Analysis and Evolution for Online Personal Collaboration Networks

By

Sarra DJEMILI

ETIS Lab, UMR 8051
Cergy Paris University, ENSEA,CNRS

A thesis submitted for the degree of

Doctor of Philosophy in Computer Science

04/12/2020

COMPOSITION OF THE JURY

Pascale Kuntz	Full Professor, Polytech Nantes, LS2N	Reviewer
Arnaud Martin	Full Professor, University of Rennes 1, IRISA	Reviewer
Florence Sèdes	Full Professor, University of Toulouse, IRIT	President
Dimitris Kotzinos	Full Professor, CY Cergy Paris University, ETIS	Thesis Director
Claudia Marinica	Assistant Professor, Polytech Nantes, LS2N	Co-supervisor
Maria Malek	Assistant Professor, CY Cergy Paris University, ETIS	Co-supervisor

This thesis is dedicated to my lovely family.

Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. They are the ultimate role models.

Most importantly, I wish to thank my loving and supportive husband, *Ramzi*, and my wonderful daughters, *Nesrine* and *Nour Yasmine*, who provided unending inspiration.

Acknowledgements

I would like to thank all the people who contributed to the realization and the writing of this thesis. This is the fruit of many years of work, spent in the company of several people who each participated in this final outcome.

The first people to whom I want to express my thanks are my thesis director Dimitris KOTZINOS, and my supervisors Claudia MARINICA and Maria MALEK. I will always be grateful to you for guiding me through these years, during which I was able to share your scientific thought. A big thank you for the exchanges and discussions that allowed me to have the necessary hindsight to progress during my thesis. I appreciate all their contributions of time, and ideas to make my PhD. experience productive and stimulating. The joy and enthusiasm they have for research was contagious and motivational.

I also thank the members of my jury who took the time to read, review and comment my work. Thank you, therefore, to Professor Pascale Kuntz and Professor Arnaud Martin for their role as reviewer and to Professor Florence Sédes for her time and the evaluation of my work.

In general, I would like to express my gratitude to my colleagues in the ETIS lab, to the MIDI team and to the ETIS lab for welcoming me warmly during this thesis, offering excellent reception conditions for doctoral students. Without forgetting Cergy Paris University, where the work was done, the ENSEA, and the Institut Polytechnique du Grand Paris, for the funding of my thesis, without whom I would not be writing this manuscript at this time.

I would like of course to thank my family for their support, my passed father, and my mother to whom I am grateful for raising me up in the way they did. They are the first reason to reach this moment. I thank them for the moral support they were able to bring me during the difficult moments, and less difficult all along my education way. And most of all for my loving, supportive, encouraging, and patient husband Ramzi whose faithful support during the final stages of this PhD. is so appreciated. Thank you.

Résumé

L’essor de grands réseaux sociaux en ligne (comme Facebook, Twitter) a permis la constitution de grands jeux de données temporelles. Ces nouveaux jeux de données ont donné l’occasion aux chercheurs de développer de nouveaux modèles pour décrire et prédire l’évolution de ces systèmes sociaux au cours du temps suivant la dynamique des acteurs sociaux et de leurs interactions. Pour résoudre ce problème, plusieurs modèles ont été proposés dans la littérature, tels que les modèles d’attachement préférentiel et les modèles probabilistes, ou d’autres intégrant des modèles de Markov ou des méthodes spectrales.

Mais une question importante qui doit être abordée est la pertinence de ces modèles d’évolution pour les réseaux sociaux personnels en ligne (OPNs), qui présentent des caractéristiques différentes à la fois structurelles et comportementales, présentant souvent le comportement opposé à celui du réseau social auquel ils participent. Plus précisément, nous sommes intéressés à explorer les caractéristiques structurelles de ces réseaux et leur relation/influence sur leur évolution.

Dans ce contexte, nous abordons deux questions principales : la caractérisation et la prédiction de l’évolution des réseaux personnels. Le premier enjeu nécessite d’explorer l’évolution des réseaux sociaux personnels au moyen d’une analyse de leur topologie basée sur des métriques structurelles. Le deuxième vise, à travers les propriétés découvertes qui régissent les réseaux personnels et leur évolution, à proposer de nouveaux modèles d’évolution adaptés aux OPNs. Pour cela, nous étudions le cas des réseaux de co-auteurs (ou de collaboration) et nous considérons le réseau personnel de chaque auteur du réseau.

Les principales contributions de cette thèse peuvent être résumées comme suit :

(i) Un ensemble de définitions des réseaux personnels en ligne. Tout d'abord, nous introduisons un ensemble de nouvelles définitions des OPNs qui prennent en compte la diversité des réseaux personnels qui existent aujourd'hui, caractérisés par la force et le sens de l'interaction entre deux acteurs sociaux.

(ii) Une étude approfondie de l'évolution des réseaux personnels en ligne. Nous étudions l'évolution d'un vaste ensemble de données de réseaux personnels de collaboration d'auteurs à partir de publications scientifiques. L'analyse a été menée en deux volets : une première analyse a été menée en étudiant l'évolution d'un ensemble de métriques décrivant la topologie des réseaux, et une seconde analyse a été menée sur l'évolution structurelle des réseaux en se concentrant sur leur décomposition en cliques maximales. Les analyses effectuées nous ont fourni une série de découvertes intéressantes nous permettant de comprendre comment ces réseaux évoluent.

(iii) L'implémentation du logiciel PERSONA qui a été utilisé pour mener l'étude précédente. Le framework PERSONA, qui a été conçu et développé dans le cadre de cette thèse, est complet et opérationnel. Il sert d'outil pour l'analyse des réseaux personnels, et comprend un ensemble de métriques pour l'analyse de l'évolution des réseaux personnels et un ensemble de modèles/algorithmes d'évolution à appliquer sur les réseaux.

(iv) Un modèle d'évolution, appelé PERSON*EM*. Les résultats de l'étude approfondie menée sur les réseaux personnels de collaboration nous ont permis de développer un nouveau modèle d'évolution des réseaux personnels, appelé PERSON*EM*, qui prédit un réseau personnel au temps $t+1$, à partir du réseau personnel au temps t et du nombre de nœuds qui doivent être ajoutés. PERSON*EM* est basé sur l'idée qu'un nouveau ou un ancien nœud se connectera toujours à une clique existante.

(v) Une étude expérimentale évaluant l'efficacité du modèle d'évolution PERSON*EM*. Pour cela, nous avons utilisé des données réelles et volumineuses. L'expérience a été réalisée en utilisant un ensemble diversifié de réseaux personnels de collaboration réels afin de prédire

leur évolution. Enfin, nous montrons que les résultats obtenus sont satisfaisants.

En résumé, cette thèse est un premier effort pour comprendre les caractéristiques et l'évolution des réseaux sociaux personnels en ligne et pour fournir des algorithmes et des modèles pour cela. Alors que nous nous sommes concentrés sur les réseaux personnels de collaboration en ligne avec des caractéristiques spécifiques (co-auteurs), les résultats produits pourraient être utilisables par la communauté dans son ensemble.

Keywords : Réseaux Sociaux, Réseaux Sociaux Personnels, Modèles d'Evolution, Réseaux de Collaboration

Abstract

The rise of numerous online social networks has led to the creation of large datasets that have given researchers the opportunity to develop new models to describe and predict the evolution of these social systems over time according to the dynamics of the social interactions. To address this problem, several models were proposed in the literature, such as preferential attachment and probabilistic models, or others integrating Markov models or spectral methods.

But one important question that needs to be addressed is the suitability of these evolution models for online personal social networks (OPNs), which carry different characteristics both structural and behavioral, many times exhibiting the opposite behavior than the social network they participate to. More precisely, we are interested in exploring the structural characteristics of those networks and their relation/influence on their evolution.

In this context, we address two main issues: the characterization and the prediction of the evolution of OPNs. The first issue requires to explore the evolution of personal social networks by means of metrics-based and structural analysis of their topology. The second issue aims at, by means of the discovered properties that govern the online personal networks and their evolution, proposing new evolution models fitting OPNs. To this end, we study the case of co-authorship (or collaboration) networks and we consider the personal network of each author in the network.

The main contributions of this work can be summarized as follows:

(i) A set of definitions for online personal networks. First, we introduce a set of new definitions for online personal networks that take into account the diversity of personal networks that exist today characterized by the strength and the direction of the interaction between two social actors.

(ii) An extensive study of online personal networks evolution. We study the evolution of a large dataset of personal collaboration networks of authors from scientific publications. The analysis was conducted twofold: a first analysis was carried by studying the evolution of a set of metrics describing the networks' topology, and a second analysis was conducted on the structural evolution of the networks focusing on their decomposition in maximal cliques. The performed analyses provided us a series of interesting discoveries allowing us an understanding of how these networks evolve.

(iii) The implementation of the PERSONA software framework that was used to conduct the previous study. The PERSONA framework, that was designed and developed as part of this thesis, is complete and operational, and it serves as a tool for personal network analysis, including a set of metrics for OPN evolution analysis and a set of evolution models/algorithms to apply on the networks.

(iv) An evolution model, called PERSONEM. The results of the extensive study conducted on personal collaboration networks allowed us to develop a new personal network evolution model, called PERSONEM, predicting a personal network in time t+1, starting from the personal network in time t and the number of nodes that have to be added. PERSONEM is based on the idea that a new or old node will always connect to an existing clique.

(v) An experimental study assessing the effectiveness of the PERSONEM evolution model. For this purpose, we used real-life and large data . The experiment was carried out by using a diverse set of real personal collaboration networks in order to predict their evolution. Finally, we show that the results obtained are satisfactory.

In summary, this thesis is a first effort to understand the characteristics and the evolution of Online Personal Social Networks and to provide algorithms and models for it. While we have focused on online personal collaboration networks with specific characteristics (co-authorships), the results produced could be usable by the greater community.

Keywords: Social Networks, Personal Social Networks, Evolution Models, Collaboration Networks

Contents

List of Figures

Chapter 1

Introduction

1.1 Research Context

Nowadays, people participate in many and diverse Online Social Networks (OSNs), which becomes a necessity, since each one serves a different objective: e.g. Facebook is used to share private information with friends, LinkedIn is used to build a professional network and Twitter is used to have access to latest news/information and to share short messages. Inside each of these networks, the user has his/her own personal connections which we call a personal (or ego) network (we will use these terms interchangeably throughout the manuscript), i.e. a network that is composed of the user as its focal point (or ego) and of these actors that the ego is interacting with directly or indirectly.

Research and applications on OSNs have been thriving the last few years following the increase of the user base and the fact that a vast amount of data is being produced based on users' interactions. Works in this area consider for the analysis a single OSN as a whole. But the OSN can also be seen from a different perspective: as a collection and connection of different smaller online personal networks (OPNs) that are brought together and share some common users. Where whole network analysts typically concentrate on uncovering the structure and composition of one big network, personal network analysts almost always study a sample of many smaller personal networks. Thus the egocentric approach fits studies about entities across different networks.

The study of OPNs started in Social Science and got extended in Computer Science. On the one hand, the available literature in the Computer Science domain concerning the study of OPNs is rather limited and mainly focuses on a representation of these networks restricting an individual's personal network to its immediate circle of acquaintances (1-level personal networks). On the other hand, literature from the Social Science didn't deal with a high volume of data, the size of the corresponding graphs and with the dynamics of such networks, making these studies widely inappropriate (while still relevant) when considering OPNs.

Thus, these establishes the fact that there is a real need in reconsidering personal networks in the online settings in order to propose new suitable definitions, making new analysis, and developing new models for this particular kind of OSNs.

In particular, we are interested to the understanding of the dynamics of OPNs, where each OPN is a result of a process of construction and re-composition that takes place over time. OSNs carry many dynamic characteristics and they change over time in terms both of structure (e.g. nodes or connections with other nodes are added/deleted) and weight of the links (e.g. strength of exchanges between two nodes). Understanding the dynamic nature of OSNs has been widely addressed in the literature and many evolution models were proposed to capture this evolution. Most of these models are based on properties that concern the OSN structure as a whole, e.g. the global clustering coefficient. But for an individual, what happens to the whole network is less important than what happens to his/her corner of this online world. So, they are more interested on how their personal network will evolve over time and how this evolution will affect them in terms of their local communities or the information that will reach them.

The dynamics of personal networks can provide insights at various levels. Firstly, at the level of the ego, it can show how the ego is affecting or is affected by his alters over time and how this affects the evolution of the entire OPN. For instance, it has been demonstrated that having more friends over time is associated with an improvement in health [102].Secondly, at the level of the personal network, we are interested in finding if and how different new sub-groups are been developing as the network evolves and how this affects the importance of the ego in the functioning of the personal network.

The study of OPNs evolution is still young and only a few set of works was devoted to, while this can offer a different view on the evolution of an OSN, by seeing it not as a global network that evolves but as a collection of individual networks that change over time both independently and collectively.

In addition, from the limited existing works on OPNs evolution, it is still not clear if their evolution is comparable to the evolution of the whole OSN, and if the traditional social network properties are applicable in the OPN setting.

Thus, we identified the need of discovering the evolution patterns that govern the change of OPNs' structure over time. The discovered properties can then be exploited to build new evolution models for OPNs.

1.2 Contributions

In order to overcome the lacks in the literature discussed in the previous paragraphs concerning OPNs formalization and evolution, we address in this thesis the issues detailed hereafter.

The first issue is the missing of an adequate formalism to define OPNs in their diversity. Indeed, 1-level personal networks' models are inadequate [109] since in OSNs, users can easily interact beyond their immediate social circle (beyond level 1) which changes extensively the notion of the personal network. In addition, existing definitions are focusing on undirected networks while current OSNs are in some cases better represented with directed graphs (allowing for unilateral communications). Another important aspect of OSNs is the intensity of a relationship between individuals, since it allows us to differentiate the connections; this is what we usually capture using the tie strength (of the link). Thus, we propose an extension of the definition initially used in social science of personal/egocentric networks for the online setting. Second, the structural properties of evolving OPNs were not enough studied by previous works, and thus, we performed a large scale analysis of OPNs extracted from real data consisting in a set of co-authoring personal networks describing the personal network of a given scientific publication's author. Third, in order to perform the analysis, we needed to use an adequate tool, which led us to implement a new software named PERSONA. Finally, based on the discovered properties revealed by the performed analysis and starting from an existing evolution model, we propose a new evolution model for OPNs that is based on *clique* dimension, where a clique is a group of individuals in the OPN where everybody is connected directly to everyone else. The clique dimension is a common pattern in co-authoring networks where a publication among a set of authors induces systematically the formation of a clique in the co-authoring network. The proposed model is among the rare efforts in the literature that tempted to model OPNs' evolution. In the following we summarize the contributions of this thesis as follows.

1. **A new set of formal definitions for OPNs**

   We provided a set of definitions that are flexible in order to cover current OPNs and to be extensible for the future. We are presenting three definitions: the first one focuses on undirected OPNs, the second definition is for directed OPNs and the third one extends the first and the second ones and defines weighted OPNs where the links are enriched by a value representing a given property.

2. **A large scale analysis on real OPNs**

   We performed an experimental analysis over a large set of real online personal networks by the mean of the computation of metrics that characterise their structure. The OPNs were extracted from a large co-authoring social network that captures collaboration among scientists who have at some point in time coauthored a publication. We examine how the computed metrics behave when the personal networks change over time in order to detect the properties driving the evolution of personal networks' structure. These conclusions would help us in providing evolution models dedicated to OPNs. To this end, we realized three types of analysis: analysis by the characterization of the edges appearing during 1-personal networks evolution, analysis via topological metrics at the ego level and at the personal network level, and analysis by considering the composition of personal networks in cliques.

3. **Implementation of PERSONA tool**

   We developed a framework named PERSONA: PERSonal Online social Networks' Analytics, where the proposed formal definitions for OPNs have been implemented. PERSONA allows us to extract OPNs based on any of the definitions presented. This allows first to validate the definitions and secondly to offer a versatile tool to anyone who wants to perform computations on OPNs. PERSONA holds various computation capabilities via a large set of metrics and algorithms allowing to understand the OPNs, and a visualization feature of the extracted OPNs.

4. **A new clique-based model for OPNs evolution**

   The results we obtained via the performed experimental analyses, comfort us with the need to propose a new model of evolution adapted to personal networks. We propose PERSON*EM*, a new evolution model for OPNs based on the clique structuring of the network. This model is an extension of the clique-superposition model proposed in [134].

5. **Experimental study**

   The experimental study evaluates the proposed evolution model on the prediction of the evolution of real OPNs of scientists extracted from the collaboration network. The results of application of the model show a good performance given that the predicted networks depicts the same topological and evolution patterns as the real networks with some exceptions that we discuss and propose to overcome in order to increase the model efficiency.

1.3 Thesis Organization

This thesis is organized as follows:

Chapter 2 is concerned with social networks representation and existing personal networks definitions both in the offline and online settings. We give here an overview of the main theoretical concepts governing offline as well as online social networks. We discuss the most important approaches proposed in the literature on the evolution of OSNs and OPNs and on the particular example of co-authorship network.

Chapter 3 introduces the set of the new OPNs' formal definitions we propose for the undirected, directed, and weighted cases with illustrative examples.

Chapter 4 describes the methodology followed to perform the different experimental analysis we performed. We give details about the used data and present the obtained conclusions.

Chapter 5 describes PERSON*EM*, the proposed evolution model for OPNs. To this end, this chapter introduces the clique superposition concept and describes the original associated model given in [134] with an evaluation of this model when considering OPNs and discusses its limits. We propose our adaptation of this model to OPNs and present the results of the model application on co-authorship personal networks.

Chapter 6 provides the architecture of the PERSONA tool that includes the implementation of the definitions, metrics, algorithms, and evolution models. Also, it details the technical choices that were made.

Chapter 7 draws our conclusions and perspectives. One of the most important perspectives is to improve the model performance from a computational point of view and to evaluate it on a different online personal networks data.

Chapter 2

Social Networks and Personal Networks

2.1 Introduction

Nowadays, the use of the term *social networks* in everyday language has become more and more common in terms of community web sites or online social networks like Facebook or Twitter or YouTube, while initially, the social network concept represented traditional means of communication such as face-to-face interaction or communication by phone. Disciplines, from social psychology to anthropology and communication, and from politics and organizational studies, have analyzed the range of applications emanating from such social networks. Researchers in these disciplines have studied not only social actors but the social relationships among these actors. Of all the research, one fundamental finding stands out: many important aspects of societal life are organized as networks [128]. The importance of networks in society has put social network analysis at the forefront of social and behavioral science research.

Online Social Networks are networks of interactions or relationships among individuals on a given internet-based system. The advent of different social media platforms, like Facebook, Twitter, LinkedIn, is creating new opportunities for the analysis of social networks. Their analysis can provide new insights into our social behaviour. Indeed, social media are generating a completely new *online* social environment, where social relationships do not necessarily map preexisting relationships established face-to-face and might have different properties. These properties are discussed later in Section 2.2.

Social networks analysis consider the social network from two perspectives developed by the late 1930s as reported by Chung et al. in [32]. The first approach, known as the *socio-centred* approach was introduced by a group at Harvard University working on ways to find subgroups of people in larger groups. This approach aims in examining the overall network structure to find out answers about the organization of the links within a social group or among social groups. The second approach, known as the *ego-centred* approach, and originated from a group of anthropologists at the University of Manchester which paved the way for community studies, focuses on the individual actor and its immediate neighborhood. The overall network structure can then be seen as a collection of (overlapping) local structures.

Thus, a given actor has his/her own personal connections which we call a *personal (or ego) network*, i.e. a network that is composed of the actor as its focal point (or *ego*) and of these actors that the ego is interacting with (called *alters*).

There is a significant literature coming from the social sciences studying personal networks, e.g. [43], [58], [125], [85] but on the contrary, we will see in this chapter that the study of personal networks in the online environment has received less attention so far. Moreover, we will show in Section 2.3 that personal networks are a good example to highlight the differences between the offline and online social environments.

Another important feature that was neglected so far concerning online personal networks is their dynamic nature since these networks are not static but they evolve over time as new people join or quit them and as new relationships are established or as old ones are broken. In Section 2.4, we give an overview of the works dealing with overall online social networks evolution and online personal networks evolution.

Finally, one example of social networks is the co-authorship network. In our work we focus on this kind of data. So, we present in Section 2.5 some existing works that have analyzed this particular social network as a whole network or have focused on the analysis of the personal networks of authors.

2.2 Social Networks

In sociology, a Social Network (an offline social network) is defined by: *a specific set of linkages among a defined set of persons, with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of the persons involved* [93]. Other definitions exist and defined it by: *a social network is a structure of relationships linking social actors* [87] or: *a social network is the set of actors and the ties among them* [129].

Thus, an *offline social network* requires some kind of *personal* or mediated connection. For example, networks built through face to face interactions and phone calls.

Studies of these classic social networks are characterized by being rather restricted to small systems, and often view the networks as static.

Online Social Networks (OSNs) are considered as a particular type of virtual communities [44] and of social software [1] [110]. Given the fact that the presence of OSNs is related to the Web 2.0, there is not one single well-established definition for OSNs. We report the one given by Schneider et al. in [114]: *OSNs form online communities among people with common interests, activities, backgrounds, and/or friendships. Most OSNs are Web-based and allow users to upload profiles (text, images, and videos) and interact with others in numerous ways.*

Thus, an OSN is the social network formed of users of a given social media, and the links between them. A large variety of OSNs is available and each OSN serves for a given purpose. For example, we have OSNs designed for building online social connections that group together social identities, individuals or organizations. This kind of OSNs can be general (e.g. Facebook, Google+) or professional (e.g. LinkedIn, Viadeo). Another kind of OSNs groups micro-blogging platforms (e.g. Twitter) or content sharing services (e.g. Youtube and Dailymotion for videos, Deezer and Spotify for music, Pinterest, Flickr and Instagram for photos, Slideshare for documents). Regardless their kind, all these OSNs consist of users' profiles, relational patterns, and features to interact with other users.

2.2.1 Social Networks Representation

A social network (offline or online) is formally represented as a social graph. In the beginning until 1946, graph theory remains the domain of mathematics. Then, with the military researches related to the world conflict of 1939-45, operational research is born causing a development of graph theory as models of concrete problems; this aspect is further reinforced in the following years by the development of economics and management sciences. On 1953, Harary and Norman [70] were among the first mathematicians who built mathematical models of offline social networks based on graph theory and established the relation between *sociograms* (a concept proposed by Moreno in 1933 [94] consisting in representing people by points and relationships by lines) and graphs.

During the 1960s, with the development of information and communication sciences, many computer scientists participated in providing algorithms for graphs as well as modeling problems generated by the new sciences in term of graphs.

[1] A social software refers to software that has been designed to support the interactions of people in enterprise work groups [110].

