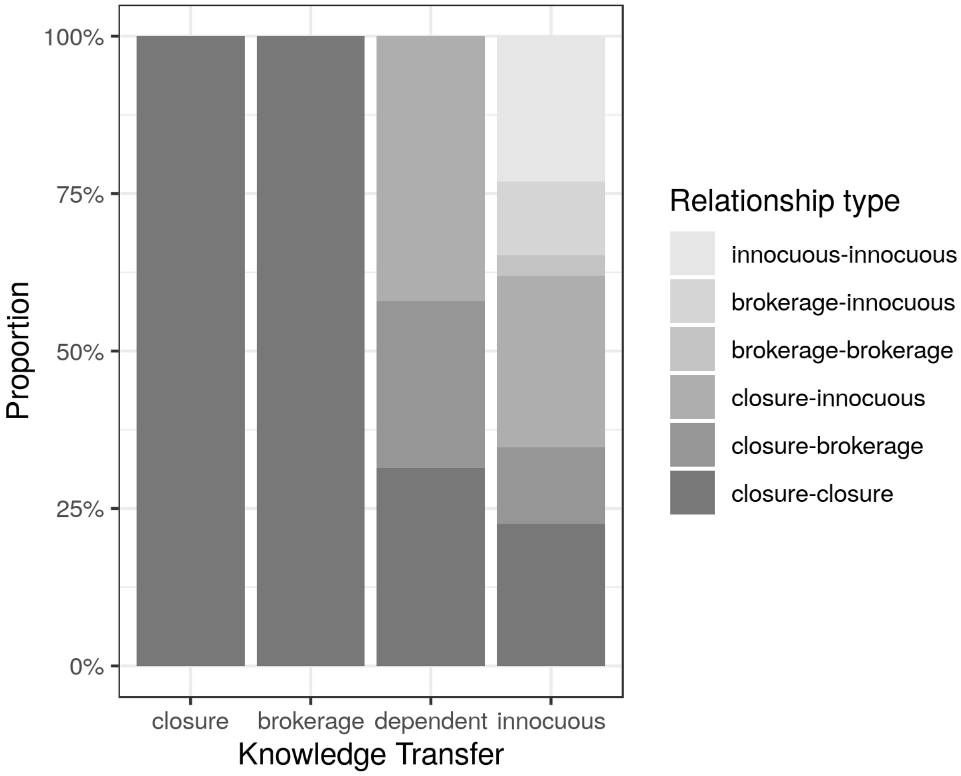


Figure 6.14: Social-based classification (relationship type) versus knowledge transfer classification (bars) for the Full Network. Darker tones represent more important social ties.

Table 6.3: RECAST classification of edges for the Full Network.

RECAST classes			
friend	bridge	acquaintant	random
2.23%	0.18%	84.53%	13.06%

transfer relationships of the most important classes (*closure* and *brokerage*) tend to be strongly associated with the most important structural relationship types (darker tones). On the other hand, the bars *dependent* and *innocuous* are more mixed, but in a less extent to relevant relationships (dark tones) than less important ones (lighter tones).

In general, even the methods based on different social perspectives, the contrast observed by the classification results shows a coherence between the methods. Thus, the greater the edge's strength of their relationships, the more diverse are their dynamic behavior.

6.3.2 Social Structure versus RECAST

We now contrast our results with those of the RECAST algorithm [Vaz de Melo et al., 2015]. As discussed in Section 2, the RECAST algorithm assigns social classes to edges in tempo-

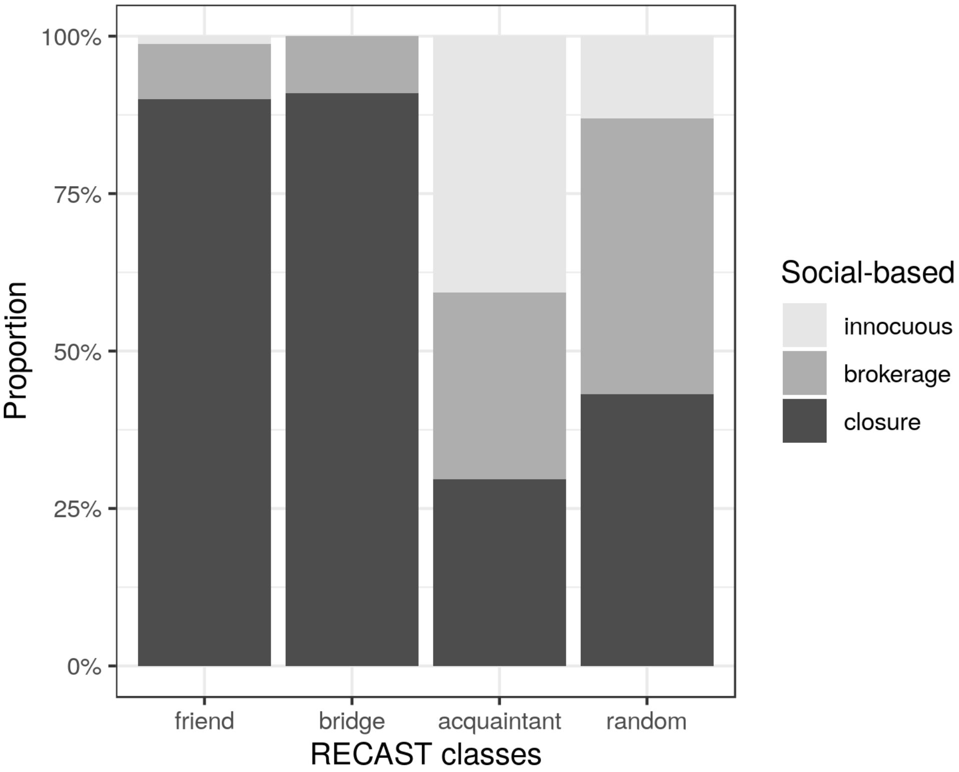


Figure 6.15: Social-based classification of the edges classified by the RECAST algorithm.

ral networks and was used in some previous works to characterize academic coauthorship networks [Brandão et al., 2017, 2018; Leão et al., 2018]. For this, it explored the regularity of relationships and the topological overlap existing among them over time. By comparing such regularities with random temporal graphs, it classifies social ties as *friend*, *bridge*, *acquaintant* and *random*. As the graph model adopted by RECAST is single-edge, we first transformed our aggregated multigraph into a single-edge one, in which the most representative edge class associated with each pair of nodes become the actual edge class in the new graph (draws were resolved considering the following importance order: *closure*, *brokerage* and *innocuous*).

First, we present the results of the RECAST classification in Table 6.3. We clearly note an unbalanced classification, showing the *acquaintant* class as the most representative (84.53%), followed by the random one (13.06%). Such results are compatible for different academic networks analysis, where in addition to relationships of type *acquaintant*, also revealed a large number of *random* ties [Brandão et al., 2017].

Considering that the RECAST algorithm separates social from casual relationships, we expect that the important social classes defined by it (*friend* and *bridge*) to be strongly associated with our most important ones (*closure* and *brokerage*). First, we analyze the pro-

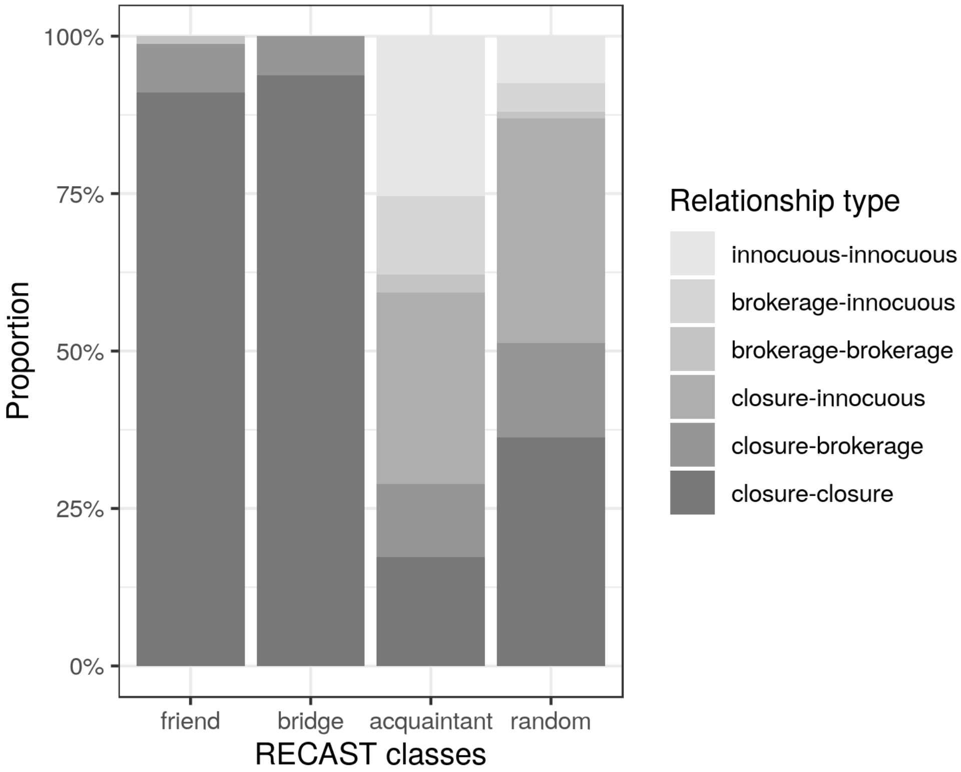


Figure 6.16: Social role of the nodes classified by the RECAST algorithm. Darker tones represent more important social ties.

portion of the social meaning of the classes grouped by the RECAST classes in Figure 6.15. Looking at the *friend* and *bridge* bars, we can see that their edges are mostly associated with the two most important classes of our method, *closure* and *brokerage*. Moreover, there is a very low overlapping of the *innocuous* edges with the *friend* and *bridge* ones. Although we expected the *innocuous* classes to be mostly associated with casual relationships (i.e., *acquaintant* or *random* classes), in contrast, the last two bars tend to be quite diverse with more equal proportions with respect to our classes. In the next section, we will better analyze the results of the *random* class according to the network properties, thus validating the consistency of our social classification.