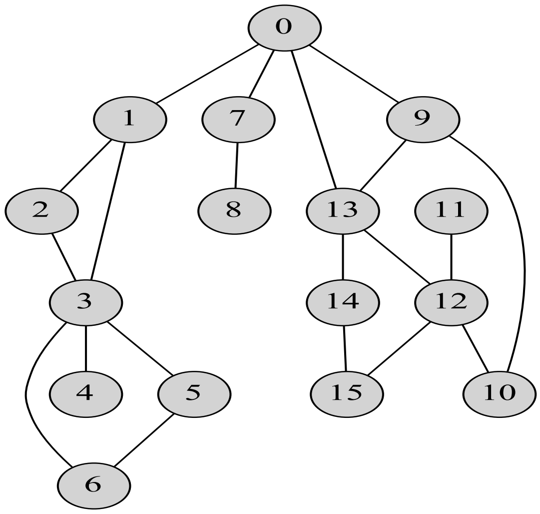


Figure 2.1: Example of a Social Network Graph.

So, with the arrival of communication networks (internet network, email network) and online social networks afterwords, graphs and graph theory background have been used to represent and study their structure.

We give in Definition 2.2.1 a definition of a social network as a graph.

Definition 2.2.1 *A Social Network is a simple graph $G(V, E)$ where V is the set of nodes representing the social actors and E is the set of edges representing the links between them (simple graph means that there is at most one edge between two distinct nodes and no loop). The number of nodes and edges composing G are noted respectively by n and m.*

Example 1 *Figure 2.1 represents a social network described by the graph $G(V, E)$ where: $V = \{0, 1, 2, \ldots, 15\}$ and E is composed of the following links between nodes in V: $E = \{\{0, 1\}, \{0, 7\}, \{0, 9\}, \{0, 13\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{3, 4\}, \{3, 5\}, \{3, 6\}, \{5, 6\}, \{7, 8\}, \{9, 10\}, \{9, 13\}, \{10, 11\}, \{11, 12\}, \{12, 13\}, \{12, 15\}, \{13, 14\}, \{14, 15\}\}$.*

2.2.2 Social Networks Analysis

During the last decades, the interest in Social Network Analysis (SNA) has increased due to the application of graph theory and statistical models in addition to the availability of various software. SNA allows to analyze relationships to understand how the actors are communicating, sharing resources and interests and collaborating together. Understanding the patterns of how individuals interact can serve many real life applications. For instance, SNA can be used for detecting human behavior as crime through people's relations and interactions given the fact that criminal organizations communicate via OSNs [91]. It can also be useful to detect when and where epidemics are spreading based on the messages shared by people evoking the disease [53]. Another important application of SNA is the marketing area [67] where the identification of communities and influencing users might represent interesting aspects for designing marketing campaigns since it allows targeting the products to the right sets of people in the OSN. Many other applications are possible using SNA as building recommendation systems (of friends, contents, trajectories, recipes, etc.) [28], proposing predictive models (e.g. for fraud detection [31], users retention), understanding online communities emergence [29], etc.

McGloin and Kirk [90] identified three categories of social network analysis: using descriptive graphs, using network measures, and using advanced network modeling. The first two methods are descriptive methods of the network's structure that translate theoretical concepts into formal measures. In contrast to a descriptive approach, the third category that uses advanced modeling techniques assume that there is some probabilistic mechanism that governs network data. We describe the three categories in the following.

- **Descriptive graphs:** Descriptive graphs often termed *sociograms* (a concept proposed by Moreno in 1933 [94] consisting in representing people by points and relationships by lines), are simple visualizations of social networks and the connections that exist among the individual actors. Sociograms operate much like a map, providing a picture of relationships between individuals.

- **Network measures:** The second category of SNA involves network measures. Network measures are central to understanding the overall structure of a network, as well as identifying key players within a network. Such analytic tasks are facilitated through a broad array of social network measures. Some measures capture qualities of the network as a whole, such as density (i.e., cohesion), while other measures concern individual's position (or whatever the node is) in the network. For instance, network measures can identify what people are the most popular members of a network (i.e., degree) and to what degree communication in a network must go through certain individuals (i.e., betweenness centrality).

- **Advanced network modeling:** Descriptive methods of network analysis are important for illuminating structural features of a given network, but they cannot be used to build and/or test theories about the generation of networks. Thus, inferential methods of network analysis can be used to find probabilistic models that accurately describe the overall features of a network [50].

In the thesis, we focus on the descriptive method using network measures that are based on theoretical concepts. These theoretical concepts governing offline as well as online social networks, are coming from social sciences that appear in most applications of social networks. Next, we discuss two main approaches to social networks with giving the related formalization from graph theory that are the *cohesion* and the *centrality*.

2.2.2.1 Cohesion

Social cohesion suggests that the members of a social network are forming cohesive subgroups (or communities, modules) consisting of densely connected sections inside the network. Members of a given subgroup interact more intensively and more frequently with each other than the members of a different subgroup. Graph theory offers different ways for identifying cohesive subgroups at the level of the overall network structure and for measuring the cohesion within a subgroup. The most important ones are the *cliques*, the *reachability*, the *density*, and the *clustering*. We describe them hereafter.

Cliques. A *clique* is a subset of V whose elements are pairwise adjacent. In other words, a clique is a complete subgraph of G. The smallest clique in a graph is composed of two actors (dyad).

Definition 2.2.2 *A graph $G(V,E)$ has a **clique** of size s if there exists a subset $V' \subseteq V$, with $|V'| = s$, such that for all u, $v \in V'$, the edge $\{u,v\} \in E$. In other words, the induced sub graph on the nodes in V' is a complete graph of s nodes [3].*

A *maximal clique* is a clique whose cardinal is the largest (no node can be added to the sub-graph without violating the criterion).

Definition 2.2.3 *A **maximal clique** in a graph $G(V,E)$ is a clique of G given by the subset $V' \subseteq V$, such that there is no other clique given by the subset $V'' \subseteq V$ that contains all of V' and at least one other node [3].*

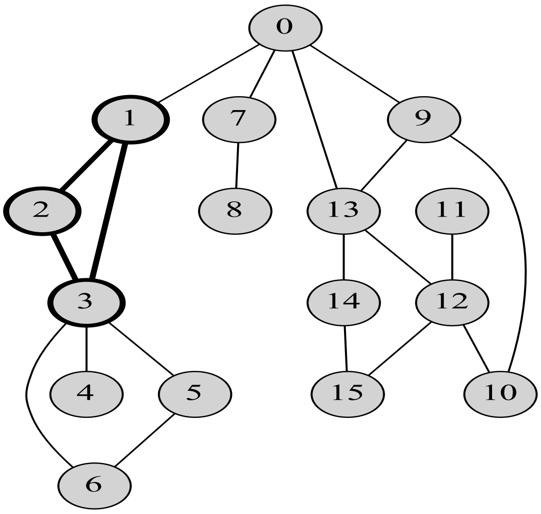


Figure 2.2: Clique and maximal clique examples in a social network graph.

The clique (and the maximal clique) constitute the most strict approach to a cohesive subgroup. There can be more than one maximal clique in a graph.

Example 2 *Based on the network given in Figure 2.2, the subgroup made of the nodes set* $V' = \{1, 2, 3\}$ *is a maximal clique of size 3, while the nodes set* $V'' = \{1, 2\}$ *represents a clique of size 2.*

Reachability. Two nodes are reachable from the one to the other if there exists a path between them. The reachability measures the dyadic cohesion (between a pair of nodes). The most fundamental reachability is the edge connecting a pair of nodes. If there is no edge connecting the pair of nodes than the *geodesic distance* that is the length of the shortest path connecting the pair of nodes is used to capture how close the two nodes are. If the graph is weighted, it is a path with the minimum sum of edge weights.

It is important to outline that there might be more than one shortest path between a pair of nodes, all of the same length. Shortest paths might be found using for example Dijkstra's algorithm [38]. For instance the geodesic distance between nodes 0 and 15 in the graph given in Figure 2.3 is equal to 3. So, the shorter the geodesics between nodes pair, the more cohesive the subgroup is. A variation on this form of cohesion is the *diameter* of the graph, which is the length of the largest shortest path (the diameter of the network in Figure 2.1 equals 6).

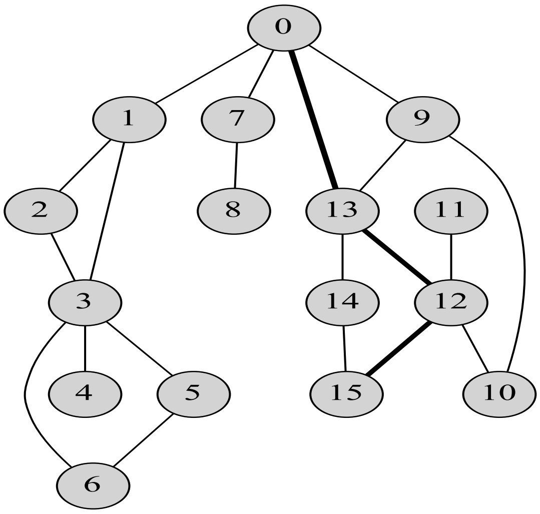


Figure 2.3: Geodesic distance example in a social network.

Density. In a network, more links among its members yield to a tighter structure. The density captures that. It measures how many edges are in the set E compared to the maximum possible number of edges among nodes in V. It can be computed on the entire network or on a given subgroup of individuals. Maximum density is found in a complete network (an example is given in Figure 2.4b), that is a network in which all pairs of nodes are linked with an edge. We give in Definition 2.2.4, the definition of the density as proposed in [58] for an undirected graph $G(V, E)$ where V represents the set of nodes, and E represents the set of bidirectional edges.

Definition 2.2.4 *The density D of an undirected graph $G(V, E)$ is:*

$D = \frac{2 \times m}{n \times (n-1)}$*, where m and n are respectively the number of edges and of nodes of the network.*

Example 3 *A complete network as the one given in Figure 2.4b has the highest density (equal to one) as all possible edges among nodes exist.*

Clustering. Clustering techniques allow to detect cohesive subgroups or clusters of nodes characterized by the presence of many edges among them and few edges with other clusters. The division of network's actors into groups or sub-structures is an important aspect of social structure. It allows to understand how the network as a whole is likely to behave. From the variety of available clustering techniques, we mention the hierarchical clustering, k-means clustering, distribution-based and

(a) Star network. (b) Complete network.

Figure 2.4: Star and Complete subgraphs.

density-based clustering [19]. These techniques allow starting from a given network to get a set of clusters.

In social networks *triads*, that are complete subgraphs of size three, are very frequent and reflect the tendency that two nodes connected to one same node have a high probability of being linked. In SNA, in order to evaluate how much clustered is a network we use the clustering (or transitivity) coefficient metric. The transitivity indicates that if a node u is connected to v and v is connected to w, then nodes u and w are likely to be connected.

The clustering coefficient metric measures the degree to which nodes in a graph tend to cluster together. There are three versions, the *local clustering coefficient* to capture the transitivity in the neighborhood of a given node, and the *average Watts-Strogatz clustering coefficient* [130] which gives an indication of the embeddedness of single nodes as it averages the local clustering coefficients overall the nodes inside the network, and the *global clustering coefficient* (transitivity index) [99] which gives an overall indication of the clustering in the network. Next, we give the definitions for the local clustering coefficient (Definition 2.2.5), the average clustering coefficient (Definition 2.2.6), and the global clustering coefficient (Definition 2.2.9).

Local clustering coefficient (Watts-Strogatz clustering coefficient [130]). The local clustering coefficient is computed at the node level to detect the transitivity inside a node's immediate neighborhood, as defined hereafter.

Definition 2.2.5 *The local clustering coefficient of a node v ($LCC(v)$) in the undirected graph $G(V, E)$ is:*

$LCC(v) = \frac{2 \times m_v}{deg(v)(deg(v)-1)}$, *where m_v is the number of edges between neighbors of v, and deg_v is the degree of v.*

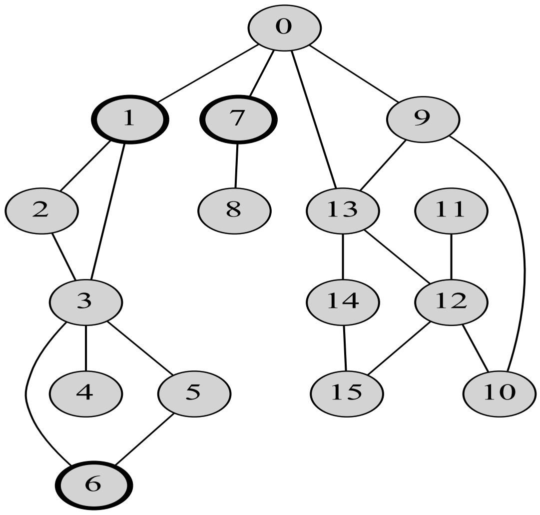


Figure 2.5: Local clustering coefficient example in a social network.

The local clustering coefficient ranges between 0 and 1. It equals 0 in case no edge exists among v's neighbors and 1 in case all v's neighbors are connected to each other.

Example 4 *As an example, if we take as node v, the node 1 in Figure 2.5, its local clustering coefficient* $LCC(1) = 0,33$*, since* $m_v = 1$ *because there is only one link between neighbors of node 1 (the edge (2,3)), and* $deg(1) = 3$ *because node 1 has three neighbors. While* $LCC(6) = 1$ *as all neighbors (nodes 3 and 5) of node 6 are connected, and* $LCC(7) = 0$ *because there is no link between nodes 0 and 8 the neighbors of node 7.*

Average Clustering Coefficient (Average Watts-Strogatz Clustering Coefficient). This metric, introduced in [130] and defined in Definition 2.2.6, allows to characterize how much, in average, a network is locally clustered. The metric is computed by averaging the local clustering coefficients (defined above) of the nodes.

Definition 2.2.6 *The average clustering coefficient* $\langle LCC \rangle$ *of an undirected network* $G(V, E)$ *is:*

$\langle LCC \rangle = \frac{\sum_{v \in V} LCC(v)}{n}$*, where* $LCC(v)$ *is the local clustering coefficient of node* v*,* $\forall v \in V$ *and* n *is the number of nodes.*

$\langle LCC \rangle$ ranges between zero and one. A value close to zero means that in average, the nodes are part of limited transitive relationships, while a value close to one indicates that the nodes participate in many transitive relationships (i.e. in complete OPNs, such as Figure 2.4b it equals 1).

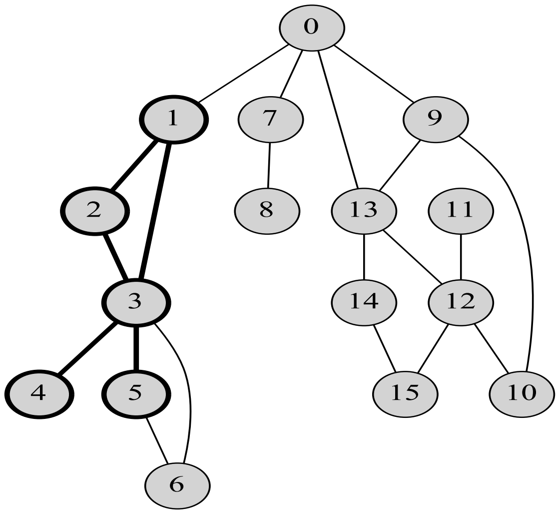


Figure 2.6: Triplet and triangle examples in a social network.

Example 5 *For example, the average clustering coefficient of the network given in Figure 2.5* $\langle LCC \rangle = 0,26$.

Global Clustering Coefficient (Transitivity Index). The global clustering coefficient (GCC), introduced in [86] and defined in Definition 2.2.9, quantifies the transitivity in graphs.

In a network, these structures are called *triangles* (or *closed triplets*). The global clustering coefficient is three times the number of triangles (or 3 X triangles) over the total number of *triplets* (or triads). Before giving the definition of GCC, we give bellow the definitions of a triplet and a triangle.

Definition 2.2.7 *A triplet consists of three connected nodes.*

Example 6 *For example, nodes {3, 4, 5} or {1, 2, 3} in Figure 2.6 form a triplet.*

Definition 2.2.8 *A triangle includes three closed triplets, one centered on each of the nodes.*

Example 7 *In Figure 2.6, nodes 1,2,3 form a triangle as it includes three closed triplets, the first is centred on node 1, the second on node 2, and the third on node 3.*

Definition 2.2.9 *The global clustering coefficient of $G(V, E)$ (undirected graph) is defined as:*

$GCC = \frac{3 \times T}{C}$, *where T is the number of triangles in G and C is the number of triplets in G [86].*

The global clustering coefficient in a complete graph equals 1.

Example 8 *For instance, the global clustering coefficient of the graph given Figure 2.6, $GCC = 0,21$ as it contains 3 triangles and 41 triplets in total.*

2.2.2.2 Centrality

Centrality is a fundamental concept in SNA and it was formally proposed by Bavelas in 1948 [17]. Being central in a social network is associated to the prominence and influence of the actors. In 1979, Freeman [57] advanced three ideas about what being central means translated into three metrics, the *degree centrality*, the *closeness centrality*, and the *betweenness centrality*. We describe them separately in the following.

Degree centrality. The degree of a node is the number of edges the node has. A node maintaining many connections indicates that it is active and important within the network. In a network where information is exchanged the actor with a high degree centrality accesses more sources of information. We give in Definition 2.2.10, the definition of the degree centrality of a node, introduced by Freeman [57], which is defined as the number of connections that the node has in the network. We can define it also as the number of paths of length equal to one that start from the node of interest.

Definition 2.2.10 *The degree centrality (or simply degree) of node u ($deg(u)$) in an undirected graph $G(V, E)$ is given by $deg(u) = \sum_v m_{uv}$, where $m_{uv} = 1$ if an edge exists between the nodes u and $v \in V$, and $m_{uv} = 0$ otherwise.*

Example 9 *For example, the degree centrality of node 3 in the graph given in Figure 2.7, $deg(3) = 5$ since it is connected to five nodes (nodes 1, 2, 4, 5, 6).*

Closeness centrality. The second centrality metric suggests that the more central a node is, the closer it is to all other nodes in the overall network. It is calculated as the reciprocal of the sum of the shortest paths length (geodesic distance) between the given node and all other nodes in the graph. The formal definition proposed by Freeman in [57] as the reciprocal of the farness, is given in Definition 2.2.11.

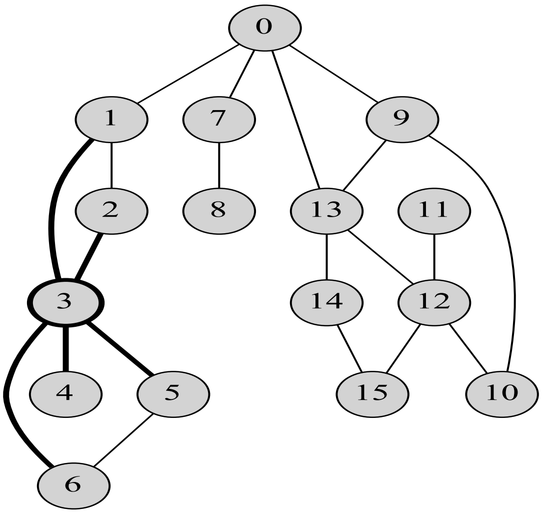


Figure 2.7: Degree centrality example in a social network.

Definition 2.2.11 *The closeness centrality of node u ($C(u)$) in an undirected graph $G(V, E)$ is given by $C(u) = \frac{1}{\sum_v d(v,u)}$, where $d(v, u)$ is the distance between vertices v and u.*

Example 10 *For example, the closeness centrality of node 0 in Figure 2.8 $C(0) = 0,48$, while the closeness centrality of node 11 $C(11) = 0,26$ which is less close to the rest of nodes than node 0 as we can observe from the figure.*

Betweenness centrality. The third centrality metric introduced in [57] argues that being important means being part of many paths between the other nodes in the network. For example, in a telecommunication network, a node having a high betweenness centrality would have more control over the network, since more information will pass through that node. Betweenness centrality is calculated as the fraction of shortest paths between node pairs that pass through the node of interest. The highest betweenness (equal to 1) is found in the star network when considering the node in the centre (an example of a star network is given in Figure 2.4a). Thus, the betweenness centrality assesses the extent to which a node is between all the other nodes in the network. A formal definition is given in Definition 2.2.12.

Definition 2.2.12 *The betweenness centrality of node u ($B(u)$) inside an undirected graph $G(V, E)$ is $B(u) = \sum_{u \neq v \neq w} \frac{S_{v,w}(u)}{S_{v,w}}$ for every pair (v, w); v, $w \in V'$; where $S_{v,w}$ is the number of shortest paths between v and w in G, and $S_{v,w}(u)$ is the number of those passing through u.*

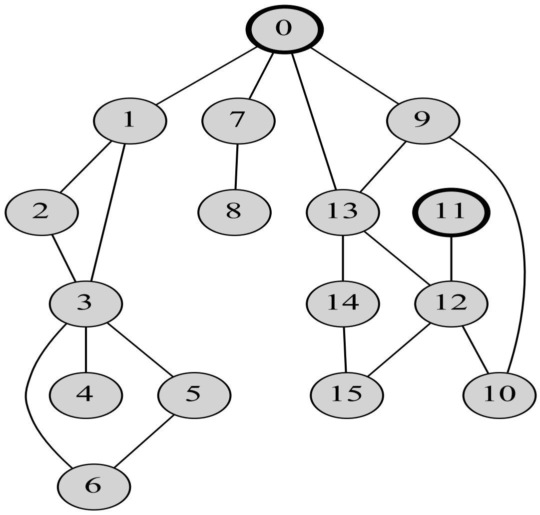


Figure 2.8: Closeness and betweenness centralities example in a social network.

Example 11 *As an example, the node 0 in the network given in Figure 2.8 have the highest betweenness centrality* $B(0) = 0,64$ *because node 0 belongs to many of the shortest paths that are connecting nodes pairs.*

In this section, two concepts conducting the majority of works on SNA were presented i.e. cohesion and centrality. More particularly, works aiming in understanding how social networks' structure changes over time and trying to propose models capturing that evolution are based on these two fundamental concepts and necessarily include some of the discussed metrics in the modeling process. In Section 2.4, we give an overview about works on networks dynamics and show how the cohesion and the centrality are exploited for expressing the change of network' structure. But, before that, we discuss next in Section 2.3 personal networks and we will see how the cohesion and the centrality are applied to this particular kind of networks (Section 2.3.4).

2.3 Personal networks

2.3.1 Offline Personal Networks

Studies in social sciences considering personal networks are interested in the social relations of individuals with their acquaintances (family, friends, etc.). Generally, to collect data, interviews are made on a set of people, and respondents are asked to list the names of their family, friends, colleagues and other eventual acquaintances.