Regarding the nodes, Figure 6.16 shows how the RECAST classes have been assigned according to the social role of the nodes involved in each relationship. In this scenario, we see clearly that the relationships of the most important nodes considered by our method (darker tones) tend to be strongly associated with the most important classes of the RECAST algorithm (*friend* and *bridge*). Again, there is more diversity in the proportion bars of the *acquaintant* and *random* RECAST classes, but with the presence of less important classes (lighter tones). Overall, even the methods based on different social perspectives, they are coherent regarding the edges' strength.

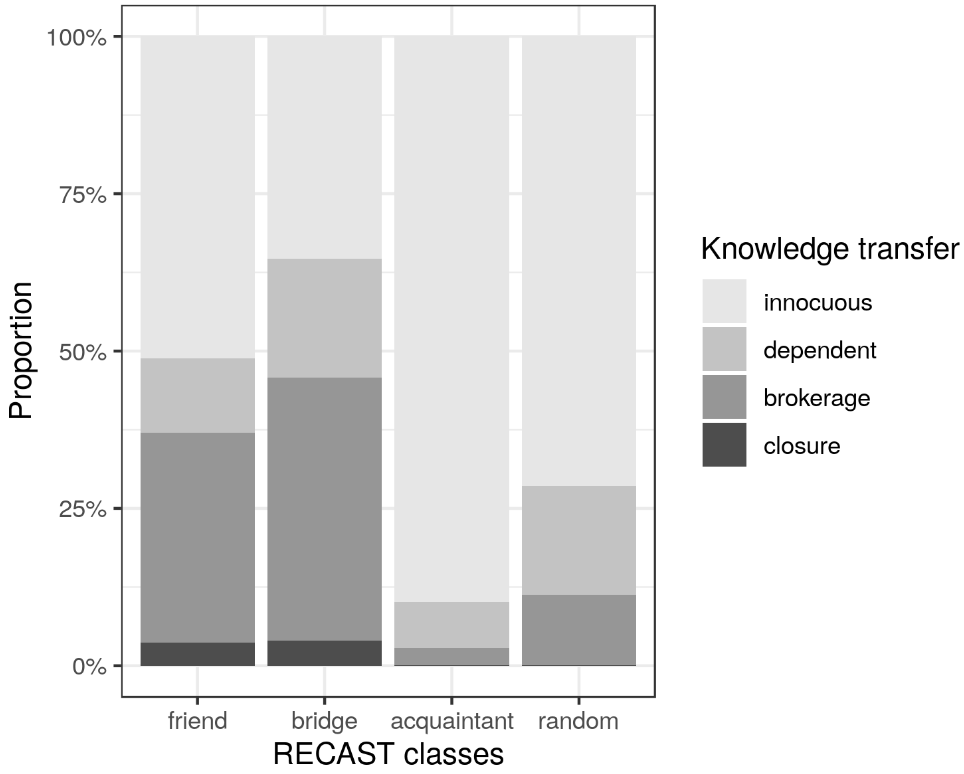


Figure 6.17: Knowledge transfer classification versus the RECAST classification.

6.3.3 Knowledge Transfer versus RECAST

Figure 6.17 shows the intersection of the classes identified by the RECAST algorithm (bars) with those classified in terms of knowledge transfer. As expected, note that the less important classes of our method (*dependent* and *innocuous*) are mostly associated with the less important classes of RECAST (*acquaintant* and *random*). Moreover, associations with bidirectional knowledge transfer (*closure*) are mostly attributed to the *friend* and *bridge* classes. We expected low proportions of associations between the least important class of our method (*innocuous*) with the two most important classes of RECAST, *friend* and *bridge*. However, there is a notable presence of that class associated with these two RECAST classes: *friend* (50%) and *bridge* (35%).

Nevertheless, given that the most important RECAST classes are *friend* and *bridge*, and considering that their correspondence with the classes of our method are, in decreasing order, *closure*, *brokerage*, *dependent* and *innocuous*, we can analyze the correlation among them. For example, there are more, in proportion, *closure* and *brokerage* classes instances for *friend* and *bridge* than for *acquaintant* and *random*. Therefore, we observe a strong coherence between the social definitions of both methods.

6.4 Summary

In this chapter, we presented the results of characterizing several social contexts by means of our proposed classification methods. We summarize this chapter as follows:

- We characterized different social scenarios as, for example, revealing a contrasting social behavior between the *Theory of Computing* and *Applied Computing* networks, and between the *Buddhism* and *Islam* communities.

- We illustrated the edges classified based on the brokerage and closure concepts in terms of the social-based classification and knowledge transfer one, thus reinforcing such edges as being better positioned in a social structure.

- We compared the differences regarding our two classification methods, as well as we contrasted them with the RECAST algorithm [Vaz de Melo et al., 2015]. In summary, we concluded that our approach provides a new social perspective to contribute to the understanding of social interactions.

In the next chapter, we shall see that the classifications proposed for nodes and edges are associated with their strategic positioning in a social structure given by network properties (e.g., the most important classes are better well-positioned than the non-relevant ones). Thereafter, we perform a sensitivity analysis to check the robustness of our classification methods.

Chapter 7

Experimental Validation

Besides analyzing how our classification methods behave in different scenarios (Chapter 6), we should validate our classification results by using an experimental methodology already established in the literature. In addition, we should statistically assess the sensitivity of our classification methods for dealing with particular time aspects and random scenarios.

For this, we first detail our experimental methodology introduced in Section 5. We follow Newman's approach, which explores the closeness and betweenness centrality metrics to determine the best positioned nodes in an academic coauthorship network [Newman, 2004]. Likewise, we employ the notion of social capital given by the strategic positioning of a particular actor in a social structure to validate the assigned classes based on the following network properties, which have been formally defined in Section 5.2 (see Table 5.3):

- *Degree Centrality.* As interpreted by Srinivas and Velusamy [2015], this metric indicates influential nodes as, for example, a node with an immediate risk of catching a virus or getting some information. Thus, a node with high connectivity is more likely to have early access to knowledge.

- *Closeness Centrality.* Nodes with higher closeness are, by definition, closer (on average) to the other nodes in the network. Then, we expect important classes (*closure* and *brokerage*) to have high values for this metric, since they have better access to knowledge from other nodes (e.g., making an opinion to reach other nodes more quickly).

- *Betweenness Centrality.* As discussed by Newman [2004], nodes with a high degree of betweenness centrality are likely to be influential, since they act as an intermediary for other nodes (e.g., in message-passing scenarios). Thus, as nodes and edges with high betweenness centrality values play crucial roles in the spread of knowledge in social networks [Mahyar et al., 2018], then we expect high values for this metric for important nodes and edges assigned to the *closure* and *brokerage* classes.

- *Clustering Coefficient.* As this metric reveals the fraction of a node's neighbors that are connected to each other (i.e., how complete the neighborhood of a node is) [Srinivas and Velusamy, 2015], we expect low clustering coefficient values for the most important classes (*closure* and *brokerage*), confirming the behavior of connecting different parts of a network.

In addition, we also use the PageRank algorithm [Page et al., 1999] by considering that more important nodes tend to make stronger endorsements due to their connectivity and ties to other important nodes. That is, we also expect *closure* and *brokerage* nodes to have high values for this metric.

For instance, let us look at the RECAST classes according to the centrality betweenness metric. As detailed in Section 2.3, we recall that the RECAST classification assigns to the edges the classes *friend*, *acquaintant*, *bridge* and *random*. In summary, the algorithm aims to separate relationships based on strong social interactions (i.e., *friend* and *bridge*) from merely casual ones (i.e., *acquaintant* and *random*). Based on this, it is expected the betweenness values to be associated with the edges according to the importance of the classes inferred by the RECAST. Thus, *friend* and *bridge* classes should have higher betweenness values than those classified as *acquaintant* and *random*.

Figure 7.1 shows the distribution of the betweenness centrality metric with respect to the RECAST classification. We clearly note that the *bridge* edges have the highest values, in accordance with the brokerage concept. However, the *friend* class, regarded as the most strong one, has very low figures. On the other hand, the *random* edges, surprisingly, have the second highest values. As the crucial task of the RECAST algorithm is to separate social from casual relationships [Vaz de Melo et al., 2015], then we would like to observe the interactions between friends and acquaintances with more importance than those regarded as *random*.

Likewise, for the next experiments, we expect the aforementioned network properties to be associated with the nodes and edges according to our proposed classification method. In this way, nodes and edges classified as *closure* and *brokerage* should have better network properties values than those of the other classes (*innocuous* and *dependent*).

Now, we consider the *Full Network*, which comprises the 24 ACM SIGs communities altogether. As the distributions of the network property values per class did not pass the normality test, we evaluated the statistical significance between each two classes by means of the non-parametric Mann-Whitney-Wilcoxon test and among all classes by means of its

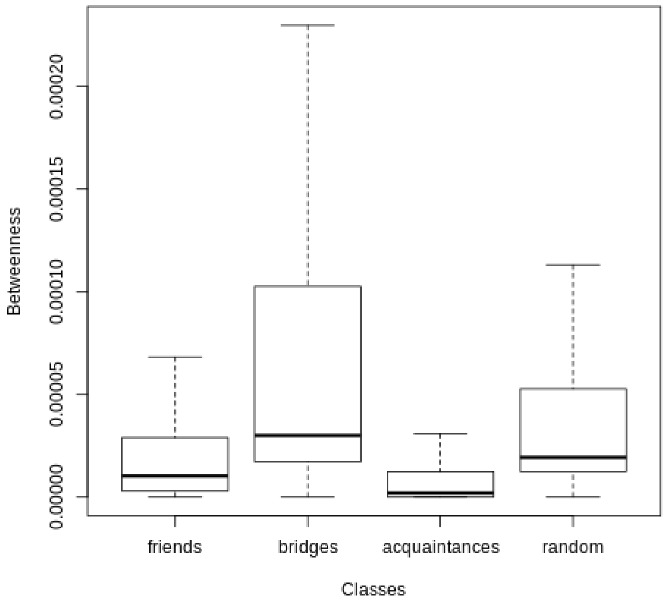


Figure 7.1: Betweenness centrality distributions per RECAST class. Outliers suppressed for better visualization.

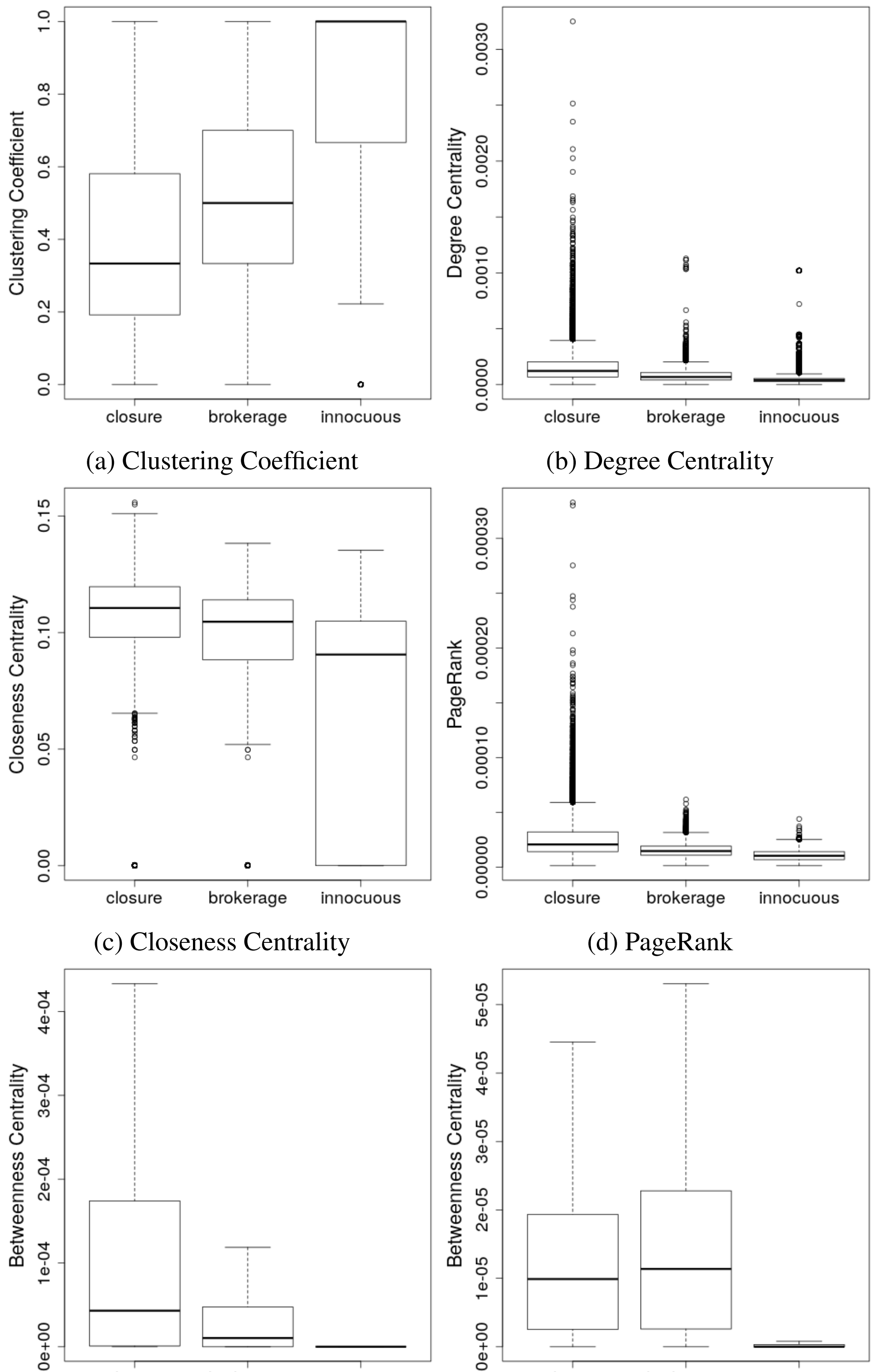
extension given by the Kruskal-Wallis test, as described by Hollander et al. [2013]. All experiments were performed with a significance level of $\alpha = 0.05$.

Next, we discuss the results for the social-based classification (Section 7.1) and the dynamics of knowledge transfer (Section 7.2). Finally, we assess the sensitivity of the proposed classifications (Section 7.3).

7.1 Social-based Classification

In this section, we investigate separately two scenarios by exploring the persistence of the researchers with respect to the relevant words from their articles' titles (tokens) and the social ties with their respective communities.

Relevant Tokens. For the classification of the nodes, Figure 7.2 presents the distribution of the network properties by node class. The clustering coefficient results (Figure 7.2a) emphasizes the characteristic of the nodes assigned as *innocuous* to be very dependent on its neighborhood. This is in stark contrast with the classes *closure* and *brokerage*, which tend to diversify their relationships. The other metrics also validate the *closure* and *brokerage* classes assigned to the nodes better positioned within a social structure. More specifically, they have more social ties (Figure 7.2b), are on average closer to other nodes (Figure 7.2c)



(a) Clustering Coefficient

(b) Degree Centrality

(c) Closeness Centrality

(d) PageRank

(e) Betweenness Centrality (nodes)

(f) Betweenness Centrality (edges)

Figure 7.2: Distribution of network properties per class for nodes (a-e) and edges (f). Outliers were suppressed from (e) and (f) for better visualization.

and topologically more important (Figure 7.2d), and have more information passing through them (Figure 7.2e). In addition, there is a clear class distinction, where *closure* has values greater than *brokerage* and *brokerage* has values greater than *innocuous*. Formally, all distributions are statistically different by means of the Mann-Whitney-Wilcoxon test (among all classes) and by the Kruskal-Wallis test (between each pair) [Hollander et al., 2013].

As for the classification of the edges, Figure 7.2f shows the distribution of the betweenness centrality metric with respect to our classification. We clearly note that the *brokerage* and *closure* classes have more expressive values for this metric. Note that the distributions of *closure* and *brokerage* distinguish less than those reported for nodes, but now the *brokerage* class is slightly superior to the *closure* one in contrast to the classification of the nodes (Figures 7.2a-e). Nonetheless, they are still statistically different according to the Kruskal-Wallis and Mann-Whitney-Wilcoxon tests. Moreover, even though the *innocuous* class accounts for 31.4% of all edges (see Table 6.1), their centrality values are very low.

Finally, comparing such results with the betweenness centrality distributions for RECAST (Figure 7.1), we clearly note that our social-based method showed some coherence in between the social concepts and the network properties, whereas RECAST classifies many structural edges (i.e., those with high network properties), as *random* as well as several edges with low figures as *friend*.

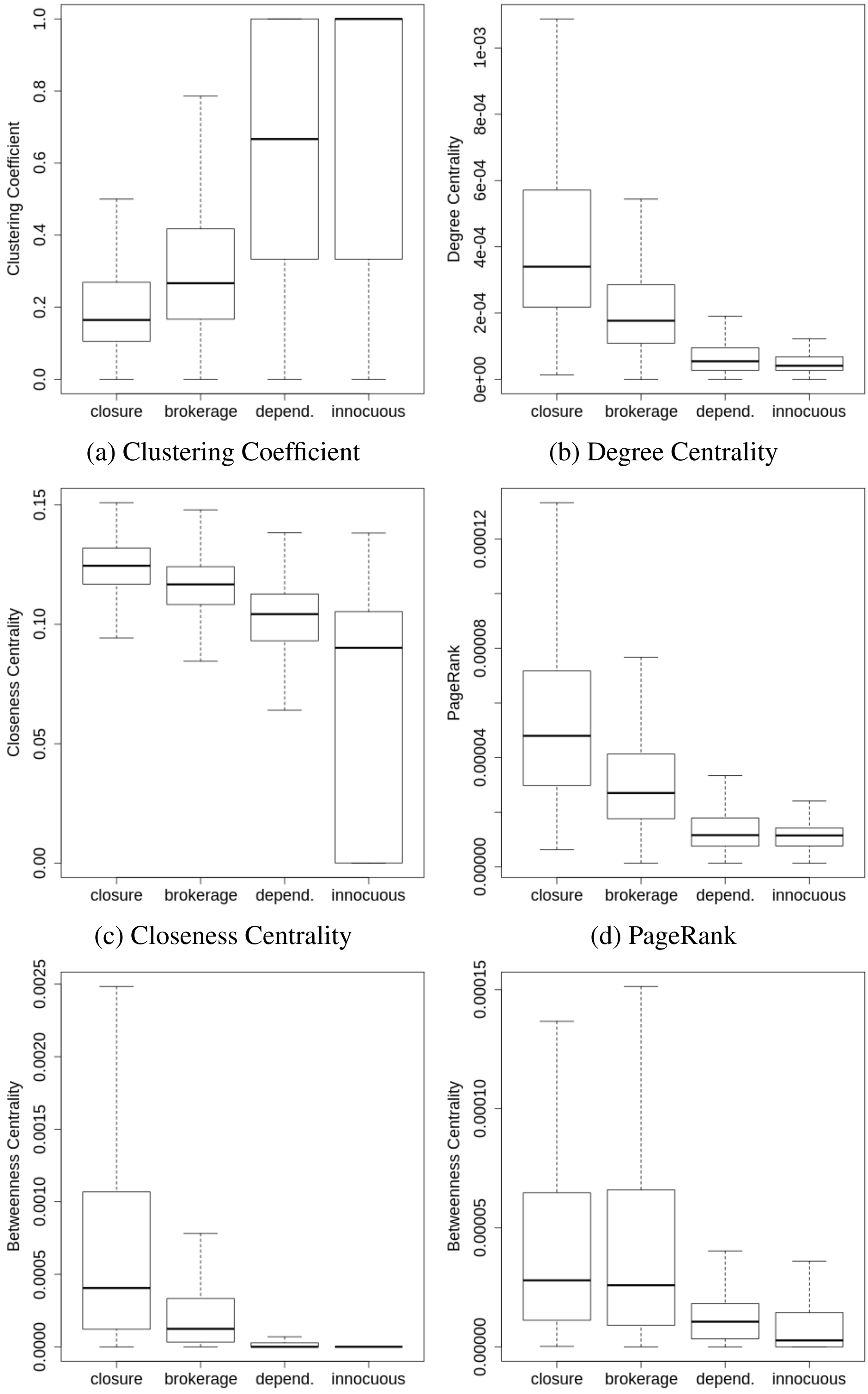
Relevant Community Ties. The results for relevant title words also hold here for community ties. Moreover, the distributions are more well-defined among the classes, i.e., there is a clear distinction among them. For instance, unlike in the relevant title words scenario, the boxes of the *closure* and *brokerage* classes do not overlap in all cases. Again, all classes are statistically valid by means of the Kruskal-Wallis and Mann-Whitney-Wilcoxon tests. As the results from this scenario are very similar to the previous one, we omit the distribution figures and refer the reader to Silva et al. [2018] for more details about them.