These respondents have no relationships between each other. For instance kinship networks were collected in [43] and [111], or individuals contacts by Christmas cards in [73]. Collecting such personal network data can be time-consuming and expensive.

From the first considerable works on personal network, we have the work of Killworth et al. in [77]. Killworth et al. conducted a series of studies that extended the work of Stanley Milgram [92], who discovered that the average number of links between any two randomly chosen Americans is on average between five and six (the known *six degrees of separation*). In this work the authors showed that the links in the chains were in fact not random, and that the choice of links made by people was typically guided by their location and occupation.

One of the most complete works on personal networks was done by Claude Fischer in [55]. Based on a large representative survey conducted in California, Fisher showed how personal networks operated in all aspects of life (work, religion and attitude formation) and discussed the ways network composition and structure change as we age.

Another line of research concerned a set of methods to estimate the size of hard-to-count populations. Using personal networks, they were able to estimate the network size of each respondent and use these results to estimate how large a hard-to-count population, such as the homeless, would have to be in line with the network-size estimate [82].

During the past three decades many researches have used personal networks as a way to render operational social support. Research into social support is one of the most important applications of personal-network methods. For instance, to explain variability in depression [18], to demonstrate how social support mediates the effects of certain psychological and physical conditions such as stroke [20], or to show the influence of the personal network on gun-carrying behavior among Black adolescents [95].

Personal-network methods are also used extensively in studies of infectious disease like the one in [124] where the authors has used variety of the characteristics of personal networks to explain contraceptive use among women in Cameroon. Also, in [96] to show the relationship between IV drug consummation and HIV virus transmission.

Another important work was done by Robin Dunbar [43]. Dunbar suggests that the ability of people to know other people is constrained by time and resources, as well as the organization of family, friends and acquaintances in memory, to an average of 150 people (known as the *Dunbar's number*). He points to the tendency for various human social groups to fission at around the same size.

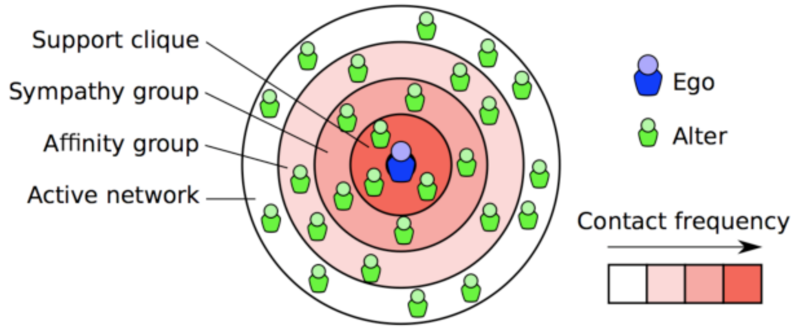


Figure 2.9: Dunbar circles [8].

More recently, topics like the effect of personal networks on voting behaviour [101] were studied.

Thus, we can see that personal networks in social sciences fields have been applied to research problems in many ways.

It is important to note that in all these studies the personal network they consider is composed of the *ego* and the *alters* directly connected to it; thus we have 1-level connections and *1-level personal networks*. An example of a 1-level personal network is given in Figure 2.10. This example reflects the friendship personal network of 'Paul' (the ego). 'Paul' has 5 friends that are 'Mary', 'Alex', 'Bob', 'Leila', and 'Sam' (the alters). Few extensions of this definition were proposed. Among the efforts done in that respect we mention the work of Dunbar who proposed a layered structure to represent personal networks. Indeed, he represented an ego-network' alters set into four layers, each layer reflects the intimacy level alters have with ego as represented in Figure 2.9. Dunbar identified several layers: support clique, sympathy group, band and active network (the whole network) with average sizes of 5, 12, 35 and 150 respectively. In this work Dunbar highlighted the fact that the structure of personal networks is determined by cognitive constraints of the human brain since maintaining an active social relationship requires time and, more generally, cognitive resources used for interacting with the alters. Moreover, he found that there is a relationship between the time invested in a relationship and the level of emotional closeness.

McCarty and Molina defined in their book [89] personal networks as in the following, *Personal networks are about the unbounded networks surrounding individuals*, but without giving additional explanation about what *unbounded* term means.

In the next section, we are looking to *online personal networks*. We will discuss applications that consider personal networks that are extracted from online settings and will focus on how personal networks are defined in the online environment, do

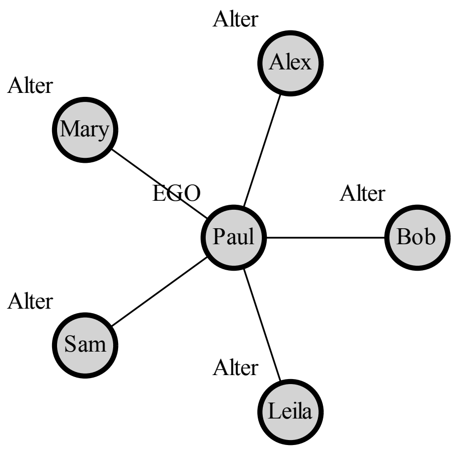


Figure 2.10: Example of the 1-personal network of Paul.

they use the same definition than the one initially provided by social science or is there a different definition? Then, we present the approaches used in SNA for analyzing OPN.

2.3.2 Online Personal Networks (OPNs)

With the arrival of online communication platforms, the question that have interested researchers regarding personal networks is whether these new means of communication modify their properties or not. Two hypothesis can be formulated. On the one hand, one could argue that OSN constitute a new form of communication which will generate new types of personal networks, possibly with totally different structural properties. On the other hand, one could in the contrary argue that OSN are just other means of communication among users, and thus will not affect the way humans interact or have an impact on the resulting structural properties of users' personal networks. This later hypothesis was supported by a set of works that compared the properties of offline and online personal networks and found some correlations between them.

Specifically, Pollet, Roberts, and Dunbar [106] wanted to discover the effect that the use of social media have on the size of each offline personal network layer based on Dunbar's definition reported in the previous section. The authors investigated whether instant messaging and Online Social Networks (OSN) increase the average size of offline personal networks based on a sample of 117 individuals aged 18 to 63 years old, and they concluded that this is not the case.

Arnaboldi et al. [9] analyzed a quite large Facebook data set (3.097.165 users), aiming at identifying whether personal networks of Facebook users present similar layered structures as the ones found in human social networks. Results show a striking similarity between the structure of ego networks in online and offline settings. This similarity was depicted by the presence of concentric layers of decreasing intimacy and contact frequency, and increasing size as explained in Section 2.3.

Other works found evidence of the presence of the Dunbar's number in Facebook [12] and Twitter [64], as a proof of the similarity of the personal networks in physical and cyber environments.

A deeper analysis of Twitter users' personal networks [10] revealed that the size of their layers is very similar to the ones found in Facebook and in offline environments. However, the contact frequency of the layers in Twitter is more than twice those for Facebook and offline personal networks which could be attributed to the nature of Twitter communication, which is characterized by short but frequent messages.

Again, based on Dunbar's definition, another kind of works have proposed generative models to produce realistic synthetic personal networks ([105], [36]). These algorithms are based on the key structural properties of personal networks model proposed by Dunbar in the anthropology literature, and the properties of the social relationships between individuals.

Apart from this kind of works on comparing offline and online personal networks, online personal networks have not received a lot of attention so far. From the rare applications in the literature, Cook et al. [37] have made regression analyses to examine the relationship between online personal networks characteristics (i.e., characteristics of the alters directly related to the focal participant plus the relationships shared among alters within the online network) and alcohol. Moreover, Chung et al. [33] developed a theoretical model to investigate the association between social network properties, *content richness* (CR) in academic learning discourse, and performance. CR is the extent to which one contributes content that is meaningful, insightful and constructive to aid learning. Analysis of data collected from an e-learning environment shows that rather than performance, social learning correlates with structural properties (as the density), with the position of individuals, and with relationship properties of social networks (tie strength).

Finally, another family of works have proposed tools dedicated to online personal networks analysis as *Ego-net digger* [80], which is a Facebook application for the analysis of personal networks. Ego-net digger collects users' social data and gives a representation of their personal social networks according to the Dunbar's circles model. *EgoNav* [71] is a visual analytic system that characterizes egos based on the relationship structure of their ego-networks and presents the results as a spatialization. *EgoNav* allows to visually summarizing a collection of personal networks. Gao and Berendt [61] proposed an exploratory tool that help its users to categorize friends more effectively in their online personal networks. At last, *E-NET* [69] is a software designed for personal network analysis. It offers a view of the key measures for personal network analysis such as size, composition and structure.

In the next section, we focus on how online personal networks are defined among all these works. Is there one common definition or different ones? Is their definition corresponding to the one provided in social sciences? In the following, we present two sets of definitions that exist in the literature: *1-personal networks* and *k-personal networks.*

2.3.3 Personal networks definitions

2.3.3.1 1- personal networks

The definition of the 1-personal network in the offline settings discussed in section 2.3.1 inspired many researches in OSNs ([35], [12], [9], [10]) who provided a series of definitions of 1-level online ego-networks, differentiated by the links that they are considering.

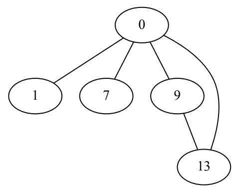
2.3.3.1.1 Ego-networks with ego-alter and without alter-alter links. The work in [9] focuses on 1-level ego-networks that do not take into account the alter-alter links.

Definition 2.3.1 *A personal network is defined as a graph $EN^e = (V', E')$ centred around the focal node e and composed of the set of nodes V' and the set of edges E'. The set of nodes of the 1-level ego-network EN^e of the ego node e is composed of the focal individual e and the individuals that are directly connected to (referred to as alters):*
$V' = \{x \in V \mid \{e, x\} \in E\} \cup \{e\}$
The set of edges of an ego-network EN^e of the ego node e having V' as set of nodes, is composed of the connections between e and all the alters:
$E' = \{\{e, x\} \in E \mid x \in V' \setminus \{e\}\}$

Example 12 *If we consider the OSN in Figure 2.1, the 1-level ego-network of the node 0 (e), according to the Definition 2.3.1, is composed of the node set V' =*

(a) 1-level ego-network of the ego node 0 without alter-alter links.

(b) 1-level ego-network of the ego node 0 without ego-alter links.



(c) 1-level ego-network of the ego node 0 with all types of links.

Figure 2.11: 1-level ego-network examples.

$\{0, 1, 7, 9, 13\}$*, and the edge set* $E' = \{\{0, 1\}, \{0, 7\}, \{0, 9\}, \{0, 13\}\}$ *representing the connections between the ego node e and his alters. This 1-level ego-network is given in Figure 2.11a*

Investigating the similarity of structures of online and offline social networks, the studies in [12] and [10] have used the above definition given initially by the human sciences' community. Firstly in [12], an ego-network in Facebook was defined by the set of nodes composed of the ego and his Facebook friends, where the friendship relation describes the ego-alter links. Secondly, in [10], an ego-network in Twitter was defined for a focal user by specifying alters as the Twitter users to whom the ego had replied at least once. It is important to note that the alter-alter links are not included.

2.3.3.1.2 Ego-networks without ego-alter and with alter-alter links. In [121], the authors propose to take into account for 1-level ego-networks the alter-alter links and not the ego-alters links. Thus, we can consider the following definition.

Definition 2.3.2 *An ego-network* EN^e *of the ego node e is composed of the set of nodes* $V' = \{x \in V \mid \{e, x\} \in E\} \cup \{e\}$ *, and the set of edges* E' *that is composed of the connections between the alters:*
$E' = \{\{x, y\} \in E \mid x \in V' \setminus \{e\} \wedge y \in V' \setminus \{e\}\}$

Example 13 *Using the example in Figure 2.1, the ego-network corresponding to the above definition is described by the set of nodes* $V' = \{0, 1, 7, 9, 13\}$*, and the set of edges* $E' = \{\{9, 13\}\}$.

Compared to Definition 2.3.1 only one link is part of the set of links, as it is the only one connecting a pair of alters. The links between the ego and his alters are not included into the ego-network (Figure 2.11b).

The authors justify their proposal under the hypothesis that the relations between the ego and its alters were not useful as the authors were interested in detecting communities inside the ego-network of a user.

2.3.3.1.3 Ego-networks with all types of links. The previous definitions were describing an ego-network where the set of edges was composed *only* of ego-alter links *or only* of alter-alter links. Other studies ([135], [109]), consider *both* types of links as part of the ego-network.

Definition 2.3.3 *An ego-network* EN^e *of the ego node* e *is composed of the set of nodes* $V' = \{x \in V \mid \{e, x\} \in E\} \cup \{e\}$ *, and the set of edges that is composed of the connections between alters, and between ego and alters:*
$E' = \{\{x, y\} \in E \mid x \in V' \wedge y \in V'\}$

Example 14 *For the OSN example in Figure 2.1, the 1-ego-network defined by Definition 2.3.3 is given in Figure 2.11c. The 1-ego-network is composed of the node set* $V' = \{0, 1, 7, 9, 13\}$*, and the link set* $E' = \{\{0, 1\}, \{0, 7\}, \{0, 9\}, \{0, 13\}, \{9, 13\}\}$*. Comparing this to the ego-network in Example 12, the set of edges has the extra link between the alters* 9 *and* 13*. Unlike to the Example 13 of the Definition 2.3.2, the ego-alter links* $\{0, 1\}, \{0, 7\}, \{0, 9\}, \{0, 13\}$ *are added to the alter-alter link* $\{9, 13\}$.

This definition is justified by the authors by the fact that the work's goal is to do quantitative analysis on OSN data to measure network metrics such as *network constraint* (reflects the extent to which ego links to other nodes that are already connected to each other) that involve interconnections between alters and those between ego and alters [109] or to check the overlap between ego and alters neighborhoods [135]. So they need for this kind of measurement to have both ego-alters and alters-alters connections.

In this section we presented three definitions for a 1-level ego-network that differ by the links that they consider, each one used for different goals. These definitions are inspired by the definition used in sociology that implies to take the immediate neighborhood of the ego node but they remains dissimilar with regard to the way they are considering the edges as part of the ego network or not.

But these definitions are inadequate for OSNs, which go beyond the 1^{st} level and where users might interact with users not directly related to them since people communicate beyond physical contact. Indeed, considering only the alters directly connected to the ego brings a strong limitation, since it will not allow to correctly represent the ego-network of a user who might interact with individuals not directly connected. For example, on Twitter, if two users do not have a direct connection (*following* link), they still can communicate via a common friend (e.g.: followed by both of them) or in an indirect way through the use of e.g. a *retweet.* Consequently, there is a real need to expand the 1-level ego-network definitions to cover these cases.

2.3.3.2 k-personal networks

The three definitions presented above have the specificity to focus on the ego-networks composed only of the ego node and the alters with whom the ego has a direct link. In addition to 1-level ego-networks, recent works ([121], [62]) propose to define ego-networks by including potential alters which are at a maximum distance k from the ego (what is called a k-ego-network).

Definition 2.3.4 *The k-ego-network of an ego e is the network of the individual e consisting of (1) e and the individuals situated at a maximum distance k of e (namely V'), and (2) all the edges between the two individuals (alters) in the k-ego-network except those with e (namely E') [121]:*
$V' = \{x \in V \mid d_G(e,x) \leq k\} \cup \{e\}$,
$E' = \{(x,y) \in E \mid x \in V' \setminus \{e\} \wedge y \in V' \setminus \{e\}\}$.

Example 15 *If we consider the example in Figure 2.12, k is equal to* 3 *(maximum distance between the ego node '0' and the other nodes is* 3*). The node set will contain all nodes ($V' = V$) and the link set will cover all the links in E' except those connecting e ($E' = E \setminus \{\{0,1\},\{0,7\},\{0,9\}\{0,13\}\}$).*

In [62], the authors also considered k-ego-networks but for Twitter's following graph. The structure adopted could be seen as a particular case of the previous definition since it includes the links with the ego and omits those between alters situated at the ego-network last level. Based on the running example in Figure 2.1, the link $(5,6)$ will be omitted. Despite extending the definitions from 1- to k-level ego-networks,

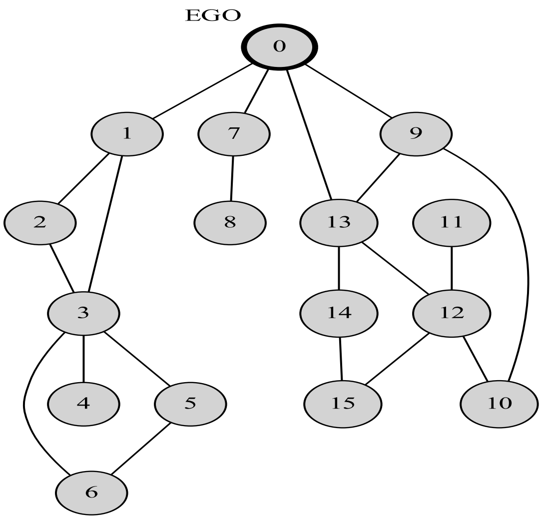


Figure 2.12: Example of a Personal Network with $k = 3$ for ego node '0'.

omitting connections or not specifying their direction renders them incomplete and limits their applicability only in specific cases.

In this section, definitions for ego-networks proposed in the literature were presented in order to assess their suitability to represent OPNs. We identified several limitations of these definitions were identified:

- Existing definitions are very often focusing on undirected networks while current OSNs are usually better represented with directed graphs (allowing for unilateral communications);

- 1-level ego-networks' models are inadequate [109] since in OSNs users can easily interact beyond their immediate social circle (beyond level 1). This changes extensively the notion of the personal network (considering it as a network of interacting actors) and this is not covered in the presented definitions;

- The intensity of a relationship/link between persons/nodes is important since it allows us to differentiate the connections; this is what we usually capture using the tie strength (of the link). The tie strength is prominent in studying social influence between individuals [7], [13], but also it affects the information diffusion process as demonstrated in [52]. The inability to express tie strength, limits our ability to properly describe the OPNs.

In the next chapter (Chapter 3), we propose a set of formal definitions for different types of OPNs, which carry characteristics that have been insufficiently or not at all addressed above.

2.3.4 Personal networks analysis

In section 2.2.2, we presented the main approaches to social networks analysis, the *cohesion*, and the *centrality*. Next, we discuss these two concepts but in the case of *personal networks analysis*.

2.3.4.1 Cohesion

Similarly to overall social networks analysis, personal networks analysis tries to capture the cohesion within personal networks via the same concepts, but projected on an individual's network. In the following, we present the set of metrics used in the literature to quantify the cohesion of personal networks as usually defined i.e. 1-level personal networks.

Cliques. A clique was involved as a personal network-level graph-based metric in [88]. A very recent work in [76], have used the clique as a new framework for representing ego-networks. Indeed, they constructed "cliques complexes" corresponding to node neighborhoods (ego-networks) and then analyzed this topology of clique complex by computing the Betti numbers [59], which is a metric that measures the number of connected component in the resulting clique complex topology, presented in that work as an alternative to the clustering coefficient metric.

Based on the definition given in Definition 2.2.2 for social networks graphs, we give the definition of a clique in a personal network:

Definition 2.3.5 *We consider a personal network as a sub-graph $G'(V', E')$ of the graph $G(V, E)$, where $V' \in V$ and $E' \in E$. This network has a* ***clique*** *of size s if there exists a subset of nodes $V'' \in V'$, with $|V''| = s$, such that for all u, $v \in V''$, the edge $(u, v) \in E'$. In other words, the sub-graph composed of nodes in V'' is a complete graph of s nodes.*

According to the same principle, we define a *maximal clique* in a personal network starting from Definition 2.2.3 proposed in [3] as a clique of the personal network whose cardinal is the largest.

Definition 2.3.6 *We consider a personal network as a sub-graph $G'(V', E')$ of the graph $G(V, E)$, where $V' \in V$ and $E' \in E$. A* ***maximal clique*** *in the personal*

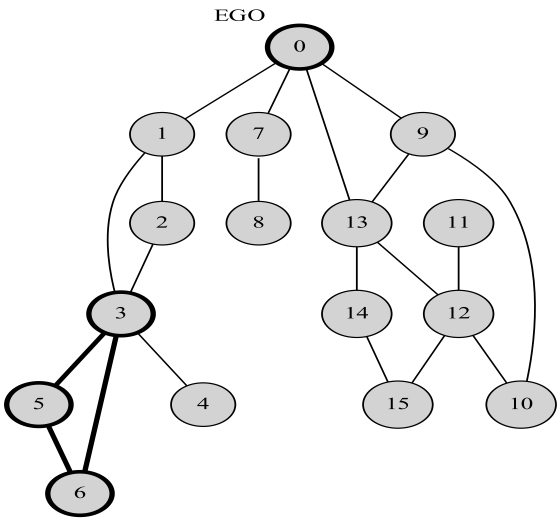


Figure 2.13: Example of a clique in the Personal Network of Ego '0', k=3.

network $G'(V', E')$ is a clique of G' given by the subset $V'' \subset V'$, such that there is no other clique given by the subset $V'' \in V'$ that contains all of V'' and at least one other node.

Example 16 *Based on the personal network of ego node '0' given in Figure 2.13, the subgroup made of the nodes set $V' = \{3,\ 5,\ 6\}$ is a maximal clique of size 3, while the nodes set $V'' = \{5,\ 6\}$ represents a clique of size 2.*

Reachability. As presented for social networks, two nodes are reachable from the one to the other if there exists a path between them. The reachability inside a personal network is restricted since in the common definition of a personal network limited to 1-level personal networks, all the nodes (alters) are known to be at distance one from the ego and at distance of maximum 2 from each other.

However, several studies tempt to compute the *two-step reach* ([72], [137]), defined as the percentage of all actors in the whole network that are within two steps of ego. The *two-step reach* can also be normalized by the size of the personal network and gives what it is called the *reach efficiency* [137]. A high reach efficiency indicates that ego's alters are influential in the network.

Example 17 *In Figure 2.14, we outlined the nodes inside the 3-personal network of ego node '0' that are at two steps from that ego which consists in the set of nodes* $\{3, 2, 8, 14, 12, 10\}$. *The two step reach is thus equal to 40% as there are 6 nodes at two steps from the ego and the network here has 15 nodes (by excluding the ego).*

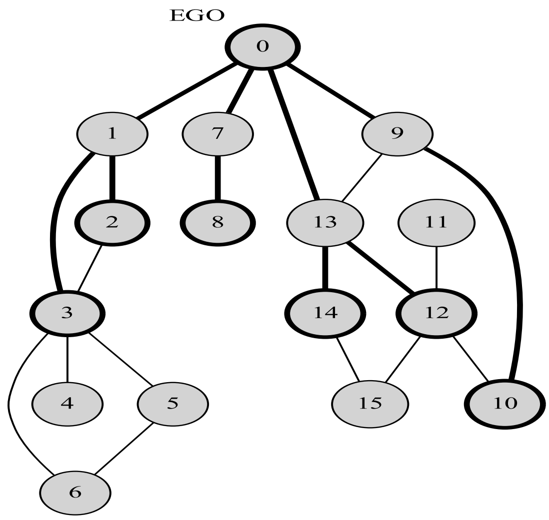


Figure 2.14: Example of Two Step Reach in the Personal Network of Ego '0', k=3.