Overall, the two scenarios (relevant title words and community ties) confirm the social role of the nodes and the strength of the social characteristics of their interactions.

7.2 Knowledge Transfer Dynamics

For the classification of the nodes, Figures 7.3a-e present the distribution of the network properties by node class. The clustering coefficient results (Figure 7.3a) emphasizes the characteristic of the nodes assigned as *dependent* and *innocuous* to be very dependent on its neighborhood. This is in stark contrast with the classes *closure* and *brokerage*, which tend to diversify their relationships.

The other metrics also validate the *closure* and *brokerage* classes assigned to the nodes



(a) Clustering Coefficient

(b) Degree Centrality

(c) Closeness Centrality

(d) PageRank

(e) Betweenness Centrality (nodes)

(f) Betweenness Centrality (edges)

Figure 7.3: Distribution of network properties per class for nodes (a-e) and edges (f). Outliers suppressed for a better visualization.

better positioned within a social structure. More specifically, they have more social ties (Figure 7.3b), are on average closer to other nodes (Figure 7.3c), are topologically more important (Figure 7.3d) and have more information passing through them (Figure 7.3e). In addition, there is clearly a class distinction, where *closure* is superior to *brokerage*, *brokerage* is superior to *dependent* and, finally, *dependent* is superior to *innocuous*. Again, all distributions are statistically different by means of the Mann-Whitney-Wilcoxon and Kruskal-Wallis tests.

Regarding the information passing through the edges, Figure 7.3f shows the distribution of the betweenness centrality metric with respect to the knowledge-transfer classification. We clearly note that the *closure* and *brokerage* classes have more expressive values for this metric. In addition, the *dependent* class values are also higher than those of the *innocuous* edges. Although the distributions of *closure* and *brokerage* distinguish less than those reported for nodes (Figures 7.3a-e), they are still statistically different according to the Kruskal-Wallis and Mann-Whitney-Wilcoxon tests. Comparing with the betweenness centrality distributions for RECAST (Figure 7.1) and our social-based method (Figure 7.2f), we highlighted that RECAST fails to separate social from casual relationships, whereas our social-based method showed some coherence between its social definition and the network properties.

7.3 Sensitivity Analysis

Since we are dealing with temporal attributed networks, the relevant attributes and time aspects must be properly analyzed regarding the effectiveness of our classification method. Next, we stress these issues in terms of: (i) the discriminatory power of their assigned attributes; (ii) the existence time of the nodes in a network; and (iii) the random node-attributes associations.

Discriminatory power of attributes. In order to measure the strength of social interactions, Algorithm 1 ensures the function Γ containing the sets of all statistically relevant attributes for each node (Section 3.2). In fact, if an attribute is associated with a node several times, then we can infer its importance.

However, a specific statistical treatment can be added to this process in order to exclude attributes that, even if randomly distributed, were erroneously considered as relevant ones. This additional statistical step consists in making the function Φ, which associates each edge e with a specific set of attributes, a random association Φ'. Then, we get Γ from different Φ' instances to measure the probability that each attribute has been erroneously classified as being *relevant*. Finally, we exclude such attributes that were considered as relevant with probability significantly higher than the level of significance α. In other words, we filter from our input the attributes that can interfere in the process of identifying the relevant ones. Even removing some of the data, we expect the proposed method to be robust enough to properly classify nodes and edges.

As a result, both configurations (without the exclusion step and with the step of excluding attributes that are not statistically valid when randomly distributed) are statistically equivalent by means of the distribution of network properties by classes. Precisely, we have noted a better distinction among the classes when such non-discriminatory attributes are excluded. In practice, this step eliminates natural evolution information of the network and, therefore, is not part of the classification process.

Existence time of the nodes. This sensitivity test consists in investigating the robustness of our approach to differentiate nodes with similar existence times. For this, we divided the nodes into the following annual time intervals: $[1,5)$, $[5,10)$, $[10,15)$ e $[15,\infty)$.

Regarding the social-based method, it was able to distinguish the distributions of all network metrics by classes for all time intervals in terms of the Kruskal-Wallis test. For the time interval $[1,5)$, the Mann-Whitney-Wilcoxon test did not differentiate the distributions between the classes *closure* and *brokerage* for the metrics *betweenness centrality* (relevant title words), *clustering coefficient* (relevant title words and community ties) and *degree centrality* (community ties). Considering the knowledge-transfer scenario, the same results hold, where again the Mann-Whitney-Wilcoxon test did not differentiate the distributions between the classes *closure* and *brokerage* for the interval $[1,5)$. Even so, the values for the classes *closure* and *brokerage* are higher than those for the *dependent* and *innocuous* ones (the values for the class *dependent* are also higher than those for the *innocuous* one).

Random node-attribute associations. To check the validity of the knowledge transfer classification on a network, we evaluate the classification robustness when the set of relevant attributes are randomly associated with the nodes (i.e., shuffling the sets $\Gamma(u)$), but keeping the same social structure (i.e., preserving the structure of nodes and edges in G). In this random scenario, our proposed classification method is expected to be robust enough not to misclassify nodes and edges as being *closure* and *brokerage*. In other words, assuming that the knowledge transfer passes through the network according to strong social ties, then we expect that such random associations will result in non-relevant classes for the nodes and edges (i.e., *innocuous*).

Tables 7.1 and 7.2 show the averages and standard deviations of the edge and node classifications, respectively, for 10 random runs. As expected, there is a rarity of nodes and edges classified as either *closure* or *brokerage*. In addition, the standard deviations are very low, thus indicating the coherence of our method. Despite that, the ideal result would be that

Table 7.1: Random Classification of Edges.

		Random	
	Real	Avg.	Std. Dev.
Closure	8.16%	0.14%	0.01%
Brokerage	9.53%	0.09%	0.01%
Dependent	31.70%	6.17%	0.30%
Innocuous	50.60%	93.61%	0.32%

Table 7.2: Random Classification of Nodes.

		Random	
	Real	Avg.	Std. Dev.
Closure	11.56%	0.55%	0.04%
Bokerage	4.55%	0.18%	0.01%
Dependent	48.95%	14.53%	0.34%
Innocuous	34.94%	84.75%	0.34%

for which all nodes and edges were classified as *innocuous*, but as we can see 6.17% of the nodes and 14.53% of the edges were classified as *dependent*. Nonetheless, in the previous section we observed that both dependent nodes and dependent edges tend to have very low importance in a social structure.

7.4 Summary

In this chapter, we validated the assigned classes to edges and nodes in terms of their importance in a social structure. For this, we based on the premise that our social-based classes determine how nodes and edges tend to be positioned in a network [Newman, 2004]. We can summarize our main contributions in this chapter as follows:

- We statistically validated the assigned classes according to network properties, thus agreeing with their expected social values.

- We stressed the proposed method in terms of its robustness for dealing with the existence time of the nodes in a network and the discriminative power of their assigned attributes.

In the next chapter, we apply our proposed model to identify influential nodes based on social concepts.

Chapter 8

Application: A Social-based Ranking

In this chapter, we address our third research goal, which regards the application of our proposed method to a real-world problem: a social-based ranking. For this, we deal with the task of identifying influential nodes in a coauthorship network based on their social influence. First, we present a semi-supervised strategy that combines other social approaches (Section 8.1). Then, we propose an unsupervised strategy based entirely on our proposed model (Section 8.2). Finally, we present the experimental methodology (Section 8.3) and the ranking results (Section 8.4).

8.1 Supervised Ranking Strategy

To train our supervised strategy, we assume that a researcher is more likely to be influential in a network if (i) the greater her participation in the network is, (ii) the better positioned in a social structure she is, (iii) the greater her social ties with the propagated relevant attributes are and (iv) the greater her dynamic diffusion behavior is [Silva et al., 2019]. Thus, based on these considerations, the ranking function of a researcher *n* is given by Equation 8.1:

$$SocialRank(n) = \alpha_1 \frac{\#publ(n)}{\underset{x \in V}{\arg\max}\ \#publ(x)} + \alpha_2 \frac{social(n)}{\underset{x \in V}{\arg\max}\ social(x)} + \alpha_3 \frac{|\Gamma(n)|}{\underset{x \in V}{\arg\max}\ |\Gamma(x)|} + \alpha_4 \frac{\sum_{(i,j) \in \mathscr{E}_d} Origin(i,n)}{\underset{x \in V}{\arg\max}\ \sum_{(i,j) \in \mathscr{E}_d} Origin(i,x)} \quad (8.1)$$

Each one of the above assumptions is represented by the terms of the Equation 8.1, which are, respectively:

- **Participation.** The function *#publ* returns the number of publications of a researcher

n. Thus, the greater the participation of a researcher in a network is, the stronger her structural ties are likely to be.