Density. In social networks, the density measures the amount of existing edges compared to the maximum possible number of edges between nodes. The density as defined in Definition 2.2.4, was presented in [88] as one of the most widely used structural metrics for personal networks. The definition of the density for personal networks is given below.

Definition 2.3.7 *We consider the personal network EN^e of an ego e. The density D of a personal network given by the undirected sub-graph $G'(V', E')$ of the undirected graph $G(V, E)$ is:*

$D = \frac{2 \times m}{n \times (n-1)}$, *where m and n are respectively the number of edges and of nodes of the personal network.*

As we can observe, while previously we defined the density for the entire graph $G(V, E)$, here we capture the density of a sub-graph $G'(V', E')$ of the graph $G(V, E)$.

Example 18 *The density of the 3-personal network of the ego node '0' given in Figure 2.15, $D = 0,16$.*

2.3.4.2 Centrality

Centrality metrics seen in Section 2.2.2.2 (eg. degree, closeness and betweenness centrality) were applied on personal networks either to compute them at the level of the ego as in [87] or at the level of a given alter as in [88]. We report hereafter the

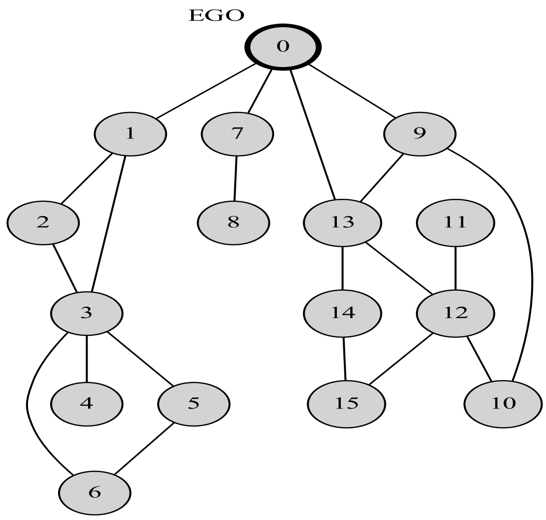


Figure 2.15: Example of a Personal Network with $k = 3$ for ego node '0'.

three definitions one for each centrality measure adapted for the personal network' ego node.

Definition 2.3.8 *We consider the personal network EN^e of an ego e. The degree centrality of the ego node e (deg(e)) in the personal network given by the sub-graph $G'(V', E')$ is given by:*

$deg(e) = \sum_y m_{xy}$, where $m_{ey} = 1$ if an edge exists between the nodes e and $y \in V'$, and $m_{ey} = 0$ otherwise.

Definition 2.3.9 *We consider the personal network EN^e of an ego e. The closeness centrality of the ego node e ($C(e)$) in the personal network given by the sub-graph $G'(V', E')$ is given by:*

$C(e) = \frac{1}{\sum_y d(y,e)}$, where $d(y, e)$ is the distance between nodes y in the personal network and ego e inside the personal network sub-graph.

Definition 2.3.10 *We consider the personal network EN^e of an ego e. The betweenness centrality of ego node e ($B(e)$) inside the personal network given by the sub-graph $G'(V', E')$ is:*

$B(e) = \sum_{e \neq y \neq z} \frac{S_{y,z}(e)}{S_{y,z}}$ for every pair of nodes (y, z), y, $z \in V'$, where $S_{y,z}$ is the number of shortest paths between y and z in G, and $S_{y,z}(e)$ is the number of those passing through e.

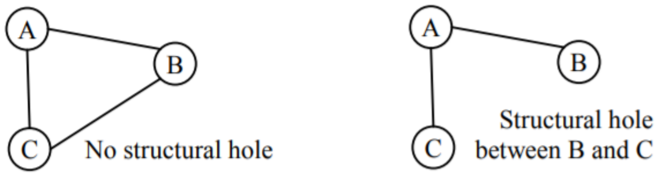


Figure 2.16: Example of a Structural Hole [137].

Example 19 *Starting from the example of the personal network of ego node '0' given in Figure 2.15, the degree centrality, the closeness centrality and the betweenness centrality of the ego node '0' are respectively:* $deg(0) = 4, C(0) = 0,03$, *and* $B(0) = 0,64$.

2.3.4.3 Structural holes

The term "structural holes" was introduced by Ronald Stuart Burt in [23]. It suggests that the ways an individual is connected with its neighborhood reveal information about holding certain positional advantages/disadvantages and so, can be useful in understanding the influence, the power, and dependency effects. A "hole" is a gap between two individuals B and C both connected to node A. In this situation, A has an advantaged position resulting from the "structural hole" between B and C since A has two alternative partners for exchanging information; while B and C have only one choice. An illustration of this example is given in Figure 2.16.

Next, we give the definition for measures related to structural holes as proposed in [137].

Effective size.

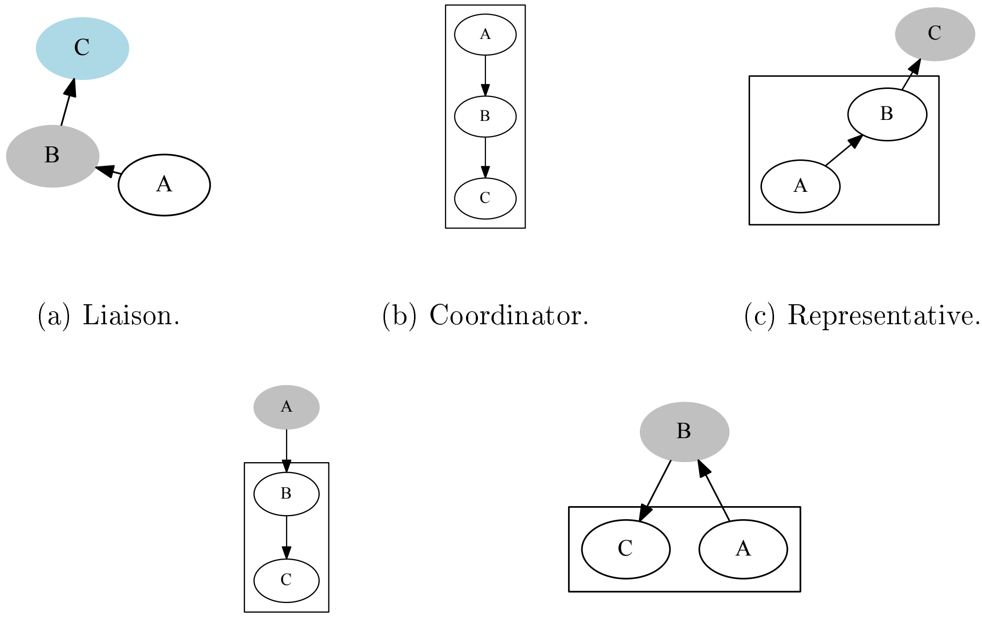
Definition 2.3.11 *We consider the personal network* EN^e *of an ego* e. *The effective size is the number of alters* n_a *the ego node has, minus the "redundancy" with alters given by* $\frac{2\times t}{n_a}$, *where* t *represents the number of ties between alters.*

Efficiency.

Definition 2.3.12 *The efficiency norms the effective size of a personal network to the size of the personal network* n_a.

Constraint. Captures the extent to which the ego is linked to alters who are also connected with each other [23].

Hierarchy. Measures how the constraint measure is distributed across neighbors [23].



(a) Liaison. (b) Coordinator. (c) Representative.

(d) Gatekeeper. (e) Consultant.

Figure 2.17: Brokerage Roles.

2.3.4.4 Brokerage

Fernandez and Gould [66] had another different approach than the one given by Burt [23]. The focus this time is on the ego and the role he plays in connecting groups. The ego "broker" is the node in the middle of a directed triad. Depending on the group to which each node composing the triad belongs, the ego broker node can act into five possible roles as presented in Figure 2.17.

In this section, we presented the different concepts used to characterize personal networks structure. In the next section, we discuss another important aspect of online social networks in general and online personal networks in particular that is their dynamic nature. Indeed, understanding OPN structure should not be restricted to their static state. It is important to discover how their structure changes over time and what properties are governing these changes. We present in the following works that discovered patterns characterizing OSN evolution and proposed a set of models. Then, we give an overview of works that have studied OPN evolution and highlight the limitations we find in that area.

2.4 Evolution of OSNs and OPNs

Social networks are dynamic. Individuals create or break relationships, communicate, join or quit the network, and so affect its structure. In the last years, numerous studies in SNA have been devoted to understanding the evolution (or dynamics) of social networks. The main issue in this area is the definition of mathematical models able to capture and to reproduce properties observed on the real networks.

Before going into the details of these models, we should first distinguish between two well studied concepts: *evolving networks* and *temporal networks*. While they seem to have the same meaning, these two concepts are actually approaching differently the time-changing nature of networks in general.

2.4.1 Evolving networks

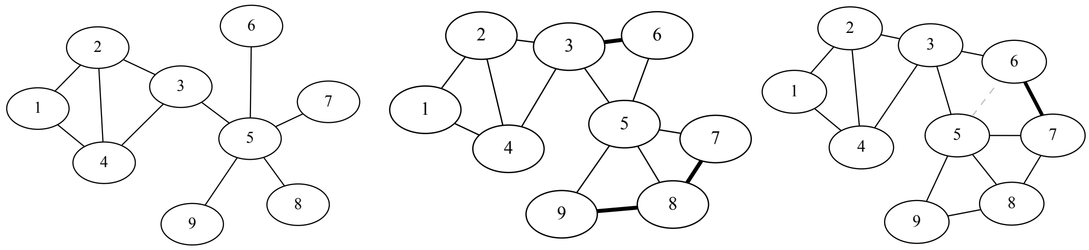
An evolving (or dynamic) network is defined as a sequence of static networks, where each network represents the state of the network at a given time step. For example in Figure 2.18, we give a sequence for a network that evolves over three time steps t, $t+1$, and $t+2$. In such networks, the change of the topology over time occur when for instance people create or break a relationship. If we consider a social network at time t with a graph $G_t = (V_t, E_t)$, the evolving graph is given by the following definition proposed in [115].

Definition 2.4.1 *An evolving graph $G[t_i, t_j]$ in a time interval $[t_i, t_j]$ is a sequence $\{G_{t_i}, G_{t_{i+1}}, ..., G_{t_j}\}$ of graph snapshots.*

Example 20 *In Figure 2.18, the evolving graph $G[t, t+2]$ from time step t to time step $t+2$ is defined by the sequence $\{G_t, G_{t+1}, G_{t+2}\}$. Where from time step t to time step $t+1$, edges $\{3,6\}, \{7,8\}, \{8,9\}$ where added to G_t. Then, from time step $t+1$ to time step $t+2$, edge $\{6,7\}$ was added, and edge $\{5,6\}$ was removed from G_{t+1}.*

2.4.2 Temporal networks

A temporal (or time-varying) network, is a network whose links last relatively for a short moment or are instantaneous (e.g. e-mails, phone calls). Temporal networks are relevant to spreading processes such as the spread of a virus or a message where the time ordering of the induced links is important. Usually, we represent a temporal network with a Time-Varying Graph (TVG). Hereafter, we report the definition proposed in [112].



(a) Network at time t. (b) Network at time t+1. (c) Network at time t+2.

Figure 2.18: Evolving Network Snapshots.

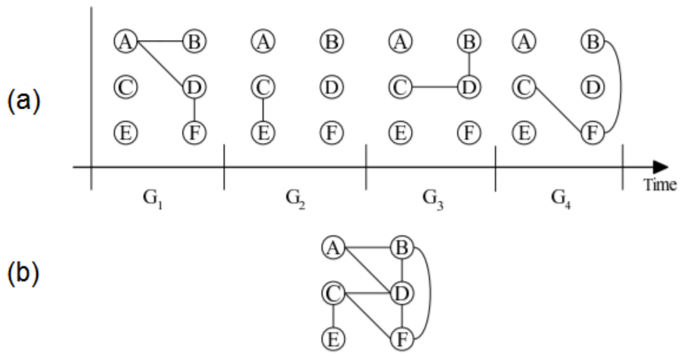


Figure 2.19: Example of a Time Varying Graph (TVG) [120].

Definition 2.4.2 *A time-varying graph is a structure $G = (V, E, \mathcal{T}, \rho)$, where $\mathcal{T}$ represents the time span over which relationships are assumed to take place, and $\rho : E \times T \mapsto \{0, 1\}$, called presence function, indicates whether a given edge is present at a given time.*

Example 21 *In Figure 2.19, we have a time varying graph G with $T = 4$ (panel a), and its projection into a static graph (panel b).*

Thus, evolving networks are those evolving slowly over time and snapshot analysis can be performed effectively, while temporal networks are created by transient interactions and require real-time methods to be analyzed [2]. In this thesis, we are interested in evolving social networks where the evolution concerns the addition or the deletion of individuals and/or relationships. Such changes occur on the time-scale of days, months or even years. So, our work is assigned to evolving networks approach rather than temporal networks approach. Next, we discuss the main findings about evolving networks models.

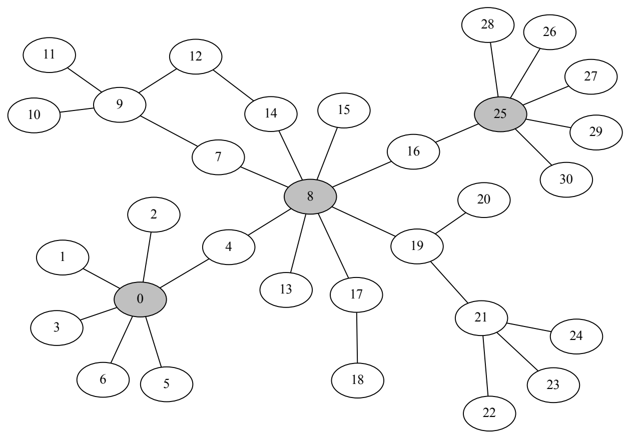


Figure 2.20: Scale-free Network.

2.4.3 First evolving network model - Scale-free networks

The first model that have been widely accepted to produce scale-free networks is the one proposed by Barabasi and Albert on 1999 [14]. This model only integrates the addition of edges by adding new nodes to the graph. A scale-free network is a network whose degree distribution follows a power law. Barabasi-Albert model is based on the *preferential attachment (PA)* where new nodes added to the network are more likely to connect to highly connected nodes with a probability proportional to their degree. We give an example of a scale-free network in Figure 2.20. This "rich get richer" principle induces power-law degree distributions as new arriving nodes will tend to connect to nodes having a high degree, which will lead to the appearance of some highly connected nodes in the network and many weakly connected nodes.

Formally, the fraction $P(k)$ of nodes in the network having k connections to other nodes goes for large values of k:

$P(k) \sim k^{-\alpha}$, where α represents the power law's exponent and depends on the details of the model whose value is typically in the range $2 < \alpha < 3$.

Many real networks have been reported to be scale-free such as the network of citations between scientific papers [107], the movie actors network [42], the biological networks [4], the World Wide Web network [14], and the Internet [49]. Moreover, it has been found that online social networks follow a power law degree distribution as proved in [83] for four OSNs: *Flickr*, *delicious*, *Yahoo! Answers* and *LinkedIn*. Thus, the majority of works that have attempted to propose models for online social networks evolution are based on Barabasi-Albert model. We give an overview of these works in the following section.

2.4.4 Evolution models for OSNs

The availability of online social networks offered new opportunities to researchers coming from different domains interested in proposing evolution models. Thus, a variety of methods has been applied to formalize the idea of evolution.

Snijders proposed in [116], [117] to use classical stochastic models, mainly the Markov model [56] to model networks dynamics. This method was extended in [26] where a multi-layer Hidden Marcov Models (HMMs) was applied as a solution to model the social dynamics of digital ecosystems (DE). Precisely, the authors tried to predict the evolution of relationships inside the overall network (outer layer), by considering the evolution of predefined sub-networks (inner layer) in terms of their nature (according to user's interest) and their intensity. The dynamics of the network are formulated by transition probabilities between hidden states of the sub-networks that have to be recognized (appearance, increase, decrease, equality or disappearing of a sub-network). The proposed model was evaluated on *Yahoo! Groups* data and has achieved a good level of prediction accuracy.

Spectral methods were also applied to model networks' evolution as in [79], where the evolution is expressed using a link prediction function that translates the eigenvalues decomposition of the adjacency matrix representing the network. So, the proposed spectral evolution model tries to predict where new edges will be added as the network grows over time. Experiments to evaluate link prediction accuracy were conducted over different social networks including online social networks as *Facebook, Epinions, YouTube* or *Flicker* on which good performances have been observed.

The remaining and prevailing set of works covers probabilistic models built on the basis of Barabasi-Albert PA model discussed in the previous section. Barabasi-Albert model provides a general mechanism to predict a network whose degree distribution follows a power law but it is insufficient to describe real online social networks. Indeed, it only focuses on the role of nodes' degree but suffers from the lack of constraints integration about links and local structures formation and of effects that influence the creation of links among users.

This fact led the researchers to integrate notions as the *transitivity* which suggests that social networks' actors will tend to create relationships with actors to whom their neighbors (immediate connections) are connected to as in [75]. In addition, other effects that might influence user's behaviour were integrated in the model proposed in [75] for social rating networks [2] as the *social influence* suggesting that users adopt

[2]A social rating network is a social network in which edges represent social relationships and users (nodes) express ratings on some items [75].

the same rating behaviour as their friends or the *selection* which hints that people form relationships with people similar to them. The proposed generative model tries to reproduce a network having the properties observed on real social rating networks (*Flicker* and *Epinions*).

Gong et al. in [65] proposed a preferential attachment and triangle closing based generative model for social attribute networks [3]. In this work, first, the authors made a set of observations on *Google+* network where by using a set of metrics like the density, the diameter, the clustering coefficient, and the degree distribution they tried to see the impact of the attribute structure on the social structure. Then, a model including two components i.e, attribute-augmented PA and attribute-augmented triangle closing was proposed on the basis of the observations derived from *Google+*.

Other works incorporated the spatial factor as an additional constraint that fosters the creation of social links in OSNs like in [5]. In this work, the authors proposed an approach to estimate for each new edge, the probability that it would be created according to a given model among different tested models: attachment by node degree (PA), age, distance or by node degree and distance, triangle creation between two nodes at random, proportionally to the number of common friends, inversely proportional to the geographic distance. The objective was to identify among these models the one with the highest likelihood to explain the data from *Gowalla* OSN. They found that the geographic distance plays an important role in the creation of new social connections since connections arise among users visiting the same places.

In addition to the cited effects influencing the evolution of social networks, users' social activities inside the OSN were considered in [118]. Enriching the social network with user activity information can provide a more fine-grained understanding of online social networks evolution. The proposed model in [118] formalizes two concepts: the social actions of hobbies searching and friend recommendation in a social network, assuming that they constitute common ways for meeting friends and forming communities. The proposed topology evolution model could reproduce several key network topological properties.

As outlined, all the models we discuss in this section are based on properties that concern the OSN structure as a whole, while in our work we are interested on how a personal network of a given individual will evolve over time. In that respect, we give

[3] A social attribute network is a social network augmented with social attributes, in which nodes represent either social actors or social attributes (e.g. affiliation, city..etc), and edges represent social links among social nodes or attribute links between social nodes and attribute nodes [65].

in the next section an overview of works aiming in understanding and studying the evolution of OPNs.

2.4.5 Evolution studies of OPNs

In the previous section, we presented the different methods used to model the evolution of OSN over time. We have seen that most of these works attempted to develop generative models for large networks that reproduce the properties revealed by the analysis of different online networks, such as their scale-free nature in degree distribution, their high clustering coefficient, and their low average shortest path length that separates nodes (so called the *small-world* phenomenon [122, 123]). However, studies on understanding the evolution of OPNs are still young and only a few works were devoted to that, while studying personal networks dynamics allows to go back to the processes that engendered the structures, the movements and particularities instead of seeking patterns and fixed forms in the overall network structure that do not capture the dynamics nor the mechanisms that generate it [104].

Personal networks' evolution has already been studied in social sciences, as done by Bidart et al. in [21] in order to explore the relation between the evolution of personal networks of a set of young people and life events by the way of a longitudinal survey on data collected via repeated interviews of the same set of individuals made every 3 years. Next, we discuss the available literature that has dealt with the evolution of personal networks in the online environment.

In [8], Arnaboldi et al. aimed to discover how the size and the structure of personal networks' layers change over time based on Dunbar's definition [119], where a layer is composed of the alters having the same degree of intimacy with the ego. In this work, the authors studied Twitter data in order to assess whether offline personal networks' properties are also valid in Twitter's OPNs. Thus, they outlined that the number of active relationships maintained by the ego remains constant due to the limited cognitive capacity of human brain. Furthermore, they noted that when the ego joins the OSN, the number of ego's relationships shows an important burst that converges to a constant value. A small set of these relationships is strong and is maintained over time, while most of them become weaker shortly after their creation. The later has also been observed by Viswanath et al. in [126] where the authors analyzed the evolution of the Facebook interaction graph and found that personal networks links are rapidly activated and deactivated and that the strength of those links decreases over time soon after they were created.