- **Social Structure.** The function *social* is expressed by

$$social(n) = \sum_{e \in \bigcup_{i=1}^{t} \mathscr{E}_i | n \in \text{ nodes(e)}} w(e)$$

and returns the aggregated weight of the edges. Here, we consider the social-based classification (Section 4.1) and divide the edges classified as *brokerage* into three as being *strong*, *regular* and *weak* bridges. Briefly, a *strong bridge* is established by two nodes that have total control over the information passing by the edge; *regular bridge* when only one node has full control over information, but the other node has control over other subjects; and a *weak bridge* when only one of the nodes has no information to be exchanged regardless of what is passing through the network. Then, the value of the *weight* function is given by

$$w(e) = \begin{cases} 3, & \text{if the edge } e \text{ is } \textit{strong bridge or regular bridge} \\ 2, & \text{else if the edge } e \text{ is } \textit{closure} \\ 1, & \text{else if the edge } e \text{ is } \textit{weak bridge} \\ 0, & \text{otherwise.} \end{cases}$$

In other words, the weights above have been defined considering the distribution of the social meaning of the relationship classes according to the betweenness centrality metric (see Figure 7.2f). This means that the edge weighting privileges the *brokerage* relationships established by strong ties, then the *closure* relationships, and finally the *brokerage* relationships in terms of weak ties. *Innocuous* relationships do not contribute to the score.

- **Relevant Attributes.** The function $|\Gamma|$ returns the number of relevant attributes. In this way, the greater is the potential for acquiring knowledge, the greater are the nodes' social ties with its set of relevant attributes, which means that their reputation is likely to be high.

- **Knowledge Transfer.** The function $Origin(i,n)$ returns 1 if the origin node i and the node n are the same, or 0 otherwise. Thus, this final term quantifies the number of out-edges in the directed knowledge-transfer graph D. Indeed, it measures the spreading of its relevant attributes to its neighborhood. Thus, the greater is the number of social ties in the directed graph, the greater is the effectiveness of its influence.

As each function term is normalized by the maximum individual score, the coefficients α_i scale the importance of each assumption by setting relative weights.

8.2 Unsupervised Ranking Strategy

The previous supervised strategy has two major drawbacks. First, only two assumptions (social structure and dynamic diffusion behavior) are based on our proposed model. Second, it associates coefficients to scale the importance of each assumption by setting relative weights, thus requiring a training phase to estimate the weights for each term. Thus, to overcome both issues, here we propose an unsupervised ranking strategy based entirely on our classification model [Silva et al., 2020] presented in Chapter 3.

As discussed by Burt [2005], the social concepts of closure (relationships within a group) and brokerage (social ties beyond a group) determine how people are connected in a social structure. In this way, when such players are well-positioned in a network (i.e., their closure and brokerage figures are high), they take advantage of early access to the information circulating around them and of a wider diversity of knowledge.

Based on that, how can we measure the node importance by means of closure and brokerage concepts? Burt [2005] addresses this question as a *structural autonomy* that informs when people are tightly connected to one another with extensive bridge ties beyond them. Specifically, maximum performance (e.g., innovation and productivity) is achieved with high closure and brokerage values (e.g., trust and cooperation), whereas minimum performance occurs when these values are lower (e.g., distrust and indifference).

Considering the above discussion, here we propose a knowledge-transfer ranking by exploring the strong ties between nodes and their relevant attributes (the closure effect), as well as the dissemination of the knowledge acquired by the nodes in the network (the brokerage effect). The former is defined in terms of $|\Gamma|$, which returns the number of relevant attributes. In this way, the greater is the potential for acquiring knowledge, the greater are the nodes' social ties with its set of relevant attributes, which means that their reputation is likely to be high. The latter is defined as the degree centrality in the directed knowledge-transfer graph D. Indeed, it measures the spreading of its relevant attributes to its neighborhood. Thus, the greater is the number of social ties in the directed graph, the greater is the effectiveness of its influence.

Formally, the ranking function of a node n is given by Equation 8.2. Note that each term is normalized by the maximum individual score and the ranking applies the geometric mean, since a node tends to be more important if she presents expressive values for both closure and brokerage (having little relevance otherwise).

$$\text{KT-Rank(n)} = \sqrt{\frac{|\Gamma(n)|}{\underset{x \in V}{\arg\max}\ |\Gamma(x)|} \times \frac{\sum_{j \in V_d} a_{ij}}{\underset{x \in V_d}{\arg\max}\ \sum_{y \in V_d} a_{xy}}}, \tag{8.2}$$

where $a_{ij} = \begin{cases} 1, & \text{if the edge } (i,j) \in V_d \\ 0, & \text{otherwise.} \end{cases}$

8.3 Ranking Validation

In academic social networks, the problem of ranking researchers is critical for a broad range of real-world problems (e.g., funding purposes) and there is no consensus on the ideal metrics for a fair decision process [Lima et al., 2013]. To overcome this, we validate our ranking strategies in terms of its results by first classifying the set of distinguished researchers who received at least one ACM award for their contributions or innovations to their specific research communities. Then, we built a ground-truth composed of 544 (out of 79,684) of such influential researchers and, following similar experimental protocols [Lü et al., 2011; Newman, 2004], applied traditional network properties to rank them. Besides considering betweenness centrality and Page-Rank, we also tested the ranking generated by degree centrality and closeness centrality, but we omitted their results because they were inferior to those of the network metrics. Also, we tested an alternative formulation for Equation 8.2 that considered the closure expressed by means of the clustering coefficient centrality and the brokerage expressed by the betweenness centrality, but it achieved inferior results.

As our ranking proposal relays on the network structure, we do not use citation-based approaches (i.e., expensive methods requiring published material content). Instead, we expanded our experimental evaluation by comparing our method with three different social approaches:

(i) **Neighborhood-based Centrality.** Considering the importance of the centrality of a node in a social structure, the H-index of a network node is defined as the maximum number h such that it has at least h neighbors with degrees greater than h [Lü et al., 2016]. Thus, this index captures the node's spreading importance, since the spread process is likely to cease if the node's neighbors have low degrees.

(ii) **Social Hierarchical Structure.** Based on the premise that nodes that are connected to other ones in lower social hierarchies cause them some kind of *social agony*, Gupte et al. [2011] proposed a metric (here called *SocialAgony*) that finds the best ranking that provokes the least agony.

(iii) **PageRank-Like Algorithm.** LeaderRank is a ranking algorithm based on random walks on a network, in which a node's score is given by the fraction of time the random walker spends on that node [Lü et al., 2011]. Thus, this index explores the leadership topology in order to identify influential nodes.

In order to evaluate our ranking and the baselines in light of the ground-truth, we use the discounted cumulative gain (DCG) metric [Järvelin and Kekäläinen, 2002], which measures the quality of a ranking by applying a log-based discount factor that privileges high relevant ranked nodes. Formally, the DCG at a rank position k can be defined as

$$\mathrm{DCG@k} = \sum_{i=1}^{k} \frac{2^{g_i} - 1}{log_2(i+1)}, \tag{8.3}$$

where g_i denotes a binary relevance associated with a node ranked at the i-th position, indicating 1 if the node represents a winner of an ACM award and 0, otherwise.

8.4 Results

Our task here is to retrieve in the first 544 positions of a list of 79,684 nodes (Computer Science researchers), those that are unquestionably considered as outstanding for their contributions to the area and have been recognized for that by receiving an ACM award or being named an ACM distinguished member (ACM Fellow or Distinguished Scientist)[1].

Before, we need to set the weights α_i of the supervised ranking (Equation 8.1). To estimate such parameters, we divided our dataset into four folds and applied the parameter tuning by grid search (values varying from 0 up to 1 by incrementing 0.01 at each iteration). By definition, all four terms of the equation are individually normalized and the average of the best configurations obtained was $\alpha_1 = 0.5\alpha_2 = \alpha_3 = 0.5\alpha_4$. This means that the number of publications (α_1) and the number of relevant terms (α_3) are twice as important as the score derived from the social capital aspect (α_2) and the knowledge transfer one (α_4). Nevertheless, all assumptions are essential to the effectiveness of our proposed ranking.

Figure 8.1 compares our supervised and unsupervised strategies, here called Social-Rank and KT-Rank respectively, with the number of publications (#publ), the network properties (betweenness centrality and PageRank), the aforementioned baselines (H-index, LeaderRank and Social-Agony) and our previously proposed social-based rank (SocialRank) in

[1] `https://awards.acm.org/advanced-member-grades.`

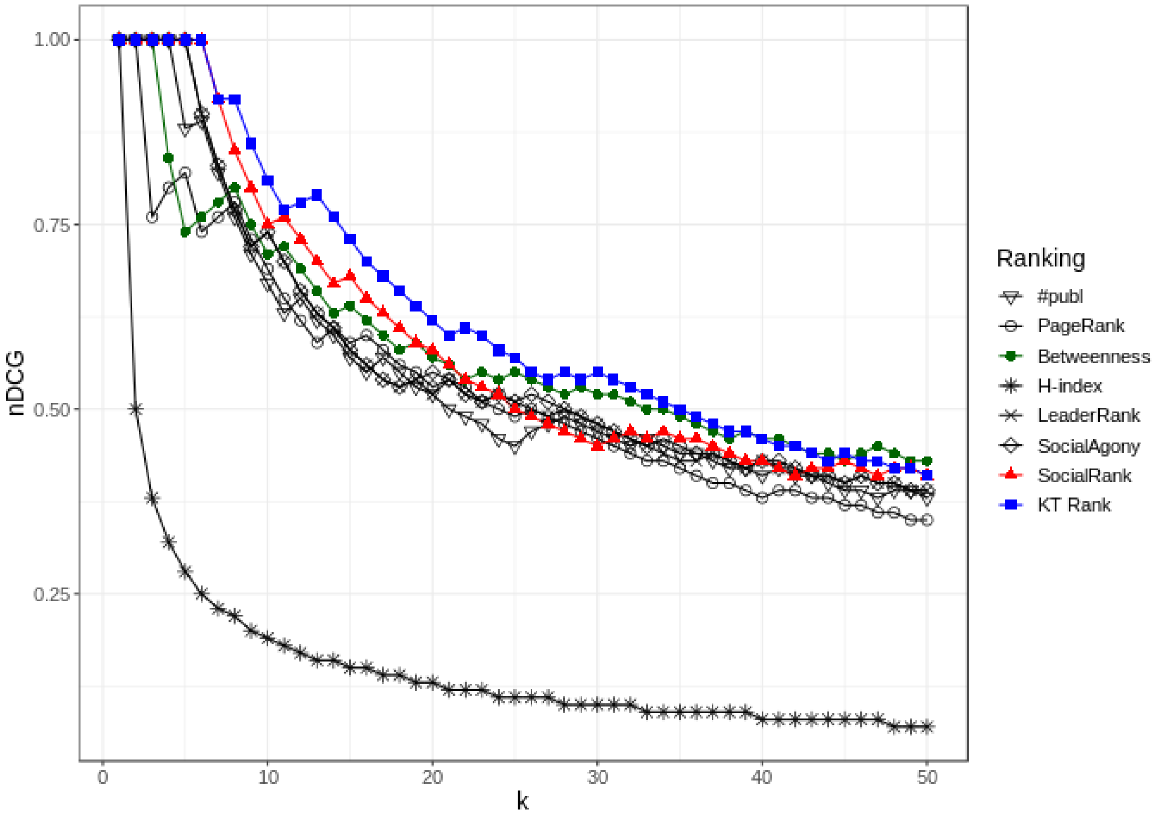


Figure 8.1: Comparison of the SocialRank and KT-Rank with the baselines according to the nDCG.

terms of the attained nDCG@k, with $1 \leq k \leq 50$, where the best metric is the one with the highest nDCG.

The results show that our KT-Rank outperforms or performs equally to the other rankings from the first position until the 40^{th} one. Then, the betweenness centrality outperforms or performs equally to the SocialRank and KT-Rank. As we are interested in the top positions, our strategies report the best results, whereas the other ranking metrics have slight oscillations in the first positions. Besides that, the time complexity of the betweenness centrality on an unweighted network is $O(|V||E|)$ and on a weighted one is $O(|V||E| + |V|^2 \log |V|)$ [Brandes, 2001], which can be prohibitive to be applied to large social networks. Finally, we clearly note an improvement in the ranking based entirely on the concepts of brokerage and closure (KT-Rank) over the supervised SocialRank.

In order to illustrate such results, Table 8.1 lists the top twenty most influential researchers according to the best ranking strategy KT-Rank ($k = 20$), showing their relative rank positions according to other social approaches (left) and network properties (right). Besides highlighting the ACM awardees, we also indicate the ACM distinguished members (ACM Fellows and ACM Distinguished Scientists). As we can see, several well known

Table 8.1: Top 20 researchers according to KT-Rank and their relative rank positions on the baselines: bold indicates ACM awardees (by innovations or contributions), ACM fellows (⋆) and ACM distinguished scientists (*).

Social Approaches						Network Properties				
KT-Rank	Researcher	*SocialRank*	*LeaderRank*	*SocialAgony*	*H-index*	*#publ*	*Degree*	*PageRank*	*Closeness*	*Betweenness*
1	**Jiawei Han**⋆	1	1	1	77	2	1	1	1	2
2	**Christos Faloutsos**⋆	3	3	3	81	4	3	4	2	1
3	**Tat-Seng Chua**	5	5	5	80	6	309	174	844	506
4	**Philip S. Yu**⋆	2	2	2	84	1	2	2	3	3
5	W. **Bruce Croft**⋆	4	28	26	399	3	21	5	192	19
6	**Scott Shenker**⋆	7	4	4	83	12	4	7	4	6
7	Hui Xiong*	8	9	8	204	20	9	11	90	36
8	**Jian Pei**⋆	15	10	10	86	28	7	8	18	15
9	Qi Tian	12	8	9	79	13	14367	13262	10658	9719
10	Maarten de Rijke	11	16	16	299	8	12	9	159	16
11	Xian-Sheng Hua*	9	36	36	307	9	2487	1691	5124	2635
12	**Shih-Fu Chang**⋆	30	19	20	210	55	73719	73544	73544	73544
13	**Susan T. Dumais**⋆	31	172	172	347	34	159	104	21	7
14	Alberto L. S.-Vincentelli⋆	13	6	6	340	7	5	3	399	4
15	Jason Cong⋆	26	30	29	438	29	25	13	196	9
16	Wei-Ying Ma*	14	13	13	82	10	31	57	140	102
17	Ryen W. White	34	49	55	230	15	45	34	38	18
18	Carl Gutwin*	23	230	234	805	36	193	70	2441	199
19	Donald F. Towsley⋆	29	75	73	679	44	85	33	220	112
20	Jieping Ye	33	17	17	397	43	14	19	379	116

Computer Science researchers, including ACM awardees, appear in the top 20 positions. The only exception in this list is *Qi Tian*, whose name is listed as ambiguous in DBLP[2].

By comparing the aforementioned ranking based on social approaches (Table 8.1, left), we notice several similar relative rank positions in the KT-Rank, SocialRank, LeaderRank and SocialAgony lists. In particular, our proposals are quite similar (range 1 to 33), and the relative rank positions showed by LeaderRank and SocialAgony are almost identical. One exception is the H-index, whose positions significantly differ from the other social approaches in the range 77 to 805, thus confirming its worst performance shown in Figure 8.1.

Regarding network properties (Table 8.1, right), only the number of publications is consistent with the social approaches (range 1 to 55), reinforcing the importance of a researcher being active on the network. The other properties sometimes agree with the social approaches (for instance, the cases of *Jiawei Han*, *Christos Faloutsos* and *Philip S. Yu*), but fail dramatically in some important cases such as those of *Tat-Seng Chua* and *Shih-Fu Chang*.

[2] https://dblp.uni-trier.de/pers/hd/t/Tian:Qi

8.5 Summary

In this chapter, we showed how to apply our proposed graph model and classification method, thus addressing our third research goal. In this regard, we proposed a strategy to rank nodes based on their social influence. Our two strategies (supervised and unsupervised) outperformed the traditional network metrics and other social-based algorithms (neighborhood-based, social hierarchical structure and PageRank-like algorithms).

In conclusion, the results demonstrate that our approaches differ slightly from the social baselines (except H-index) and, in some cases, in terms of network properties, thus providing new social perspectives for a more robust evaluation of the importance of nodes in a social structure.

Chapter 9

Conclusions and Future Work

This thesis addressed the problem of characterizing actors and their social interactions in dynamic social networks. As additional complexity, the general scenario comprised parallel relationships over time, as well as attributes associated with each social interaction. As a scientific contribution, we reinforced the importance of the network theory paradigm for understanding the complexity that involves real world actors and their relationships [Barabási, 2009; Watts, 2004]. More specifically, we emphasized the social roles of actors in dynamic attributed networks in order to determine the social meaning of their multiple dynamic interactions. For this, we relied on Burt's definition of two social concepts [Burt, 2005], *closure* as the ability of aggregating individuals with similar social patterns, and *brokerage* as the ability of creating bridges with diversified social patterns.

Next, we summarize our main findings (Section 9.1) and provide some considerations for future work (Section 9.2).

9.1 Summary of Results

As discussed in Section 1.2, our main goal in this thesis was to provide a novel understanding about real world complex systems. Next, we present our findings according to the following three associated research goals.

RG1 - Modeling Dynamic Attributed Interactions. We explored node-attributes associations to model social interactions over time based on social concepts. Then, we defined three graph representations to deal with dynamic interactions, their associated attributes and how knowledge transfer flows across the network. More specifically, our strategy explored the persistence (i.e., long-lasting associations) of the nodes with specific attributes to determine the social importance underlying the dynamicity of the interactions between the nodes over time. We showed that these models provide a new perspective for characterizing social networks (Chapter 6), as well as to be used in practice (Chapter 8).

RG2 - Node and edge classifications. Regarding our second goal, we proposed a new method to classify the dynamic interactions in attributed social networks. For this, we relied on Burt's definition of *closure* and *brokerage* [Burt, 2005], thus dealing with strong and weak diversified interactions between the nodes along the time, respectively. Then, we based our strategy on the persistence of the attributes in the nodes' interaction history.

Considering the social-based classification and the dynamics of knowledge transfer:

- We defined three types of node behavior in a social structure, thus characterizing as (i) *closure* those that have authority on certain attributes and, therefore, have a great potential to diffuse knowledge from their domain; (ii) *brokerage* those that have a weak association with their attributes; and (iii) *innocuous* those that have occasional presence in the network.

- We classified the edges based on the same three social concepts. More specifically, such classes emphasize the strength of relationships as strong ties (*closure*), weak ties (*brokerage*) and non-relevant information passing through the edge (*innocuous*).

- We explored the dynamics involving individuals in terms of how their attributes are transferred across the network. Then, we proposed a knowledge-transfer method that captures the social tie of individuals and their associated attributes over time by exploring how the attributes pass through the network in order to assign social classes to the nodes and edges.

We can summarize our experimental analysis as follows:

- We dealt with different social scenarios (academic coauthorship networks and Q&A communities), which provided multiple social facets. Specifically in the academic context, two scenarios were characterized by exploring the persistence of relevant tokens extracted from the article titles and the social ties with academic communities. Additionally, in the Q&A scenario we characterized 28 communities divided into six categories from Stack Exchange. In summary:

  - Our social-based classification method characterized the social role of the nodes and the strength of the social meaning of their multiple interactions. For instance, members of the STOC (Theory of Computing) community have a tendency to show more competence in specific topics related to computation theory, thus the higher number of nodes of the class *closure*. On the other hand, SAC (Applied

  Computing) is a community mainly focused on applied computing, thus covering a wide range of topics, which justifies the high number of *innocuous* nodes.

  - Regarding the knowledge-transfer perspective, the communities from the *Technology* category have high values of the *dependent* and *innocuous* classes, which is a behavior similar to that of the academic communities like ISSAC (Symbolic and Algebraic Computation), STOC (Theory of Computing) and PODC (Principles of Distributed Computing). Moreover, we presented contrasting social behaviors by comparing, for example, the *Theory of Computing* and *Applied Computing* networks, and the *Buddhism* and *Islam* communities.

- Based on Newman's experimental methodology [Newman, 2004], we statistically validated the assigned classes according to network properties, thus agreeing with their expected social meaning. Overall, our proposed method agrees with their expected social behavior, e.g., the most important classes (i.e., based on the *closure* and *brokerage* concepts) are better well-positioned than the other ones in the social structure. Also, our proposed method was stressed in terms of its robustness for dealing with the existence time of the nodes in a network and the discriminative power of their assigned attributes.