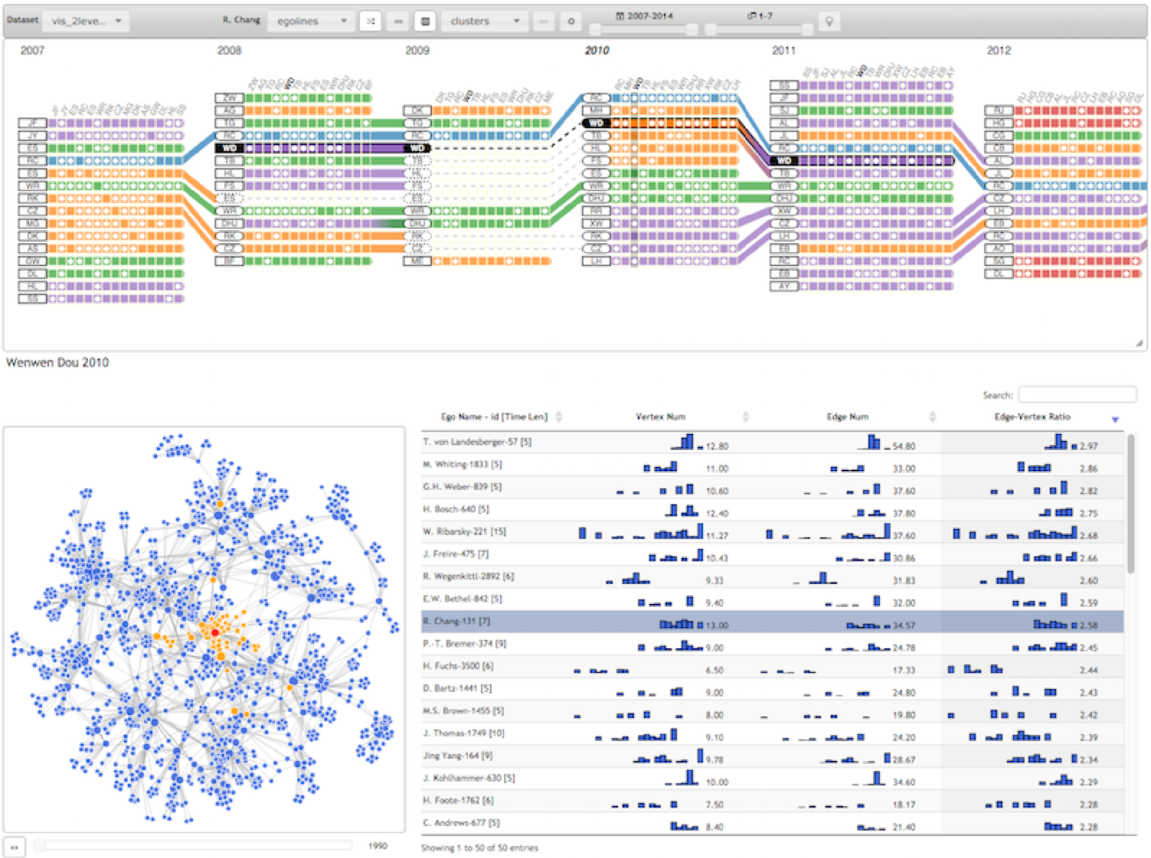


Figure 2.21: EgoLines visualization tool.

Although these conclusions can be helpful when modeling the evolution of relationships' strength in personal networks, they remain limited since they do not reflect how OPNs graph structure is affected over time when nodes and edges are added or deleted. Moreover, by considering only the relationships between the individual (ego) and his direct connections (personal networks with $k = 1$), these studies restrict the information about the evolution of the personal network since many times the evolution appears not on the level 1 but on levels 2 and above.

Visualization techniques help in understanding the evolution of OPNs. In that respect, in [113] the authors propose techniques to visualize large scale personal networks evolution by considering the data as continuous streams. The visualization software *EgoLines*, presented in [136] (Figure 2.21), proposes a dynamic analysis of personal networks. These tools allow to isolate a personal network and analyze its evolution with a visual support, but they are lacking the capabilities of performing massive scale analysis. The system *EgoSlider* [132] (Figure 2.22) allows to analyze a set of dynamic personal networks by summarizing their properties at the network, in-

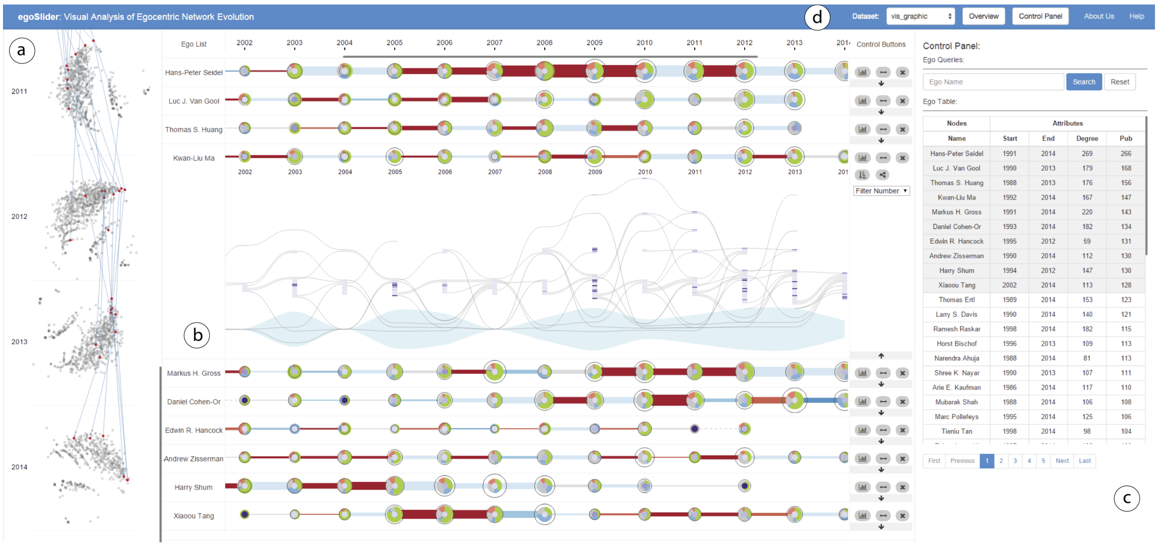


Figure 2.22: EgoSlider visualization tool.

dividual or temporal-based level. Unfortunately, the integrated metrics (e.g. number of alters, edges between alters, etc.) are limited and provides only generic information about the OPN.

2.5 Co-authorship networks topology and evolution studies

A co-authorship network is a social network made of researchers connected with each other if they have co-authored one or more papers together. These networks were studied by different disciplines including physics, bio-medical research and computer science.

The co-authorship networks are of general interest for understanding the topological and dynamical laws governing complex networks, as they represent one of the largest publicly available computerized social networks. One of the very first studies conducted on co-authoring networks was performed by Newman in 2001 [97] and refined in 2004 [100]. These two first works studied the structure of the network by computing the number of authors, the mean papers per authors and the authors' degree. Three of the main observations were outlined in this paper. First, the networks are scale-free with some deviations because the degrees' distribution follow a power law with some deviations. Second, the networks are small world networks as the average separation of the network (average of length of the shortest path between pairs of nodes) is of 6. And last, the networks have a very strong clustering effect given that

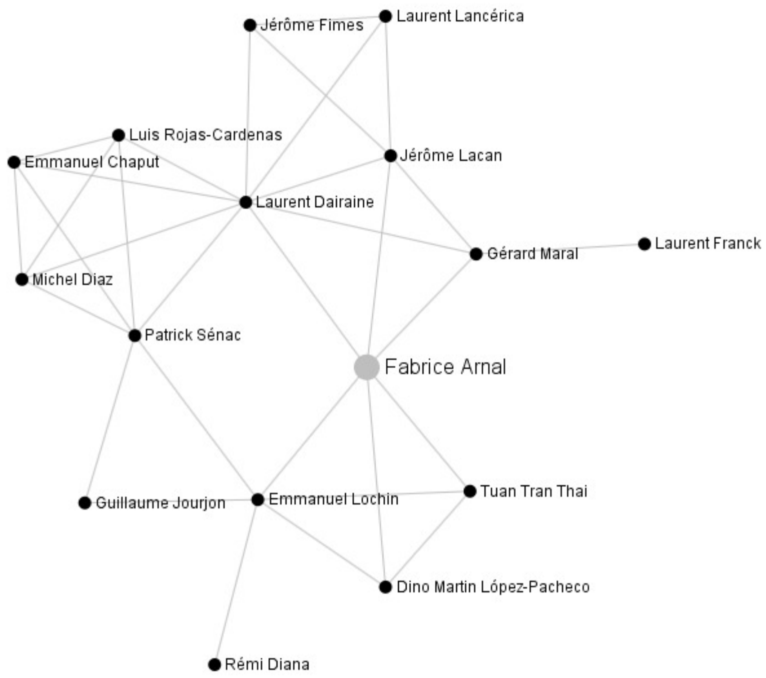


Figure 2.23: Personal network of author "Fabrice Arnal".

the global clustering coefficient or transitivity index [99] is very high. Later, other works studied the structure of static co-authoring networks producing comparable results, such as [133], [127].

In his studies, Newman was only interested in the study of the structure of the static co-authoring network. The work of Barabasi et al. in 2002 [15] was the first one to study the evolution of such networks where the evolution of a collaboration network is seen as the possible addition of new nodes and new links in the network. In this work, topological characteristics of two collaboration networks in Mathematics and Neuro-Science were discovered. Indeed, they found that these networks are scale-free and their evolution is driven by the *preferential attachment* (explained earlier) and that the average degree of the nodes increases over time while the node separation, given by the average shortest path between pairs of nodes, decreases. A model that captures the network's time evolution was proposed. Another work in [74], found that the growth of the collaboration network for the Computer Science field is governed by the scale-free degree distribution and the small world phenomenon.

The studies presented above dealt with the co-authoring networks in their totality when studying its structure or evolution.

Co-authorship personal networks were also of interest for a set of studies that aimed to discover static co-authoring ego-networks properties. For example, the re-

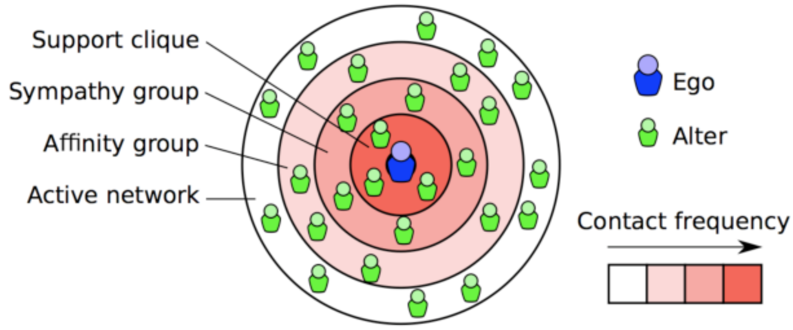


Figure 2.24: Dunbar circles [8].

cent work in [11] aimed to discover if the structure from sociological ego-networks following Dunbar's definition [8] (which represents an ego-network' alters set into four layers, each layer reflects the intimacy level alters have with ego, as represented in Figure 2.24) are present in the co-authoring ego-networks. In this work, they base their research on the limitation imposed on humans by cognitive and time constraints. Notice that here an ego-network of an author is limited to its co-authors and without integrating the links between its co-authors (as in Definition 2.3.1) weighing each link with a measure of strength of their collaboration.

Another work in [103] was interested in analyzing the relationship between research impact and the structural properties of personal co-author networks. In contrary to the considered personal network in [11], the links among an author collaborators (alters) were part of the personal network (as in Definition 2.3.3). A set of measures including centrality measures as the betweenness and the degree were performed on a set of 500 authors with some other measures at the overall network level (average clustering coefficient, density, average path length). They found that in sparse ego-networks characterized by a high ego betweenness centrality, the central authors have more control on their collaborators being more selective in their recruitment and concluding that this behaviour has positive implications in the research impact.

A similar work in [54] provided, after the visualization of the co-authorship network, a set of descriptive statistics at three levels; i.e. the individual level where ego-networks according to Definition 2.3.3 were considered, the group level and the network level. Then, they developed and tested seven hypotheses associated to the researchers' embeddedness in the co-authorship network with the number of the researcher's citations. For example, one of the seven hypothesis suggest that the higher an author's degree centrality, the more that author is cited. Finally, the work by

Chakraborty in [27], proposed an approach to automatically detect circles in ego-networks where a circle represents a densely knit community of researchers. The considered ego-networks follows Definition 2.3.2, i.e. the links between the ego and its co-authors were excluded from the ego-network. Other similar research exploring ego-networks structure in co-authorship network can be found in [1], [45], [22], [133], [30].

While the structure of personal co-authorship networks has been investigated over the set of studies cited in the previous paragraph, few works were devoted to the understanding and modeling of their evolution over time. This question was addressed in [98] by Newman for estimating the number of neighbors at distance 2 from the ego (2-level ego-networks) based on the degree distribution and clustering coefficient estimated on the 1-level personal network of a given author. The proposed formulas were able to make accurate estimates, on real social networks, of how many people your friends will be friends with.

2.6 Conclusion

In this chapter, we presented offline and online social networks and highlighted the differences between them with a projection on both overall OSNs and OPNs. We gave an overview of SNA' concepts for these two related but different perspectives of social networks. We demonstrated that the available literature is lacking of adequate definitions that capture the diversity of today's OPNs. In the next chapter (Chapter 3), we provide formal definitions for different types of OPNs. Our contribution is to provide inclusive and usable definitions that suit different needs.

In addition, as stated earlier, we are interested in the evolution of OPNs and at the best of or knowledge there is no evolution model dedicated to these particular networks and we believe that the existing literature is missing such model. Thus, understanding how online personal networks are evolving is still missing in the current literature, while social networks evolution in general was widely addressed and many models were proposed.

Moreover, from the set of evolution models for OSNs we discussed in this chapter, it is clear that all these models are exploiting properties about the whole OSN structure and might not work when considering OPNs. Indeed, we will see in Chapter 4, where we report the results of an experimental analysis over a large set of real online personal networks by the mean of the computation of metrics that characterizes their structure, that OPNs show different properties.

Chapter 3

Online Personal Networks Definitions: A proposal

3.1 Introduction

In the previous chapter, and precisely in section 2.3.3, we presented an overview of how personal networks are defined in the literature and discussed two sets of definitions for: *1-personal networks* and *k-personal networks*. On the one side, 1-personal networks constitute the most common way to define personal networks in the existing works restricting an individual's personal network to its immediate acquaintances. On the other side, a couple of works used the second set of definitions for k-personal networks which constitute an extended form in the sense that more people are part of an individual's personal network, who are not necessarily directly connected to him/her. However, we outlined important limitations of all these existing definitions. We remind them in the following:

- Restricting an individual's personal network, to its direct connections as given by the definition set of 1-personal networks constitutes an old model attributed to social sciences. The characteristics of nowadays OSNs have changed the way people interact, making possible a communication with others that are not necessarily in our immediate circle of friends. In addition, restricting a personal network at the first level might hide useful information when trying to understand a variety of research problems, as how the information flows within the personal network, how people are getting influenced, or how the personal networks evolve over time. So, extending the current definition is a necessity.

- Still, in social sciences, studies on personal networks concern most of the time kinship, friends, acquaintance relations which are symmetric relationships, which means that if individual a is friend with individual b, then individual b is also friend with individual a. Users' relationships in OSNs are not always symmetric as it is the case for example in Twitter where the users follow each other, but the follow relationship can be in only one direction. More precisely, for a given user u we might have two distinct types of relationships: the first one for the users that user u is following (*followees* of u), and the second for the users that follow the user u (*followers* of u). Thus, we need to define a personal network accordingly to the type of the relationship among users.

- The relationships in real OSNs can be quantified with a weight reflecting a given characteristic of the underline relationship or the exchange between users. Thereby, when considering OPNs we should also take into account that the links might be weighted. This allows us to capture links of varying types and strengths reflecting more about a relationship than just saying that it either exists or not. For example, if we take two Facebook users, the weight of their connection given by the frequency of contact between them can tell us about how close to each other these two users are.

In this chapter, we introduce a new set of definitions; the definitions are flexible enough to cover current OPNs, while they remain easily extensible for the future. We are presenting three definitions:

1. the first one focuses on undirected OPNs
2. the second definition is for directed OPNs
3. the third definition extends the first and the second and defines weighted OPNs where the links are enriched by a weight representing a given property.

3.2 From Dunbar's model for personal networks to new definitions of online personal networks

Let's consider the Dunbar model for 1-personal networks given in Figure 3.1. In this model a personal network is made of layers where each layer consists of a set of users that are the alters directly connected to the ego. Each layer in Dunbar's model reflects the intimacy level alters have with ego. The inner layer is composed of individuals who are the most intimate with the ego, while the outer layer is made of those who are less intimate with the ego.

Based on this model, instead of having layers built on the level of intimacy between the ego and its alters, we consider them as referring to the distance of the ego with the rest of the nodes in the personal network computed as the shortest path length. Thus, the first inner layer groups 1-level alters that are directly connected to the ego, the second layer contains alters at the second level that are not directly connected to the ego but to his 1-level alters in the former inner layer and so on. An illustration of this modeling is given in Figure 3.1. Notice that, for instance, an alter that is in the outer layer is at distance 4 from the ego which means that there is no other way for reaching that alter by traversing from the ego to an alter in layer A then to an alter in layer B then to an alter in layer C then to the final alter in layer D rendering the distance less than 4. However, many different ways of distance 4 might exist from the ego to that alter in layer D through layers A, B, and C.

With this representation, we can see that we are extending the notion of a personal network in the sense that not all the alters are necessarily connected immediately to the ego. We can define this way 1-personal networks, but also 2-personal networks, 3-personal networks, 4-personal networks, and so on.

We give bellow, a formalization of this modeling by distinguishing 3 different definitions. The 3 definitions depend on the type of the connections inside the personal network. The first one concerns undirected personal networks where the connections are represented by symmetric edges. The second definition groups two distinct types, one type for representing connections emanating from the ego given by outgoing edges and the second type for representing connections having as target the ego given by incoming edges. At last, the third definition corresponds to the case where personal network connections are characterized by a given property and thus weighted edges are used to represent such connections.

3.3 Undirected online personal networks

Let's consider an online social network given by a graph $G(V, E)$ where V is the set of nodes representing the social actors and E is the set of (undirected) edges representing the links between them. Next, we give the definition of an undirected personal network of a given user $e \in V$.

Definition 3.3.1 *We define a k-*undirected personal network $k - PN^e$ *of an ego individual e as being a sub-network of the online social network defined by $G(V, E)$ and composed of the ego and the individuals who are connected to it directly or indirectly*

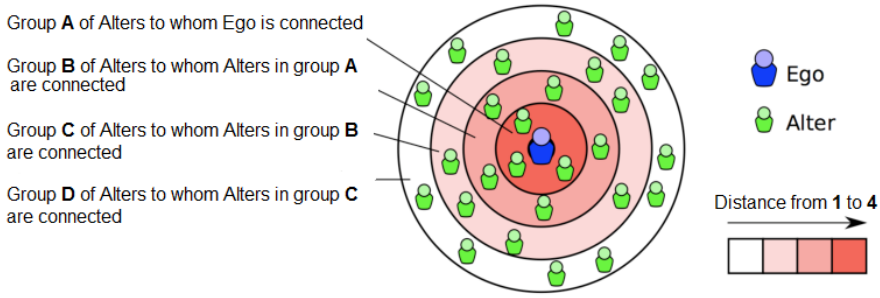


Figure 3.1: From Dunbar circles to OPNs definition.

(named alters), and of all the connections between e and his alters, and between the alters. The personal network $k - PN^e$ is defined by:

$k - PN^e = G'(V', E')$ *with* $V' \subset V$, *and* $E' \subset E$

$V' = \{u \in V \mid d_G(e,u) \leq k \ \wedge \exists v \in V,\ d_G(e,v) = k\} \cup \{e\}$

$E' = \{\{u,v\} \in E \mid u \in V' \wedge v \in V'\}$

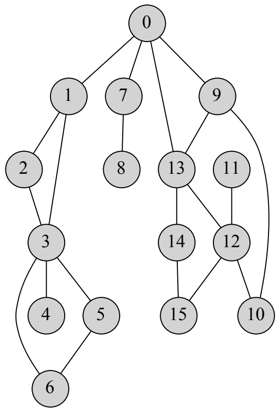
where;

- V' *represents the set of nodes composed of the ego node* e *and all nodes that are connected to* e *via a shortest path of maximum length* k*, with the condition that at least one node* v *is at a* k *distance from* e*,*

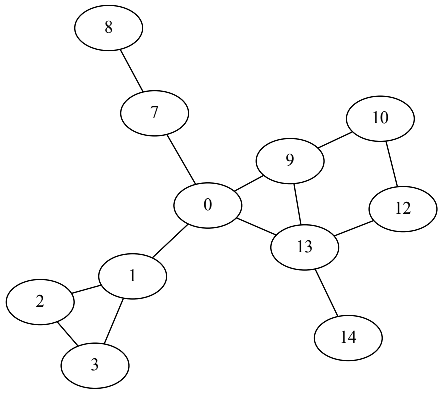
- E' *is the set of all edges connecting the couples of nodes in* V'*. The shortest path is computed by* $d_G(e,u)$ *as the number of edges contained in the shortest path connecting* e *to* u*.*

It is important to note that we use here the function d_G in order to restrict the selection of set of nodes of the personal network of an ego e; nevertheless, once the set of nodes V' selected, we will consider in the personal network all the existing edges between the selected nodes. This means that the resulting network might contain nodes that are connected to the ego e via a path longer than k ($d_G > k$), but via a shortest path shorter or equal to k ($d_G \leq k$). The function d_G will have the same role in the following definitions.

Definition 3.3.1 is close to the one given in [62] for Twitter, where Maira et al. analyzed information diffusion, without the restriction of excluding links between last-level alters. Compared to the definition given by [121] for k-personal networks including ego-alter links but excluding alter-alter links that we presented in section 2.2.2, our definition considers in addition to alter-alter links, the links between the ego node and its direct alters as part of the ego-network. This is important if we aim



(a) Example of an undirected OSN.



(b) Extracted 2-personal network of ego node "0".

Figure 3.2: Undirected Personal Network Example.

for example to study information flow between nodes in the OPN, then we need to have the ego-alter connections. Thus, our definition expands the notion of a personal network allowing to better represent diverse current OPNs unlike [62], where the definition was solely for Twitter.

Example 22 *Based on our new definition, we can extract from the undirected network in Figure 3.2a, the personal network of the ego node* 0 *with k=2. The set of nodes V' will be composed of:*

- *the ego node* 0.
- *all the nodes that are at distance* ≤ 2 *from ego node* 0 *with the condition that there is at least one node that is at distance* $= 2$ *from ego node* 0*. Thus, nodes 1, 7, 9, 13 are part of V' since they are at distance 1 from ego node* 0*, in addition to nodes 2, 3, 8, 10, 12, 14 that are at distance 2 from ego node* 0.

The set of edges E' is then composed of the links that existed in E between nodes in V'.

The resulting 2-personal network is given in Figure 3.2b. This ego-network consists of: the node set $V' = \{0, 1, 2, 3, 7, 8, 9, 10, 12, 13, 14\}$*, and the link set* $E' = \{\{0, 1\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{0, 7\}, \{7, 8\}, \{0, 9\}, \{9, 10\}, \{0, 13\}, \{9, 13\}, \{10, 12\}, \{12, 13\}, \{13, 14\}\}$.

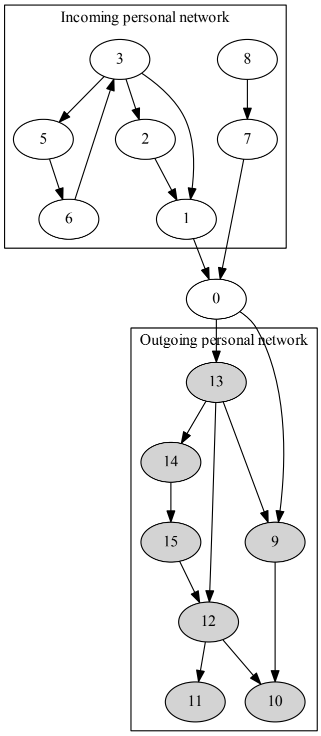


Figure 3.3: Directed (incoming and outgoing) 3-personal network of ego "0".