RG3 - Application. Finally, we showed how to apply our proposed graph model and classification method. Specifically, we proposed two social strategies for identifying influential nodes. For this, we focused on exploring the importance of nodes and their relationships according to the classes assigned to them, thus weighting how better they are positioned in a social structure. As a result, our strategy outperformed traditional network metrics and other social-based approaches (neighborhood-based, social hierarchical structure and PageRank-like algorithms).

9.2 Future Work

Given that our study presents a new perspective for analyzing edges and nodes based on social concepts, the next step is to investigate to what extent we can combine our strategies with other works proposed in the literature. In this regard, although the RECAST algorithm [Vaz de Melo et al., 2015] has not shown robust results to explicitly separate social from random interactions by means of a good positioning in a social structure (see Figure 7.1), it was extremely adept at identifying relationships underlying acquaintances. Thus, we can now use such an outcome to better quantify which relationships are likely to be extremely strong, as well as to discard those that may not have had positive indicators. Alternatively, we can use the RECAST to preprocess the data to, for example, filter out clearly sporadic social interactions. In fact, as a first idea, we intend to explore concepts of the RECAST algorithm, specially for detecting relationships with *no social value*. More precisely, we intend to improve this strategy by separating strong social relationships from merely casual interactions.

In another perspective, we can incorporate the social bond between the actors and their communities into our problem. We have already tested the social tie *actor-community* by means of attributes and the results have shown that there is a structural strength between them over time (i.e., better network properties). But now, we can formalize a community-based strategy in order to identify a set of nodes (social circles) that have some structuring power in certain communities. By grouping similar nodes instead of considering them individually, we can provide a better assessment when establishing the strength of the relationships. For instance, bridges between groups can be considered more important than those connecting two individual nodes in the same group. We also intend to apply our social concepts to the problem of community detection by weighting the edges in terms of their assigned social classes. In this way, we plan to investigate the persistence of the nodes with their neighborhood to better assess each associated social class [Silva et al., 2020].

Finally, considering that nodes positioned in different parts of a network can have similar structural roles within their local network topology [Donnat et al., 2018], we can learn such structural representations to better assess the social role of the nodes and the meaning of their interactions. Particularly, we can investigate how to consider small patterns (i.e., motifs [Paranjape et al., 2017]) involving the nodes and their associated attributes. In conclusion, discussing such issues corroborates to provide insights for characterizing complex networks.

Bibliography

Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211–230.

Aggarwal, C., He, G., and Zhao, P. (2016). Edge classification in networks. In *Proceedings of the International Conference on Data Engineering*, pages 1038–1049, Helsinki, Finland. IEEE Computer Society.

Aggarwal, C. C., Li, Y., Philip, S. Y., and Zhao, Y. (2017). On Edge Classification in Networks with Structure and Content. In *Proceedings of the International Conference on Data Engineering*, pages 187–190, San Diego, CA, USA. IEEE Computer Society.

Alves, B. L., Benevenuto, F., and Laender, A. H. F. (2013). The Role of Research Leaders on the Evolution of Scientific Communities. In *Proceedings of the International Conference on World Wide Web (Companion Volume)*, pages 649–656, Rio de Janeiro, Brazil. International World Wide Web Conferences Steering Committee / ACM.

Aral, S. (2016). The future of weak ties. *American Journal of Sociology*, 121(6):1931–1939.

Asiri, S. and Miri, A. (2016). An IoT trust and reputation model based on recommender systems. In *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, pages 561–568, Auckland, New Zealand. IEEE.

Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. *Science*, 325(5939):412–413.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4):590–614.

Barrat, A., Fernandez, B., Lin, K. K., and Young, L.-S. (2013). Modeling temporal networks using random itineraries. *Physical review letters*, 110(15):158702.

Benevenuto, F., Laender, A. H. F., and Alves, B. L. (2015). How Connected are the ACM SIG Communities? *SIGMOD Record*, 44(4):57–63.

Beutel, A., Xu, W., Guruswami, V., Palow, C., and Faloutsos, C. (2013). Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the International Conference on World Wide Web*, pages 119–130, Rio de Janeiro, Brazil. International World Wide Web Conferences Steering Committee / ACM.

Bhagat, S., Cormode, G., and Muthukrishnan, S. (2011). Node classification in social networks. In *Social network data analytics*, pages 115–148. Springer.

Bhat, S. Y. and Abulaish, M. (2013). Community-based features for identifying spammers in Online Social Networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 100–107, Niagara, ON, Canada. ACM.

Braha, D. and Bar-Yam, Y. (2009). Time-Dependent Complex Networks: Dynamic Centrality, Dynamic Motifs, and Cycles of Social Interactions. In *Adaptive Networks*, pages 39–50. Springer.

Brandão, M. A., de Melo, P. O. V., and Moro, M. M. (2018). STACY: Strength of Ties Automatic-Classifier over the Years. *Journal of Information and Data Management*, 9(1):52–68.

Brandão, M. A. and Moro, M. M. (2015). Analyzing the Strength of Co-authorship Ties with Neighborhood Overlap. In *International Conference on Database and Expert Systems Applications*, pages 527–542, Valencia, Spain. Springer.

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

Brandão, M. A., Vaz de Melo, P. O. S., and Moro, M. M. (2017). Tie Strength Dynamics over Temporal Co-authorship Social Networks. In *Proceedings of the International Conference on Web Intelligence*, pages 306–313, Leipzig, Germany. ACM.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality. *Journal of mathematical sociology*, 25(2):163–177.

Burt, R. S. (2005). *Brokerage and Closure: An Introduction to Social Capital*. Oxford University Press.

Burt, R. S. (2009). *Structural holes: The social structure of competition*. Harvard University Press.

Chaturvedi, S., Iyyer, M., and Daume III, H. (2017). Unsupervised Learning of Evolving Relationships Between Literary Characters. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 3159--3165, San Francisco, California. AAAI Press.

Chen, Y. and Liu, J. (2019). Becoming Gatekeepers Together with Allies: Collaborative Brokerage over Social Networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 81–88, Vancouver, British Columbia, Canada. ACM.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94:S95–S120.

Coleman, J. S. (1994). *Foundations of social theory*. Harvard university press.

Dai, H., Dai, B., and Song, L. (2016). Discriminative Embeddings of Latent Variable Models for Structured Data. In *Proceedings of the International Conference on Machine Learning*, pages 2702–2711, New York City, NY, USA. JMLR.org.

Donnat, C., Zitnik, M., Hallac, D., and Leskovec, J. (2018). Learning Structural Node Embeddings via Diffusion Wavelets. In *Proceedings of the International ACM Conference on Knowledge Discovery and Data Mining*, pages 1320--1329, London, UK. ACM.

Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press. ISBN 0521195330, 9780521195331.

Feng, J., Shi, D., and Luo, X. (2018). An identification method for important nodes based on k-shell and structural hole. *Journal of Complex Networks*, 6(3):342–352.

Freire, V. P. and Figueiredo, D. R. (2011). Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society*, 17(4):255–266.

Gilbert, E. and Karahalios, K. (2009). Predicting Tie Strength With Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220, Boston, MA, USA. ACM.

Gilpin, S., Eliassi-Rad, T., and Davidson, I. (2013). Guided learning for role discovery (glrd): framework, algorithms, and applications. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 113–121, Chicago, IL, USA. ACM.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380.

Guimera, R., Uzzi, B., Spiro, J., and Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702.

Gupte, M., Shankar, P., Li, J., Muthukrishnan, S., and Iftode, L. (2011). Finding Hierarchy in Directed Online Social Networks. In *Proceedings of the International Conference on World Wide Web*, pages 557–566, Hyderabad, India. ACM.

Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Faloutsos, C., and Li, L. (2012). RolX: Structural Role Extraction & Mining in Large Graphs. In *Proceedings of the ACM SIGKDD International conference on Knowledge Discovery and Data Mining*, pages 1231–1239, Beijing, China. ACM.

Henderson, K., Gallagher, B., Li, L., Akoglu, L., Eliassi-Rad, T., Tong, H., and Faloutsos, C. (2011). It's Who You Know: Graph Mining Using Recursive Structural Features. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 663–671, San Diego, CA, USA. ACM.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*, volume 751. John Wiley & Sons.

Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97–125.

Huang, H., Dong, Y., Tang, J., Yang, H., Chawla, N. V., and Fu, X. (2018). Will Triadic Closure Strengthen Ties in Social Networks? *ACM Transactions on Knowledge Discovery from Data*, 12(3):30.

Iglewicz, B. and Hoaglin, D. C. (1993). *How to detect and handle outliers*, volume 16. Asq Press.

Inkpen, A. C. and Tsang, E. W. (2005). Social Capital, Networks, and Knowledge Transfer. *Acad. Manage. Rev.*, 30(1):146–165.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Jiang, J., Shi, P., An, B., Yu, J., and Wang, C. (2017). Measuring the social influences of scientist groups based on multiple types of collaboration relations. *Information Processing & Management*, 53(1):1–20.

Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Data Mining*, pages 219–230, Palo Alto, California, USA. ACM.

Kajdanowicz, T., Kazienko, P., and Doskocz, P. (2010). Label-dependent Feature Extraction in Social Networks for Node Classification. In *International Conference on Social Informatics*, pages 89–102, Laxenburg, Austria. Springer.

Kim, J. (2019). Author-based analysis of conference versus journal publication in computer science. *Journal of the Association for Information Science and Technology*, 70(1):71–82.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.

Laender, A. H. F., de Lucena, C. J. P., Maldonado, J. C., de Souza e Silva, E., and Ziviani, N. (2008). Assessing the Research and Education Quality of the Top Brazilian Computer Science Graduate Programs. *SIGCSE Bull.*, 40(2):135–145.

Leão, J. C., Brandão, M. A., Vaz de Melo, P. O., and Laender, A. H. F. (2018). Who is really in my social circle? Mining Social Relationships to Improve Detection of Real Communities. *Journal of Internet Services and Applications*, 9(20):20:1–20:17.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010). Predicting Positive and Negative Links in Online Social Networks. In *Proceedings of the International Conference on World Wide Web*, pages 641–650, Raleigh, North Carolina, USA. ACM.