3.4 Directed online personal networks

The previous definition introduces undirected ego-networks allowing to describe their nodes and connections. In these networks, the connections are symmetric (i.e. in Facebook, if you have a friend in your list, this person will have you in his/her list), but if we take Twitter as example, the connections are not following anymore a symmetric model as you can follow someone without him/her following you back. Thus, in case of Twitter, we have a directed relationship and so, we propose a definition based on directed graphs. To this end, we need to define two distinct concepts: incoming and outgoing OPNs.

The incoming OPN of a given individual e is the network made of e and other individuals in the network such that it exists a directed path starting from each of these individuals and ending at the ego e.

Similarly, the outgoing OPN of an ego e is the network made of e and other individuals in the network such that it exists a directed path starting at the ego e and ending at each of these individuals. The union of the incoming and outgoing personal networks of e gives its full directed OPN. Next, we give the formal definitions of incoming and outgoing OPNs.

3.4.1 Incoming online personal networks

Definition 3.4.1 *An k-incoming personal network $k - PN_{in}^{e}$ of an ego e is a directed sub-graph $G'(V', E')$ of the directed graph $G(V, E)$, composed of the set of individuals u that are connected to e via a shortest path of maximum length k, where u is the start node and e the end node, and the set of links composed of the links being part of the incoming paths to e:*

$k - PN_{in}^{e} = G'(V'_{in}, E'_{in})$ *with* $V'_{in} \subseteq V$, *and* $E'_{in} \subseteq E$

$V'_{in} = \{u \in V \mid d_G(u, e) \leq k \wedge \exists v \in V,\ d_G(v, e) = k\} \cup \{e\}$

$E'_{in} = \{(u, v) \in E \mid u \in V'_{in} \setminus \{e\} \wedge v \in V'_{in}\}$

where:

- V'_{in} *is composed of e and all the nodes (1 or more) with the condition that for each node, there exists an incoming (i.e. a path for which the end node is the ego) shortest path of maximum length k from the given node to e,*
- E'_{in} *is the set of directed edges linking nodes in* V'_{in}.

We notice that, in the definition above, we cannot have cycles involving the ego node. This is guaranteed by the exclusion of the ego from the starting endpoints of the edges in $E'_{in}(x \in V'_{in} \setminus \{e\})$. This avoids the confusion with outgoing edges having as starting point the ego that forms the outgoing OPN that we will define in the next subsection. However, cycles that do not involve the ego are possible inside the incoming OPN.

Through the definition of an incoming ego-network, we can model the flow of the information *explicitly received* by a particular user (the ego). For example, we could be interested in investigating whether this particular user is getting information from conflicting sources or if the obtained information affects its participation in various communities or even we could determine the path followed by the information. It is important to note that this type of analysis was not possible with the definitions presented in Section 2.3.3 of Chapter 2 due to the lack of definition of directed ego-networks.

Example 23 *Based on the graph in Figure 3.3, the directed sub-graph with white nodes and directed edges situated in the upper frame on top of the ego node* $e = 0$*, in addition to the ego node* $e = 0$ *and its incoming edges, corresponds to an incoming ego-network using the definition of* $k - PN^e_{in}(V', E')$ *with k=4. It is given in Figure 3.4.*

The node set V'_{in} *is composed of:*

- *the ego node 0*
- *nodes 1 and 7 that are at distance 1 from ego node 0*
- *nodes 2,3, and 8 that are at distance 2 from ego node 0*
- *node 6 that is at distance 3 from ego node 0*
- *node 5 that is at distance 4 from the ego node 0*

Thus $V'_{in} = \{0, 1, 2, 3, 5, 6, 7, 8\}$*; all the nodes in* V'_{in} *are at a shortest path with maximum length 4 from the ego.*

The set of directed edges E'_{in} *is composed of the directed edges among nodes in* V'_{in}*. Thus,* $E'_{in} = \{(1, 0), (2, 1), (3, 1), (3, 2), (6, 3), (3, 5), (5, 6), (7, 0), (8, 7)\}$.

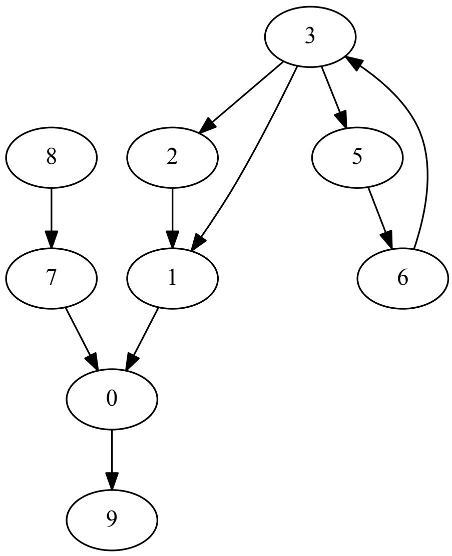


Figure 3.4: The incoming personal network of ego "0", k=4.

3.4.2 Outgoing online personal networks

Definition 3.4.2 *A k-outgoing ego-network PN^{e}_{out} of an individual ego e is a directed sub-graph $G'(V',E')$ of the directed graph $G(V,E)$ composed of the set of individuals u that are connected to e via a shortest path of length of maximum k going from e to u, and the set of links composed of the links being part of the outgoing paths from e:*

$k - PN^{e}_{out} = G'(V'_{out}, E'_{out})$ *with* $V'_{out} \subseteq V$, *and* $E'_{out} \subseteq E$

$V'_{out} = \{u \in V \mid d_G(e,u) \leq k \wedge \exists v \in V,\ d_G(e,v) = k\} \cup \{e\}$

$E'_{out} = \{(u,v) \in E \mid u \in V'_{out} \wedge v \in V'_{out} \setminus \{e\}\}$

where:

- V'_{out} *is composed of the ego node e and all the nodes at a maximum outgoing shortest path length k from it such that there exists an outgoing shortest path from e to each one of these nodes,*
- E'_{out} *is the set of directed edges linking nodes in* V'_{out}.

With the definition of outgoing ego-networks, we provide a way to consider only the outflow of information from a single ego node. This model can be used for understanding the spread of information starting from a specific source.

Example 24 *Given the example in Figure 3.3, the directed sub-graph with grey nodes and edges situated in the frame that is under the ego node e = 0, in addition to the ego node e = 0 and its ongoing edges, corresponds to an outgoing ego-network* $k - PN^{e}_{out}(V'_{out}, E'_{out})$ *with k=3. It is given in Figure 3.5.*

The node set V'_{out} *is composed of:*

- *the ego node 0*
- *nodes 9 and 13 that are at distance 1 from ego node 0*
- *nodes 10, 12, and 14 that are at distance 2 from ego node 0*
- *nodes 11 and 15 that are at distance 3 from ego node 0*

Thus, $V'_{out} = \{0, 9, 10, 11, 12, 13, 14, 15\}$*; since these nodes are at a maximum shortest path length 3 from the ego.*

Also, the set of directed edges is composed of directed edges between nodes in V'_{out}. *Thus,* $E'_{out} = \{(0,9), (0,13), (9,13), (9,10), (13,14), (13,12), (14,15), (15,12), (12,10), (12,11)\}$.

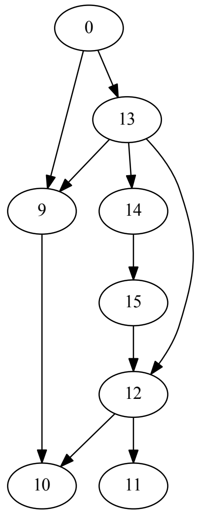


Figure 3.5: The outgoing personal network of ego "0", k=3.

Thus, the k-directed ego-network $k - PN_d^e(V', E')$ of the ego node e is given by the following property.

Property 1 *A directed ego-network $k - PN_d^e(V', E')$ of an individual e is the union of the incoming and outgoing ego-networks:*
$k - PN_d^e = k - PN_{in}^e \cup k - PN_{out}^e$,
which is translated by: $V' = V'_{in} \cup V'_{out}$, $E' = E'_{in} \cup E'_{out}$.

Example 25 *The example given in Figure 3.6 constitute the directed OPN of the ego node 0 with* $k = 4$. *The set of nodes* $V' = V'_{in} \cup V'_{out} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$ *and the set of edges* $E' = E'_{in} \cup E'_{out} = \{(1, 0), (2, 1), (3, 1), (3, 2), (6, 3), (3, 5), (5, 6), (7, 0), (8, 7), (0, 9), (0, 13), (9, 13), (9, 10), (13, 14), (13, 12), (14, 15), (15, 12), (12, 10), (12, 11)\}$.

Thereby, we proposed a definition for directed ego-networks, while distinguishing between incoming and outgoing ego-networks which was not possible with previous definitions and thus capture the complex reality of OSNs.

3.5 Weighted online personal networks

Definition 3.5.1 *A k-undirected and weighted ego-network is the undirected ego-network* $k - PN^e$ *of an ego node e given in Definition 3.3.1 having as an extra element*

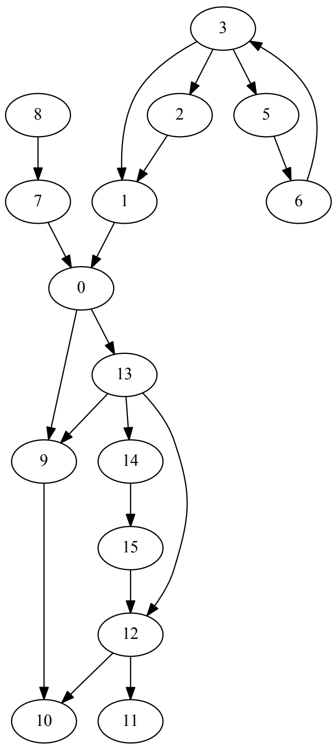


Figure 3.6: Directed 4-personal network of ego "0".

a function f which assigns to each undirected edge in G' a value representing the tie strength of the edge.

$k - PN^{e}_{uw} = G'(V', E') \subseteq G(V, E) \mid V', E'$ *as defined in Definition 3.3.1*

$\exists f,\ f : E' \to \mathbb{R}$

$f(\{u, v\}) = a,$ *where* $\{u, v\} \in E'$ *and* $a \in \mathbb{R}$.

It is important to note that the definition of d_G remains the same as in Definition 3.3.1. This means that we do not use the weights of the edges in order to compute the shortest paths when building the set V', but we only count the number of edges which are part of the shortest path.

Example 26 *The ego-network graph in Figure 3.2b can be seen as a weighted graph corresponding for instance to the 2-personal undirected Facebook graph of the user e="0" and that on each edge between two users in the social graph we have an integer value representing the number of private messages exchanged between them during a*

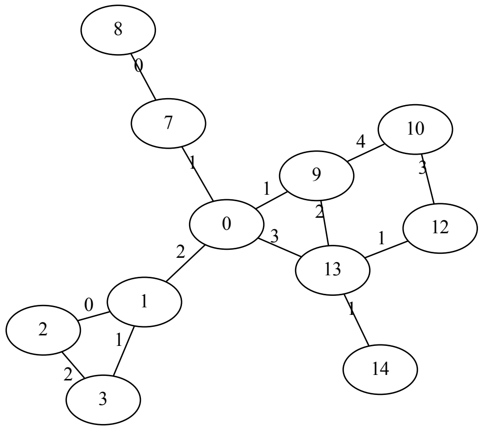


Figure 3.7: Weighted 2-personal network of ego node "0".

certain period. The resulting graph is given in Figure 3.7. For example, users 9 and 10 have exchanged four messages on Facebook.

The same process can respectively be applied to define a weighted *and* directed ego-network. We give bellow, the definition of a directed and weighted ego-network.

Definition 3.5.2 *A k-directed and weighted ego-network is the directed ego-network PN^e of an ego node e given in Property 1 having as an extra element a function f which assigns to each directed edge in G' a value representing the tie strength of the edge.*

$k - PN^e_{dw} = G'(V', E') \subseteq G(V, E) \mid V',E'$ *as defined in Property 1*

$\exists f,\ f : E' \to \mathbb{R}$

$f(\{u, v\}) = a,$ *where* $\{u, v\} \in E'$ *and* $a \in \mathbb{R}$.

Thus, these definitions provide a tool to study topics around the strength of the connections among online social network users.

3.6 Conclusion

In this chapter 2, we proposed a set of formal definitions for online personal social networks (OPNs). The definitions are independent of any application or online social network and capture all diverse existing cases, including weighted, directed (incoming and outgoing) and undirected OPNs. To the best of our knowledge there is no other systematic and formal recording of these concepts in a way that provides a universal framework for studying personal online social networks in general. The definitions were published in [39]. In the following chapters, we use these definitions for extracting and studying personal networks.

Chapter 4

Co-authorship Personal Networks Evolution: Analysis

4.1 Introduction

As pointed out in Chapter2, understanding how online personal networks are evolving is still missing in the current literature, while the evolution of networks as a whole was widely addressed and many models were proposed. In this chapter, we propose to fill this gap by studying the evolution of OPNs by means of 3 different analysis. First, we present analysis that concern 1-level personal network. In these analysis, we wanted to characterize the new links made among alters as the personal networks evolve which is important to understand in order to know how links are organized among alters.

Second, we perform an experimental analysis on a large set of real online personal networks by the mean of the computation of metrics that characterize their structure. We examine how these metrics behave when the personal networks change over time in order to discover the properties driving the evolution of personal networks' structure.

Finally, we present the analysis we did on the same set of real online personal networks to discover the patterns governing these evolving personal networks when considering maximal clique. Indeed, with the second bloc of analysis, we found that cliques can explain many of the metric observations.

Before detailing these 3 different analysis, we present next the data on which we worked to perform the analysis.

4.2 Data set description

4.2.1 Data collection

We have chosen to perform the analysis on data from DBLP (Digital Bibliography & Library Project) available on *aminer* plateform[1]. The dataset we selected is a DBLP citation network in its 7th version initially composed of 2.244.021 papers and 4.354.534 citation relationships collected on 2014-05-25. This dataset groups a set of 10 instances depending on the domain of the papers (Artificial Intelligence, Computer Graphics and Multimedia, Computer Networks, Database with Data-mining and information retrieval, High Performance Computing, Human Computing Interaction and Ubiquitous Computing, Information Security, Interdisciplinary Studies, Software Engineering and Theoretical Computer Science). We selected to work on the DBLP citation network on the domain of Computer Networks containing 16222 papers presented in a text file where each paper is represented with some information in the following format:

#* — Paper title
#@ — Authors
#t — Year
#c — Publication venue
#index — Index id of this paper
#% — The id of references of this paper
#! — Abstract

An example for a paper is given bellow:

#*Introduction to the User Requirements Notation: learning by example.
#@Daniel Amyot
#t2003
#cComputer Networks
#index793184
#%118406
#%1099051
#%1127648
#!

In the text file containing the papers, an empty line separates the papers.

[1]https://aminer.org/citation

The issue we had with this instance of dataset of papers in Computer Networks domain is that the majority of references of papers are not necessary part of the field of Computer Network and thus are not represented in the data and only 5.980 were there. In order to have a complete data, we extracted and added the entries corresponding to those referenced papers from the other domains but also their respective references that are not represented in the initial citation network in Computer Networks. At the end we obtained a dataset of 38.134 papers.

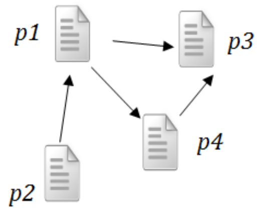
4.2.2 Data storage

Starting from the file containing the 38.134 papers, we built an SQL database named *dblp*. The database groups a set of relational tables. In the following, we describe these tables and their corresponding attributes:

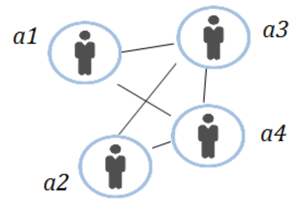
- *papers: idPaper, yearPaper, title, venue*
- *citations: citedPaper, idPaper*, where each entry represents a citation relationship. We obtained in total 60.950 citation relationships.
- *authors: idAuthor, nameAuthor*. The idAuthor attribute was created since only the names of authors are given in the original data file. The table contains 23.010 authors.
- *author_paper: idAuthor, idPaper*. An entry here links each author to the paper she/he participated in. A given author might participate to more than one paper.
- *coauthors: idAuthor, idCoAuthor, weight*. In this table, we represented the co-authoring relationship among the authors of the same paper. Each record reflects a co-authoring relationship between two authors of a same paper with an attribute *weight* representing the number of papers co-authored by these two authors. This table contains 101.204 co-authoring ralationships.

Using this database, two distinct networks can be built:

1. **The citation network** based on the *citation* table that stores citation relationships. The corresponding graph is a directed graph where nodes represent the papers and a directed edge from a given paper $p1$ to a given paper $p2$ reflects the fact that paper $p1$ cites paper $p2$. An example is given in Figure 4.1a.



(a) Citation network is represented with a directed graph.



(b) Co-authorship network is represented with an undirected graph.

Figure 4.1: Citation and Co-authorship Networks.

2. **The co-authorship network** is built starting from *authors* table. The corresponding graph is an undirected graph where nodes represent the authors and an undirected edge between two given authors $a1$ and $a2$ means that authors $a1$ and $a2$ have written at least a paper together. An example is given in Figure 4.1b.

In the rest of this chapter, we focus on the co-authorship network.

4.2.3 The co-authorship network and its evolution

Co-authorship network to study. As described in the previous section, we constructed the network of co-authorships with connections between pairs of authors who share at least one publication in the field of Computer Networks. The corresponding co-authorship network graph that is undirected, since relationships are symmetric, was initially composed of 23.010 nodes (authors) and 101.204 edges (co-authorships).

The co-authorships were made on papers published in the period 1971-2013. The first analyse of this graph have allowed to identify a set of 2348 separated components. The largest graph component that constitute the giant component of the graph is composed of 13854 nodes (authors) and 32946 edges (co-authorships). Figure 4.2 gives an overview of the giant component. The next large component have a size of 55 nodes followed by smaller components. Thus, the co-authorship network graph we are studying in the following is made of authors who do not belong to independent communities but belong to the giant component of the whole graph of scientific collaborations in Computer Networks area from 1971 until 2013.

Co-authorship network evolution. The evolution of a scientific collaboration network from time t to time $t+1$ consists in the addition of new authors (nodes) that join the network and create new collaborations (edges) with existing authors

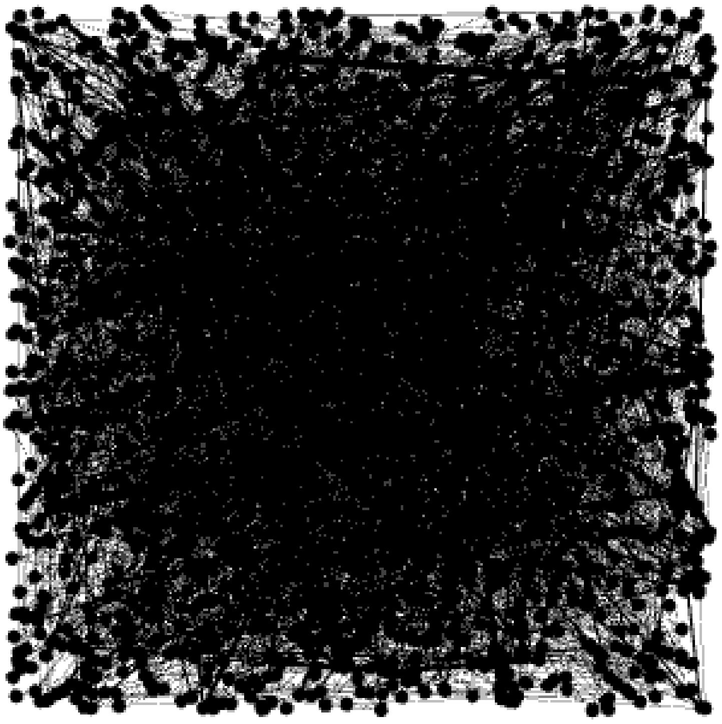


Figure 4.2: Giant Component of Computer Networks Co-authoring Graph.

that were in the network at time t. New edges can also appear between two authors already in the network at time t, as well as between two new authors joining the network at time $t + 1$. In this context, the t parameter corresponds to the year in which a co-authorship was established (year of the published paper). Thus, one can capture the co-authorship network at a given starting time point and observe how this network evolves by the addition of nodes and edges until a final time point.

Co-authorship personal networks evolution. In our case, the aim is not to study the evolution of the collaboration network as a whole, but to study the evolution of the personal collaboration networks of a set of individual authors (egos) over time in an effort to understand for example if the evolution of the personal networks shares common characteristics, patterns or trends or if the behavior is specific to some particular personal networks. The personal co-authoring network of a given author based on Definition 3.3.1, given in Chapter 3, of an undirected personal network, is made of that focal author (ego) and all those authors with whom she/he co-authored at least one paper in addition to their respective co-authors and so on depending on parameter k. Thus, a personal co-authoring network of a given author is extracted from the entire co-authoring network in Computer Networks field stored in *dblp* database we described above.

Dataset	Year of first appearance	Number of analyzed authors	Time window
dataset 1	2004	560	2006 to 2013
dataset 2	2005	594	2007 to 2013
dataset 3	2006	1096	2008 to 2013
dataset 4	2007	1029	2009 to 2013
dataset 5	2008	1256	2010 to 2013

Table 4.1: The description of the different parts of the used DBLP data set.

Dataset of co-authoring personal networks to study. In order to obtain a cohesive set of co-authoring personal networks for studying the evolution, we first have fixed the period on which we want to study the personal networks evolution to be between 2004 and 2013. Thus, we have split our data into 5 parts; each part is composed of a set of authors (egos) that have started publishing on a given year. For example:

- $dataset_1$ holds authors that had their first publication in 2004,
- $dataset_2$ contains authors that had their first publication in 2005, and so on.

In Table 4.1, we give a description of the 5 datasets. The third column of the table gives the number of authors analyzed per dataset, while the last column contains the time window on which we studied the evolution for each dataset.

We started studying the evolution of each dataset two years after the dataset's authors joined the network since we have observed that during the first two years a considerable fraction of personal networks remains unchanged (do not evolve in terms of nodes' and edges' addition).

In the next section, we describe the methodology we used to analyze the described data and present the obtained results and the corresponding conclusions for the three following types of analysis:

1. Analysis of 1-level co-authorship personal networks evolution via links characterization
2. Metric-driven analysis, which relies on most of the metrics we presented in Chapter 2, focusing on understanding personal collaboration networks' evolution.
3. Analysis of co-authorship personal networks evolution via cliques.

4.3 Analysis of 1-level co-authorship personal networks evolution via links characterization

4.3.1 Methodology

As a first step to understand the dynamics of personal networks, we have focused on 1-level personal network in order to study the types of new edges that can be created among the alters during the evolution from time t to time $t+1$.

Given the evolution of the co-authorship network that is characterized by only the addition (deletion is not possible) of new nodes (authors) and thus new edges (co-authoring between two authors), every new node at time $t+1$ will connect to the *ego* in a 1-level personal network. Thus, in addition to the edges made between new nodes joining the 1-level personal network at $t+1$ and the *ego*, we identified three possible types of new edges that can be made among the alters within the 1-level personal network:

1. Edge between 2 new nodes: where two authors a_1 and a_2 are added to a personal network at time $t+1$, because they co-authored a paper with an existing author a_3 at time t. Thus, authors a_1 and a_2 join the personal network already connected with an edge.

2. Edge between a new node and old node: based on the same example as for the first type of edges, this time we are referring to the edge that is created between a new author at $t+1$, for example a_1, and an old one, here a_3.

3. Edges between old nodes: the last type of edges that we can identify is the edge made at time $t+1$ between two old nodes a_0 and a_3 existing in the personal network at time t.

In the following, we report the results of the computation of the proportions of these 3 types of edges on the 1-personal networks of authors in $dataset_1$ that started publishing on 2004.

4.3.2 Results

On the set of 560 authors from $dataset_1$, we computed, in each time period (2006-2007, 2007-2008, 2008-2009, 2009-2010, 2010-2011, 2011-2012, and 2012-2013), the percentage of each type of edge for each personal network and averaged the percentages overall the personal networks in a given time period. The results are represented in Figure 4.3. On the period 2012-2013 only a few set of 1-personal network evolve, so, we decided to not present the corresponding graphic.

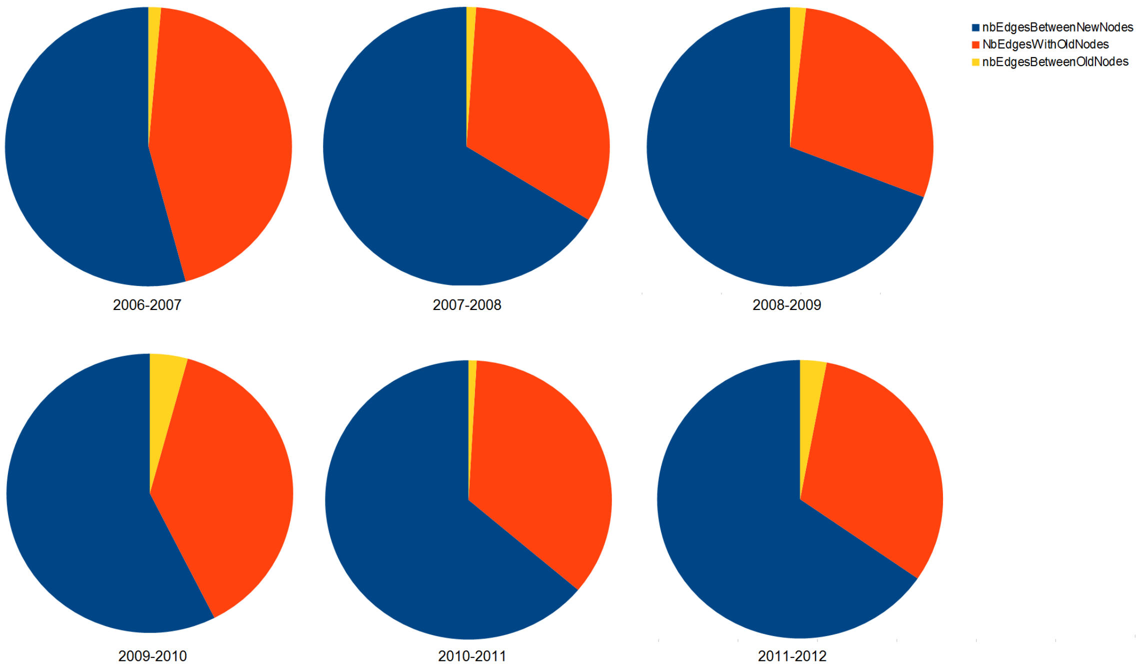


Figure 4.3: Types of edges, connecting the alters, in 1-level personal networks of authors in $dataset_1$.

We can note, for each period, that the proportion of edges between new alters joining the 1-level personal network is the largest with more than 50%, followed by the proportion of edges between a new alter and an existing alters (old) with less than 40%. Finally, edges made between old alters are very few since their proportion does not exceed 4% (the highest proportion registered was from 2009 to 2010 with 3.8%).

To conclude, the analysis of new links among alters in 1-level co-authorship personal networks shows that most of the new links are formed among new coming nodes which reveals that nodes join the network under the form of cliques. Thus, a new collaboration involves most of the time new authors that were not in the network at time t with inevitably the ego author (because we consider 1-level personal networks) and eventually his alters (old nodes) which explains the second important proportion of edges made between new and old nodes. However, we found that it is unlikely for two old authors not connected at time t to connect at time $t+1$ given the very small proportion of links made between old alters.

In the next chapter, we will see that such knowledge is helpful for an evolution model since it allows to guide the creation of new links among nodes in the personal network. However, characterizing the evolution based on edges is not enough and do not capture the evolution of the global structure of personal networks especially if we look at levels above ($k = 2, 3...$). Thus, in the next section, we perform experimental analysis based on a set of metrics that describes better the structure of the personal networks.

4.4 Metrics-driven Analysis

The most common way of analyzing personal networks is attribute-based. Indeed, socio-demographic variables such as age, sex, and race are often collected for personal networks members, and these variables are then summarized as averages or percentages. For example, the average age of alters or, the proportion of alters who are women.

The application of structural analyses that are traditionally used on whole (sociocentric) networks data may prove fruitful. The utility of this approach becomes apparent when the sample of network members considered is relatively large as we will see further in this section.

4.4.1 Chosen metrics and motivations

From the set of metrics available in the literature and from the once discussed in Chapter 2 for capturing the structure of social networks in general and personal networks in particular, we selected those that we believe better represent the structure of online personal networks, but also the ones that can provide some insights or patterns on their dynamics. The choice of these metrics is discussed hereafter.

4.4.1.1 Metrics at ego level

Ego degree centrality. We compute the degree centrality for the ego node (Definition 2.3.8, Chapter 2, Section 2.3.4) in order to capture how the egos' direct connections number is evolving over time.

Betweenness centrality. When the personal network is evolving, we would like to observe whether the importance of the ego is affected (if it decreases or increases) via the computation of its betweenness centrality (Definition 2.3.10, Chapter 2, Section 2.3.4). This will give us more information about the structural changes in the network while evolving, because, for example, an ego with a decreasing betweenness implies that in its personal network alters become more and more connected over the years (thus ego becomes less important over the years).

Local clustering coefficient (Watts-Strogatz clustering coefficient [130]). The local clustering is computed at a node's level to detect the transitivity inside a node's immediate neighborhood (Definition 2.2.5, Chapter 2, Section 2.3.4). Thus, when the personal network evolves, our aim by computing this metric is to observe the changes in the 1st level ego's alters connections rate (given that all 1-level alters are connected directly with the ego) which affects ego's importance.
Effective size. As for the local clustering coefficient, the effective size/efficiency (Definitions 2.3.11 and 2.3.12, Chapter 2, Section 2.3.4) allows us to capture the degree of connectedness between ego's alters and so the loss or gain of importance of the ego when the 1-personal network evolves over time.

4.4.1.2 Metrics at personal network level

Number of nodes and edges.

Definition 4.4.1 *If $G'(V', E')$ is an OPN as defined in Definition 3.3.1, then $n = |V'|$ and $m = |E'|$ are, respectively, the number of nodes and the number of edges composing the OPN.*

We compute the number of nodes and edges to capture the size of the personal network and to detect its change over time as it is modified by the addition of nodes and edges. In this way, we can assess the evolution of the personal network in terms of the number of nodes and edges.
Density. The utility of computing the density (Definition 2.3.7, Chapter 2, Section 2.3.4) on personal networks is to evaluate the connectedness of nodes composing the personal network. A high density means that the nodes are well connected, while a low density reflects the presence of a low amount of edges among the nodes composing the personal network. Thus, the value of the density when the personal network evolves provides us an insight on how their connectedness is changing depending on whether the value of the density increases or decreases.
Global Clustering Coefficient (Transitivity Index). We want to compute the global clustering coefficient on personal networks (Definition 2.2.9, Chapter 2, Section 2.3.4) in order to see how it is structured when considering triplets of nodes and if the *friend of my friend is likely to be my friend* principle, proved for random graphs, is also valid for personal networks we are investigating. The triangle structure was observed as being present in many real social networks and represent one of the main properties that researchers consider when providing models for social networks

evolution. Our aim is to verify if it is also a property that can characterize personal networks, and if so, how?

Average Clustering Coefficient (Average Watts-Strogatz Clustering Coefficient). For personal networks, it is interesting to perform the computation of both clustering coefficients, i.e., the global clustering coefficient (transitivity index) and the average clustering coefficient (Watts-Strogatz average clustering coefficient), which both reflect the personal network clustered architecture, to see if they behave similarly when the personal network is evolving or if they evolve differently over time. Indeed, the Watts-Strogatz average clustering coefficient (Definition 2.2.6, Chapter 2, Section 2.3.4) might transpose more precisely the connectivity of nodes in personal networks because it focuses on the local connectivity, while the transitivity index is computing the connectivity over the whole network.

Degree centrality and average degree centrality. We are interested in computing the degree centrality (Definition 2.2.10, Chapter 2, Section 2.3.4) over all personal network nodes for two purposes:

1. to check the presence of power law distribution for personal network nodes degrees, the importance of this element is discussed hereafter,
2. to compute the average degree of all the nodes inside the personal network.

Power law distribution. The fact that nodes'degrees follow a power law distribution is a property that appears in many real OSNs. For a network, it consists in having few nodes with a high degree and many nodes with low degree. This property implied a new one named the *preferential attachment*, largely used in the majority of evolution models.

In order to verify if the nodes' degrees are following a power law distribution, we compute the degrees (as described in Definition 2.2.10, Chapter 2, Section 2.3.4) of all the nodes in the OPN. Then we verify that nodes' degree distribution takes the form $P(x) = cx^{-\alpha}$, where $P(x)$ denotes the fraction of nodes in the OPN having degree x, c is a normalization constant and α represents the exponent of the power law distribution function. If an OPN has a power law degree distribution, it means that the number of nodes with degree x is proportional to $x^{-\alpha}$ and α is the slope of the distribution and ranges typically around 2 or 3. When α is high, the number of nodes with high degree is smaller than the number of nodes with low degree.

For OPNs, our aim is to check if the power law distribution holds, as in the case of OSNs. This information will allow us to validate whether the evolution models based on preferential attachment are suitable for OPNs and to what extend, or not.

Ego's maximum degree of separation (*k-max*). In the Definition 3.3.1, Chapter 3 of OPNs, the parameter k is used to limit the nodes part of the OPN, since only the alters that are at a maximum distance k from the ego can be part of the k-personal network of the ego. For a 2-personal networks, $k = 2$ and the alters are at a maximum distance of 2 from the ego. Note that by distance between two nodes, we mean the shortest path length between those nodes. By *k-max*, defined below in Definition 4.4.2, we capture the maximum distance between the ego node and all alters reachable from the ego, and we observe how *k-max* changes when the OPN evolves. Moreover, we check if the 6-degree of separation principle [122, 123], validated in real world OSNs [81, 63], is satisfied for OPNs. The 6-degree of separation, known as the *small world* phenomenon, suggests that each pair of nodes inside an OSN is connected via a shortest path of average length 6.

Definition 4.4.2 *The ego's e maximum degree of separation (k-max) in a undirected personal network $G'(V', E')$ is given by $k - max = \max_{u \in E'}(d(e, u))$, where $d(e, u)$ is the shortest path length between the ego node e and each node u reachable from e.*

4.4.2 Methodology

The analysis of the set of selected online personal networks is performed via the computation of the set of metrics described in the previous section over the different time-steps (each time-step corresponds to a year in our case). At the end, we consolidate the observed behavior of the metrics over the personal networks and over the years in order to reach, wherever possible, a common conclusion.

To this end, we use PERSONA (PERSonal Online social Networks' Analytics) platform that will be described in details in Chapter 6, in order to extract from the entire network of collaborations, the desired personal network of a given ego, and then to compute all metrics on the extracted network. To extract a personal network, we need to specify the dataset of which the ego is part (e.g. $dataset_1$), the *year* (possible values depending on the dataset, as presented in Table 4.1) and the k value (we distinguish five values for k, from 1 to 5).

We chose to use for k, values from 1 to 5 because we observed that there are very few personal networks with *k-max* values of 6 and 7, no personal networks with *k-max* between 8 and 13, and for *k-max* > 13, the personal networks join the giant component. This element will be detailed in the next section when discussing the

	Number of computed metrics	Number of k values	Number of authors	Number of years	Number of computations per metric
Metrics computed at 1-level OPNs	3	1	560	8	4480
Metrics computed at k-level OPNs	7	5	560	8	22400

Table 4.2: Summary of metrics' computation methodology on *dataset 1*

evolution trend of ego's maximum degree of separation. This observation means that the ego-alter distance of 5 seems the best threshold up to which one can capture the characteristics of an individual's personal network. As outlined in [76], where the distance 5 from a given central node was set as a limit for analyzing road networks justifying this choice by the fact that a distance beyond five would capture non-local properties, since neighbourhoods would include a large fraction of nodes belonging to the entire network.

Thus, for a given dataset and a specific metric, we will compute all the values of the metric for the authors (i.e. the ego/personal networks) for each *year* value and each k. For example, if we consider $dataset_1$, the density is computed for the personal network of each author inside $dataset_1$ on each year from 2006 to 2013 and for each $k = 1, 2 \ldots, 5$. For the case of the effective size, the local clustering coefficient and the degree centrality, we use only $k = 1$ since these metrics are relevant only for 1-level personal networks. Table 4.2 gives a summary of how we performed metrics' computation on $dataset_1$. The last column of the table shows the number of computations we obtain per metric.

Then, in order to assess the evolution of each metric through the years for the personal networks of the same dataset, we make two types of plots (for each couple $(year, k)$):

- The first plot is a cloud of dots where on the x axis we have the ids of egos (authors) representing the personal networks, and on the y axis we have the value of the metric for each personal network. An example of this type of plots is given in Figure 4.10. The plot represents the value of the density for all personal networks belonging to $dataset_1$, on 2006 for $k = 4$.

- The second type of plot is a bar plot for describing the distribution of the value of a given metric among the personal networks. Each bar describes the proportion of personal networks having that value of metric. As an example, we give in Figure 4.5, the distribution of the density over the personal networks belonging to $dataset_1$, on 2006 for $k = 4$.

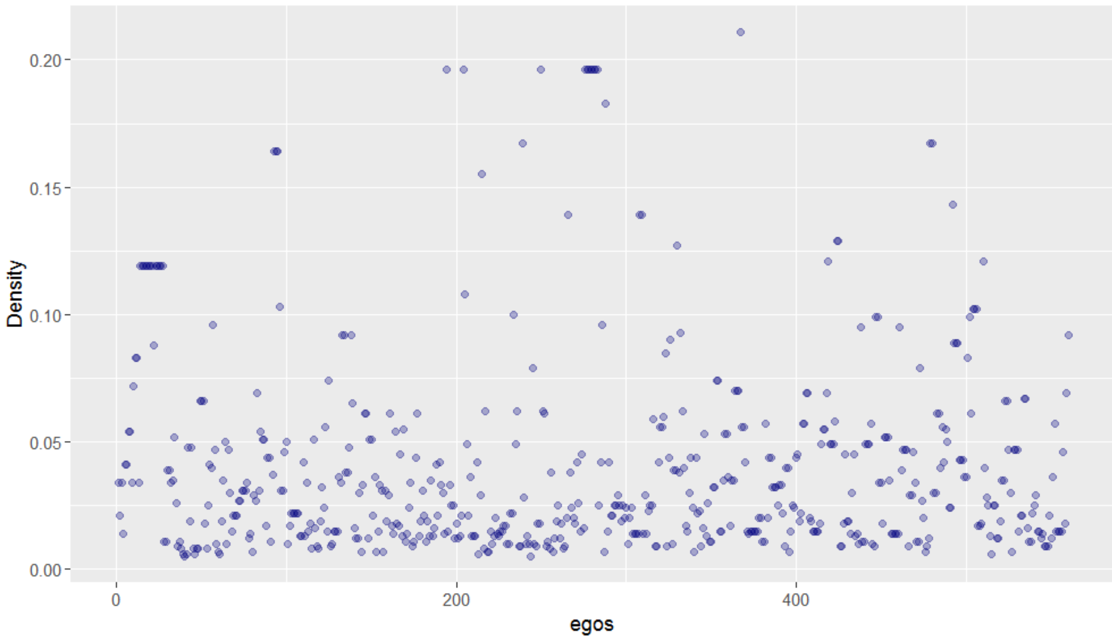


Figure 4.4: Density values on 2006, $k = 4$,
for authors starting publishing on 2004 ($dataset_1$).

By using both types of plots, we can observe the tendency that a given metric has in terms of increasing or decreasing through the years and observe the behavior when k changes.

Then, in order to consolidate and confirm the visual observed evolution trends from the graphics, we performed for each metric a Signed Rank Wilcoxon test at three points in time: 2006, 2008, and 2010 for $k = 2$ to 5. The Signed Rank Wilcoxon test is a non parametric test that compares two paired samples (paired means that both samples consist of the same test subjects). The test essentially calculates the difference between sets of pairs and analyzes these differences to establish if they are statistically significantly different from one another.

More precisely, we consider two populations X and Y of respective size n_x and n_y. We assume that the observations are independent and have an order relation. We want to test the following hypothesis:

$H0$ (equality of the two samples): The probability that an observation of population X is greater than an observation of population Y is equal to the probability that an observation of population Y is greater than an observation of population X: $P(|X| > |Y|) = P(|Y| > |X|)$.

Using the software R, we are able to perform the test by providing the vectors

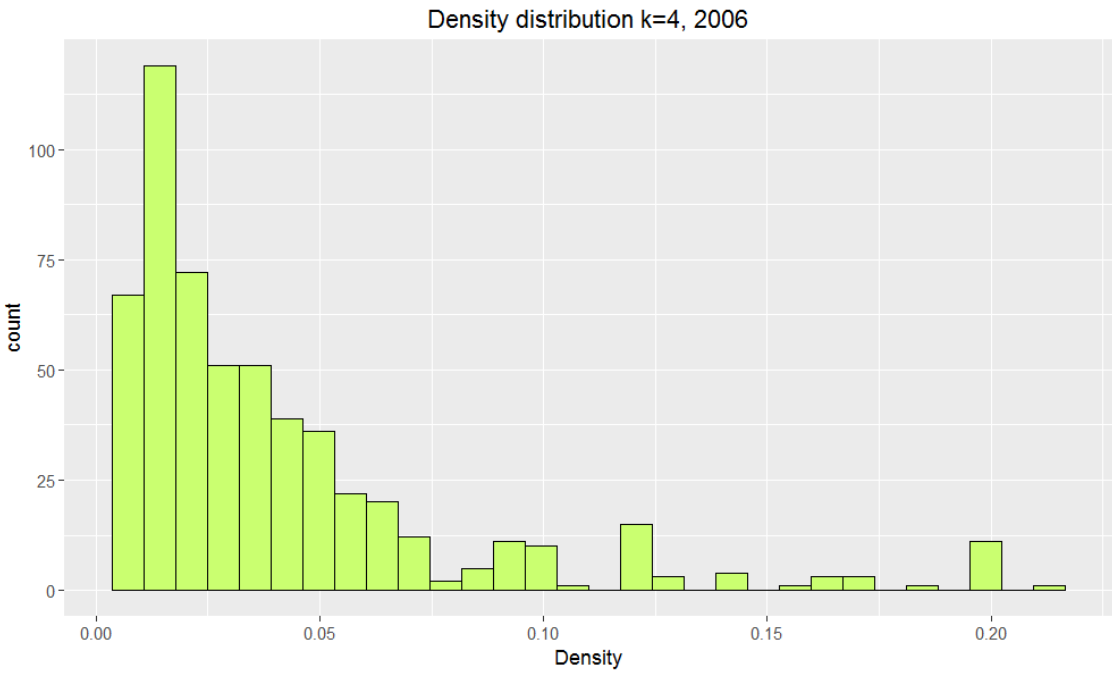


Figure 4.5: Density distribution on 2006, $k = 4$,
for authors starting publishing on 2004 ($dataset_1$).

representing both paired samples. We set the parameter *(paired)* at TRUE. In addition, we can specify the alternative hypothesis to the null hypothesis (that assumes that both samples are equal) which must be one of "two.sided" (default), "greater" or "less" as in the following:

wilcox.test(X, Y, alternative = c("two.sided", "less", "greater"), paired = TRUE)

The output of this function in R consists of V representing the value of the test statistic, and the *p-value* that allows to reject or not the null hypothesis.

In our case, we apply the Wilcoxon test on the values of a given metric for a given k at two points in time t and $t+1$ (vectors X and Y) to confirm whether it increases or decreases using the alternative hypothesis "less" (which means that the metric increases from time t to time $t+1$), and "greater" (which means that the metric decreases from time t to time $t+1$).

In the next section, we provide the results of these analyzes and discuss the observed behavior of each metric.

4.4.3 Metrics evolution trends

Before presenting the findings for each metric, we precise that the obtained results for all the metrics are the same for all the 5 datasets. Thus, in the following, we choose to report the results observed on $dataset_1$ which has the largest time window for studying the evolution.

Number of nodes and edges. In co-authorship networks nodes and edges are only added over time and cannot be removed. By analyzing the co-authorship personal network of $dataset_1$, we found that the distribution of nodes and edges had the same shape whatever k and whatever the year. We give in Figures 4.6 and 4.7 an example of the distribution of nodes and edges respectively, for $k = 3$ on 2009 for authors in $dataset_1$ that started publishing on 2004. In order to have an idea on the size of the personal networks, in term of number of nodes and edges, of the authors in $dataset_1$ and how it evolves, we recap in Table 4.3 the minimum and maximum number of nodes (n) and edges (m) over the personal networks for all k values and all the years. We can notice that the smallest OPN has 2 nodes and one single edges when $k = 1$, and that the biggest OPN is composed of 8650 nodes and the number of edges can reach 22668 when $k = 5$ on 2013.