Levchuk, G., Roberts, J., and Freeman, J. (2012). Learning and Detecting Patterns in Multi-Attributed Network Data. In *Social Networks and Social Contagion, Papers from the 2012 AAAI Fall Symposium*, Arlington, Virginia, USA. AAAI Press.

Levin, D. Z. and Cross, R. (2004). The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management Science*, 50(11):1477–1490.

Li, E. Y., Liao, C. H., and Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9):1515–1530.

Li, J., Cheng, K., Wu, L., and Liu, H. (2018). Streaming Link Prediction on Dynamic Attributed Networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 369–377, Marina Del Rey, CA, USA. ACM.

Liao, L., He, X., Zhang, H., and Chua, T.-S. (2018). Attributed Social Network Embedding. *IEEE Trans. on Knowl. and Data Eng.*, 30(12):2257–2270.

Lima, H., Silva, T. H., Moro, M. M., Santos, R. L., Meira Jr, W., and Laender, A. H. (2013). Aggregating Productivity Indices for Ranking Researchers across Multiple Areas. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 97–106, Indianapolis, IN, USA. ACM.

Lü, L., Zhang, Y.-C., Yeung, C. H., and Zhou, T. (2011). Leaders in social networks, the delicious case. *PloS one*, 6(6):e21202.

Lü, L., Zhou, T., Zhang, Q.-M., and Stanley, H. E. (2016). The H-index of a network node and its relation to degree and coreness. *Nature communications*, 7:10168.

Mahyar, H., Hasheminezhad, R., Ghalebi, E., Nazemian, A., Grosu, R., Movaghar, A., and Rabiee, H. R. (2018). Identifying central nodes for information flow in social networks using compressive sensing. *Social Network Analysis and Mining*, 8(1):33:1–33:24.

Medo, M., Mariani, M. S., Zeng, A., and Zhang, Y.-C. (2016). Identification and impact of discoverers in online social systems. *Scientific Reports*, 6:34218.

Moreira, C., Calado, P., and Martins, B. (2015). Learning to rank academic experts in the DBLP dataset. *Expert Systems*, 32(4):477–493.

Neshati, M., Fallahnejad, Z., and Beigy, H. (2017). On dynamicity of expert finding in community question answering. *Information Processing & Management*, 53(5):1026–1042.

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

Newman, M. E. (2004). Who is the best connected scientist? A study of scientific coauthorship networks. In *Complex networks*, pages 337–370. Springer.

Newman, M. E. (2010). *Networks: an introduction.* Oxford University Press, Oxford.

Newman, M. E., Barabási, A.-L. E., and Watts, D. J. (2006). *The structure and dynamics of networks.* Princeton University Press.

Orman, G. K., Labatut, V., Plantevit, M., and Boulicaut, J.-F. (2014). A Method for Characterizing Communities in Dynamic Attributed Complex Networks. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 481–484, Beijing, China. IEEE.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Paranjape, A., Benson, A. R., and Leskovec, J. (2017). Motifs in temporal networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 601–610, Cambridge, United Kingdom. ACM.

Paruma-Pabón, O. H., González, F. A., Aponte, J., Camargo, J. E., and Restrepo-Calle, F. (2016). Finding Relationships between Socio-Technical Aspects and Personality Traits by Mining Developer E-mails. In *Proceedings of the International Workshop on Cooperative and Human Aspects of Software Engineering*, pages 8–14, Austin, Texas, USA. IEEE.

Posnett, D., Warburg, E., Devanbu, P., and Filkov, V. (2012). Mining Stack Exchange: Expertise Is Evident from Initial Contributions. In *Proceedings of the International Conference on Social Informatics*, pages 199–204, Lausanne, Switzerland. IEEE.

Rezaei, A., Perozzi, B., and Akoglu, L. (2017). Ties That Bind: Characterizing Classes by Attributes and Social Ties. In *Proceedings of the International Conference on World Wide Web (Companion Volume)*, pages 973–981, Perth, Australia. International World Wide Web Conferences Steering Committee.

Sanz-Cruzado, J. and Castells, P. (2018). Enhancing Structural Diversity in Social Networks by Recommending Weak Ties. In *Proceedings of the ACM Conference on Recommender Systems*, pages 233–241, Vancouver, British Columbia, Canada. ACM.

Shah, C. and Kitzie, V. (2012). Social q&a and virtual reference—comparing apples and oranges with the help of experts and users. *Journal of the American Society for Information Science and Technology*, 63(10):2020–2036.

Shah, N., Beutel, A., Hooi, B., Akoglu, L., Gunnemann, S., Makhija, D., Kumar, M., and Faloutsos, C. (2016). EdgeCentric: Anomaly Detection in Edge-Attributed Networks. In *Proceedings of the 16th IEEE International Conference on Data Mining Workshops*, pages 327–334. ISSN .

Shi, X., Adamic, L. A., and Strauss, M. J. (2007). Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 378(1):33–47. ISSN 0378-4371.

Silva, A., Jr., W. M., and Zaki, M. J. (2012). Mining Attribute-structure Correlated Patterns in Large Attributed Graphs. *PVLDB*, 5(5):466–477.

Silva, T. H., Laender, A. H., Davis, C. A., da Silva, A. P. C., and Moro, M. M. (2017). A profile analysis of the top brazilian computer science graduate programs. *Scientometrics*, 113(1):237--255.

Silva, T. H., Laender, A. H., and de Melo, P. O. V. (2019). Characterizing Knowledge-Transfer Relationships in Dynamic Attributed Networks. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 234–241, Vancouver, British Columbia, Canada. ACM.

Silva, T. H., Laender, A. H., and de Melo, P. O. V. (2020). On knowledge-transfer characterization in dynamic attributed networks. *Social Network Analysis and Mining*, 10(1):1–16.

Silva, T. H., Moro, M. M., and Silva, A. P. C. (2015a). Authorship Contribution Dynamics on Publication Venues in Computer Science: an Aggregated Quality Analysis. In *Proceedings of the Annual ACM Symposium on Applied Computing*, pages 1142–1147, Salamanca, Spain. ACM.

Silva, T. H. P. and Laender, A. H. F. (2018). Uma Abordagem para Classificação de Interações Sociais Dinâmicas a partir de seus Atributos. In *Proceedings of the Symposium on Knowledge Discovery, Mining and Learning*, São Paulo, Brazil. IBM Research Brazil.

Silva, T. H. P., Laender, A. H. F., and Vaz de Melo, P. O. S. (2018). Social-Based Classification of Multiple Interactions in Dynamic Attributed Networks. In *Proceedings of the IEEE International Conference on Big Data*, pages 4063–4072, Seattle, WA, USA. IEEE.

Silva, T. H. P., Moro, M. M., Silva, A. P. C., Meira, Jr., W., and Laender, A. H. F. (2014). Community-based endogamy as an influence indicator. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 67–76, London, United Kingdom. IEEE Press.

Silva, T. H. P., Rocha, L. M., Silva, A. P. C., and Moro, M. M. (2015b). 3c-index: Research Contribution across Communities as an Influence Indicator. *Journal of Information and Data Management*, 6(3):192–2015.

Srinivas, A. and Velusamy, R. L. (2015). Identification of influential nodes from social networks based on Enhanced Degree Centrality Measure. In *Proceedings of the IEEE International Advance Computing Conference*, pages 1179–1184, Banglore, India. IEEE.

Srivastava, S., Chaturvedi, S., and Mitchell, T. (2016). Inferring Interpersonal Relations in Narrative Summaries. In *AAAI Conference on Artificial Intelligence*.

Sun, X., Kaur, J., Milojević, S., Flammini, A., and Menczer, F. (2013). Social Dynamics of Science. *Scientific Reports*, 3:1069.

Tang, J. (2017). Computational Models for Social Network Analysis: A Brief Survey. In *Proceedings of the 26th International Conference on World Wide (Web Companion)*, pages 921–925, Perth, Australia. International World Wide Web Conferences Steering Committee.

Trevithick, P. and Clippinger, J. H. (2008). Method and system for characterizing relationships in social networks. US Patent 7,366,759.

Valverde-Rebaza, J. C., Roche, M., Poncelet, P., and de Andrade Lopes, A. (2018). The role of location and social strength for friendship prediction in location-based social networks. *Information Processing & Management*, 54(4):475–489.

Vasilescu, B., Serebrenik, A., Devanbu, P., and Filkov, V. (2014). How social Q&A sites are changing knowledge sharing in open source software communities. In *Proceendings of the ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 342–354, Baltimore, Maryland, USA. ACM.

Vaz de Melo, P. O. S., Viana, A. C., Fiore, M., Jaffrès-Runser, K., Mouël, F. L., Loureiro, A. A. F., Addepalli, L., and Chen, G. (2015). RECAST: Telling Apart Social and Random Relationships in Dynamic Networks. *Perform. Eval.*, 87:19–36.

Wang, W., Yu, S., Bekele, T. M., Kong, X., and Xia, F. (2017). Scientific collaboration patterns vary with scholars' academic ages. *Scientometrics*, 112(1):329–343.

Watts, D. J. (2004). The "New" Science of Networks. *Annual Review of Sociology*, 30:243–270.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440.

Xu, L., Wei, X., Cao, J., and Philip, S. Y. (2017). Multiple Social Role Embedding. In *Proceedings of the IEEE Conference on Data Science and Advanced Analytics*, pages 581–589, Tokyo, Japan. IEEE.

Yang, J. and Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213.

Yang, J., Tao, K., Bozzon, A., and Houben, G.-J. (2014). Sparrows and Owls: Characterisation of Expert Behaviour in StackOverflow. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*, pages 266–277, Aalborg, Denmark. Springer.

Yang, Y. and Xie, G. (2016). Efficient identification of node importance in social networks. *Information Processing & Management*, 52(5):911–922.

Yo, T. and Sasahara, K. (2017). Inference of Personal Attributes from Tweets Using Machine Learning. In *Proceedings of the IEEE International Conference on Big Data*, pages 3168–3174, Boston, MA, USA. IEEE.

Zaki, M. J., Meira Jr, W., and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

Zhang, G., Liu, L., and Wei, F. (2019). Key nodes mining in the inventor-author knowledge diffusion network. *Scientometrics*, 118(3):721–735.