With analysing the co-authorship personal network of $dataset_1$, we couldn't identify a single rate with which nodes and edges are added over the years. However, the role of edges in shaping the evolving structure of a personal network, will be further clarified as we discuss the rest of the metrics that follow.

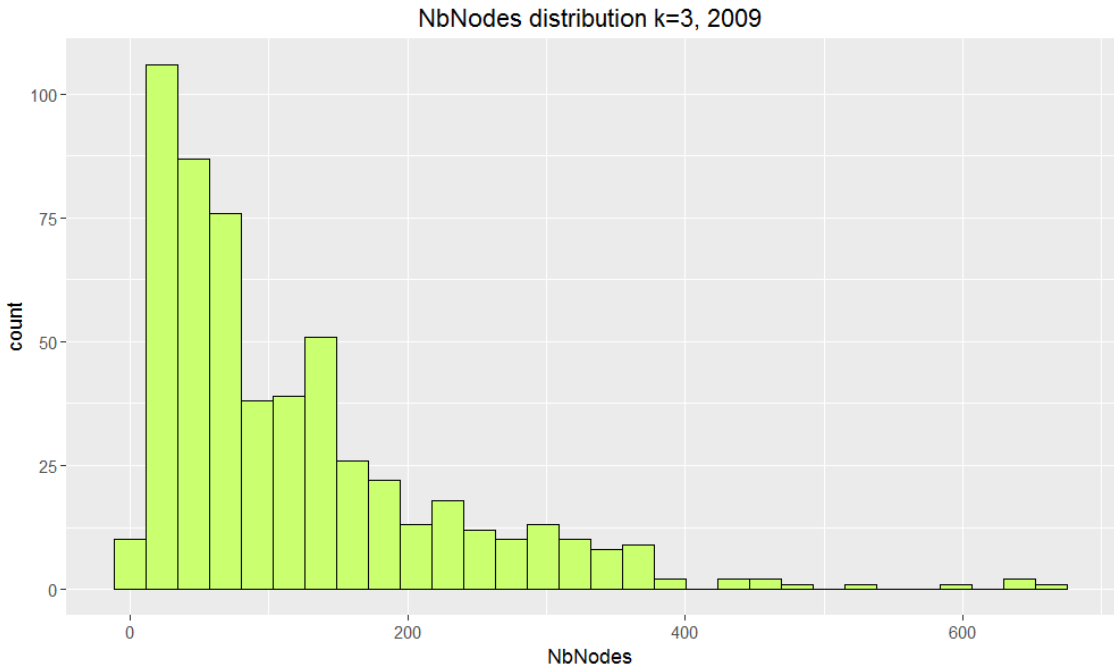


Figure 4.6: Distribution of the number of nodes on 2009, $k = 3$, for authors starting publishing on 2004 ($dataset_1$).

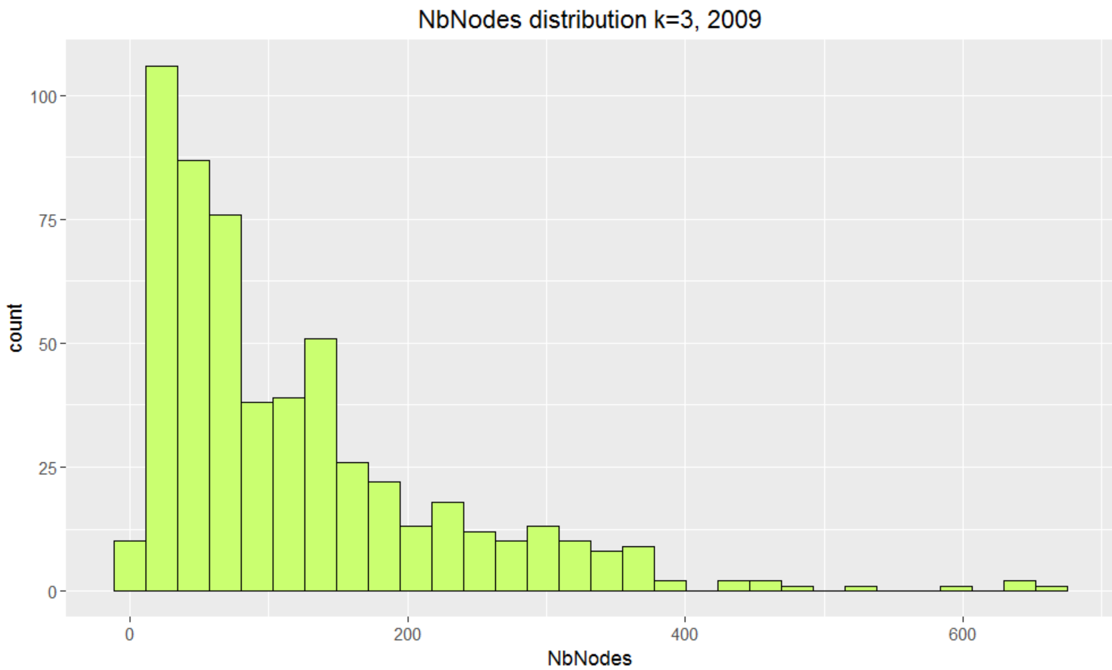


Figure 4.7: Distribution of the number of edges on 2009, $k = 3$, for authors starting publishing on 2004 ($dataset_1$).

Density. The density quantifies how well connected the nodes composing the personal networks are. In the following, we split the observations we draw for density evolution into two parts: the first part concerns the evolution of the personal networks regarding the value of k, and the second part concerns observations over the years.

- **Observations regarding k:** When $k = 1$, which is a particular case as we will see that through the evolution of many of the metrics we are analyzing, we observe a large proportion of personal networks with a density equal to 1, as we can see in Figure 4.8, with density distribution on 2006 (we obtained the same distribution on the other years) which means that all the nodes are connected with each other, forming complete networks. At this level, personal networks are frequently composed of only few nodes.

- **Observations over years:** When k=1, the proportion of personal networks with a density value equal to 1 decreases as the personal networks grow with nodes that are joining them but that do not necessarily connect to all the existing nodes. This can be observed on Figure 4.9 where the number of personal networks on 2009 having a density equal to 1 is about 300, while in 2006 it was about 400 as we can see it in Figure 4.8. Even if in 2013, we still have density

Year	metric	k=1	k=2	k=3	k=4	k=5
2006	min n	2	3	6	8	10
	max n	22	69	197	469	1020
	min m	1	2	5	8	14
	max m	165	328	570	1123	2229
2007	min n	2	3	6	8	12
	max n	22	93	290	689	1476
	min m	1	2	5	8	14
	max m	165	335	721	1604	3414
2008	min n	2	3	6	8	12
	max n	29	119	382	1064	2444
	min m	1	2	5	8	14
	max m	221	378	951	2528	5876
2009	min n	2	3	6	9	13
	max n	33	213	670	2018	4049
	min m	1	2	5	10	19
	max m	221	562	1651	4869	9891
2010	min n	2	3	6	9	13
	max n	33	240	794	2487	5050
	min m	1	2	5	10	19
	max m	269	600	2013	6127	12630
2011	min n	2	3	6	9	13
	max n	47	360	1261	3736	7386
	min m	1	2	5	10	19
	max m	269	1020	3326	10408	19205
2012	min n	2	3	6	9	14
	max n	47	414	1518	4539	8394
	min m	1	2	5	10	19
	max m	269	1054	3906	12213	22149
2013	min n	2	3	6	9	14
	max n	47	425	1561	4676	8650
	min m	1	2	5	10	19
	max m	269	1076	4021	12631	22668

Table 4.3: Minimum and maximum number of nodes (n) and edges (m), $dataset_1$.

value of 1 for a set of personal networks, in general the density value decreases with years.

For $k = 2, 3, 4, 5$, we observed that the density decreases over the years. This tendency is also valid where no personal network has a density equal to 1 and all of them get it decreasing to reach very low values at the last years (Figures 4.10, 4.11 4.12).

The Wilcoxon test confirms the decreasing tendency of the density over years as we can observe it in Table 4.4 where we give the results of the test applied on the density of personal networks between 2006 and 2008, and between 2008 and 2010 for $k = 2$ to 5. In Table 4.4, we give the value of V the statistic value of the Wilcoxon test and the values $p - value_{decreasing}$ and $p - value_{increasing}$ representing the p-value of the Wilcoxon test whether we give as alternative hypothesis "greater" or "less" respectively. We can see that $p - value_{increasing}$ equals always 1 which indicates that the increasing trend cannot explain the evolution of the density, while $p - value_{decreasing}$ that is $<$ 2.2e-16 consolidate the observed decreasing trend as it is significantly lower than 5% (the threshold we use to reject the null hypothesis).

Year	2006-2008	2008-2010
k=2	$V = 94067$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 89291$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$
k=3	$V = 132480$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 134110$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$
k=4	$V = 151380$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 148780$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$
k=5	$V = 155690$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 138660$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$

Table 4.4: Wilcoxon test on the Density, $dataset_1$

This decreasing trend is justified by the fact that generally when authors publish for the first times (two first years), they link to the set of authors with whom they share these first publications which explains at the beginning the emergence of complete networks at $k = 1$. Then, as years pass, if an author has new collaborations, they will not include necessarily all the previous collaborators of the author, and so the density decreases, and this happens regardless of the k. The behaviour is the same with the one reported in [6] for the whole collaboration network extracted from a collection of papers in High Energy Physics Theory.

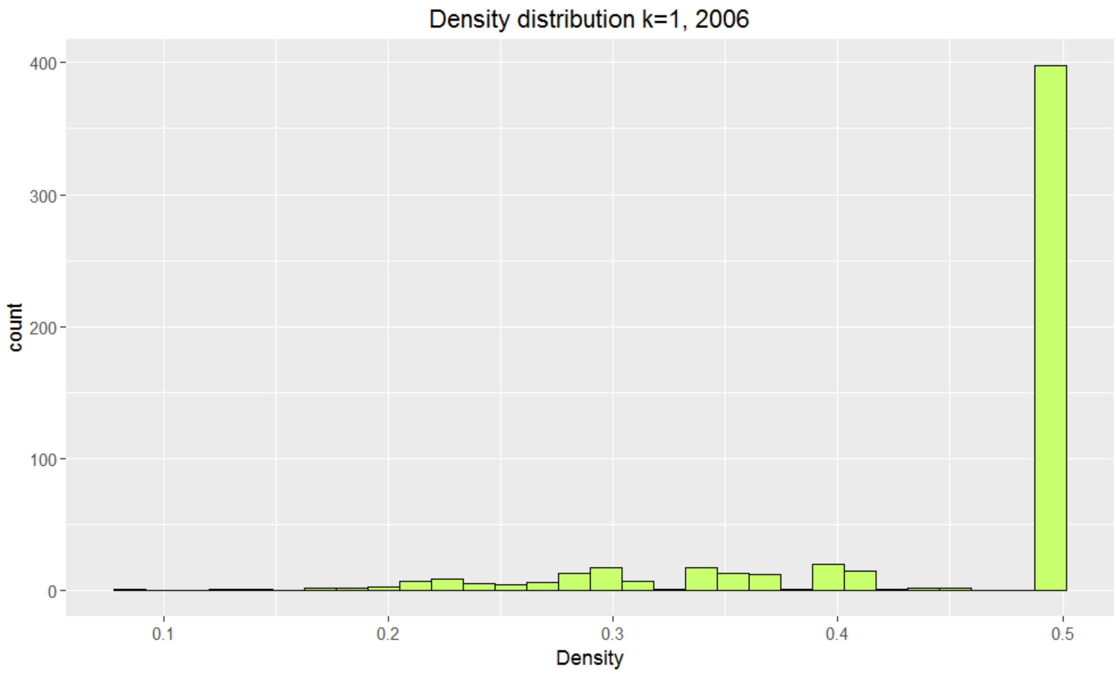


Figure 4.8: Density distribution on 2006, $k = 1$,
for authors starting publishing on 2004 ($dataset_1$).

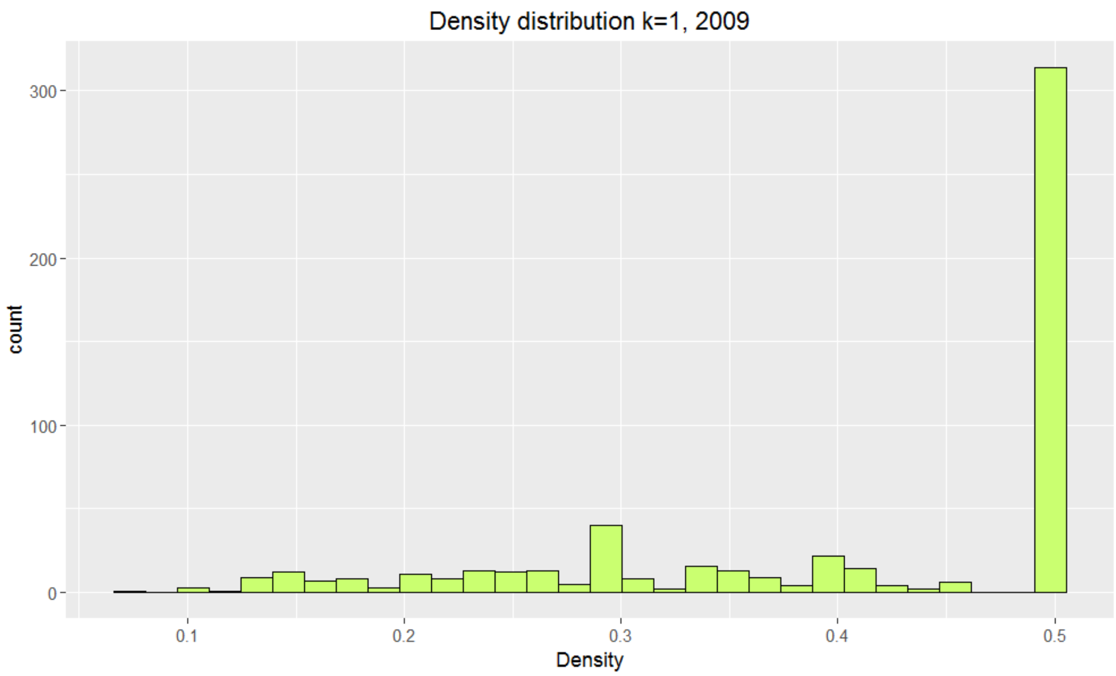


Figure 4.9: Density distribution on 2009, $k = 1$,
for authors starting publishing on 2004 ($dataset_1$).

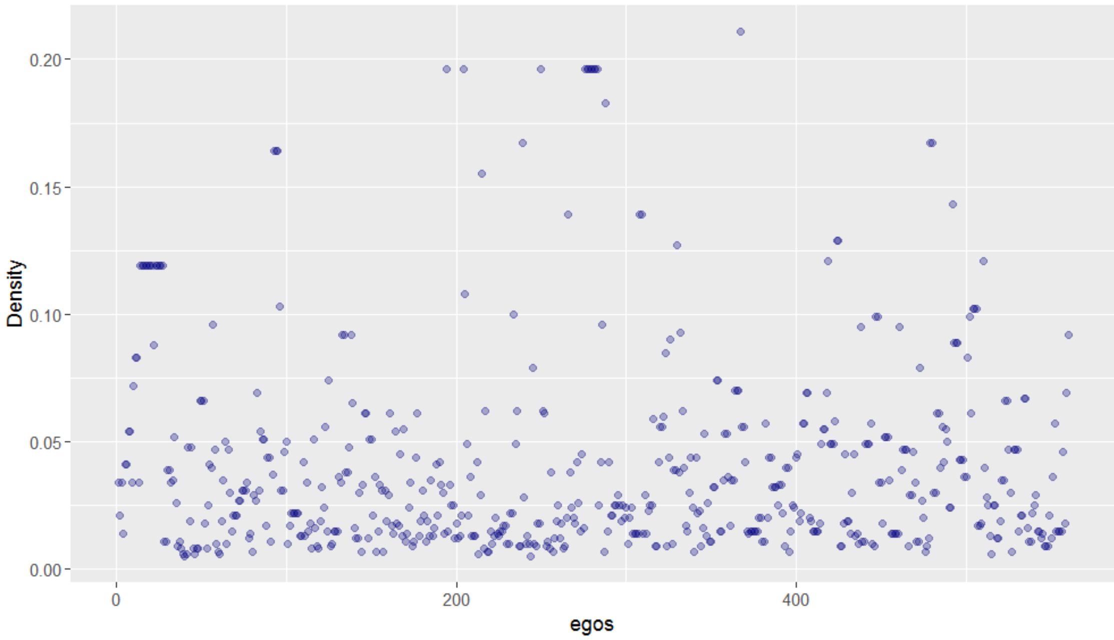


Figure 4.10: Density values on 2006, $k = 4$,
for authors starting publishing on 2004 ($dataset_1$).

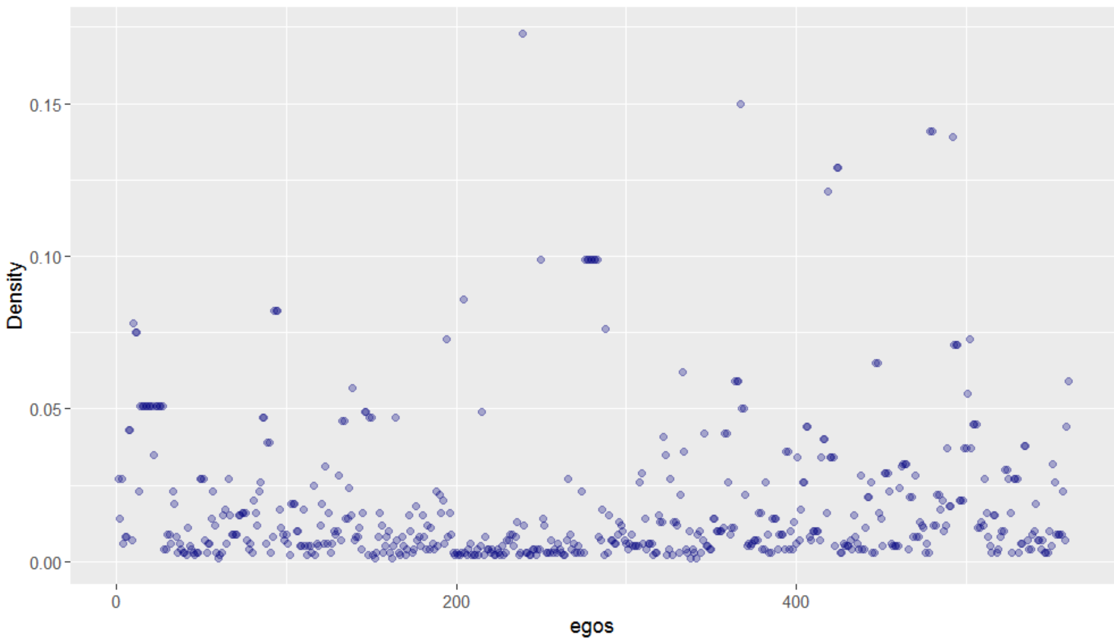


Figure 4.11: Density values on 2009, $k = 4$,
for authors starting publishing on 2004 ($dataset_1$).

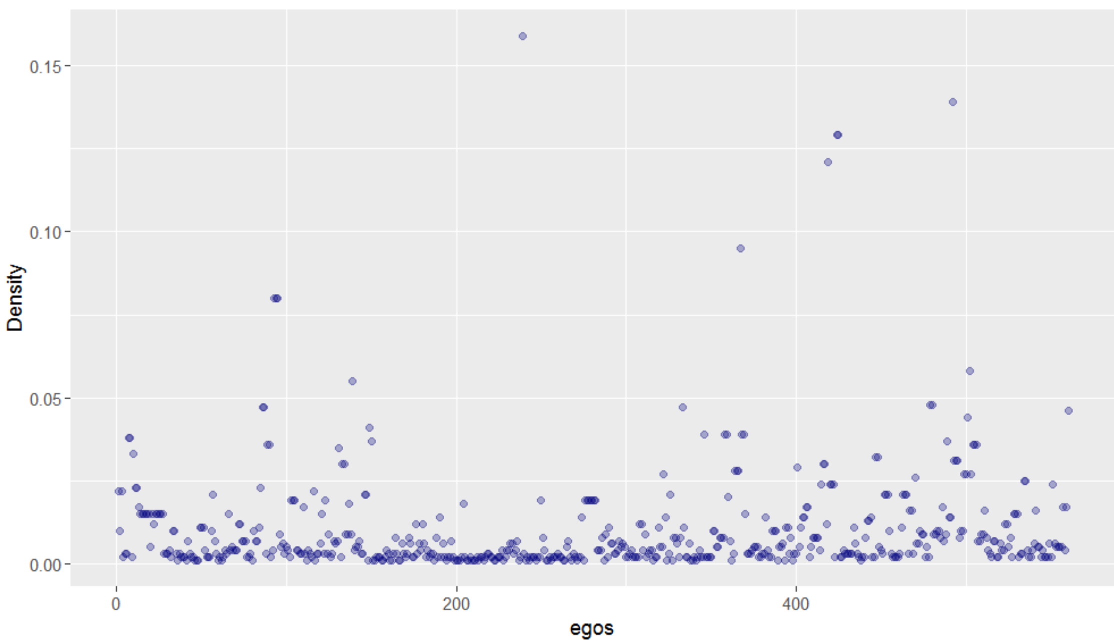


Figure 4.12: Density values on 2012, $k = 4$,
for authors starting publishing on 2004 ($dataset_1$).

Global clustering coefficient and average clustering coefficient.

As presented in Section 4.4.1, we want to measure the transitivity inside co-authorship personal networks via two metrics: (1) the global clustering coefficient (transitivity index) and (2) the average Watts-Strogatz clustering coefficient. While the first one captures the global transitivity of the personal network and the second one expresses the transitivity around the nodes, we want to assess if both metrics behave the same or not.

Thus, in our work, we aim to test if for co-authoring personal networks the two metrics behave in the same way.

Figure 4.13: Global clustering coefficient for $k = 1$, $year = 2006$, for authors starting publishing on 2004 ($dataset_1$).

We will discuss the observed trend for each metric separately regarding the value of k and over the years.

- **Global clustering coefficient evolution:**
  - **Observations regarding k:** As observed for the density, at $k = 1$, a high proportion of author's personal networks have a global clustering coefficient that is equal to 1 because personal networks at this level are usually complete. This proportion, as again observed for the density, is not present anymore for $k = 2$. We also distinguish at $k = 1$, a set of

personal networks with a global clustering coefficient of 0 as presented in Figure 4.13. This is due to the absence of triangles because we are in the case of a personal network with only two nodes and one edge connecting them. But, as the personal networks grow over time, triangle are formed and so, the global clustering of such personal networks will get a value different than 0.

– **Observations over years:** Concerning the trend over the years, the global clustering coefficient decreases which means that there is less transitivity caused by the fact that there are fewer connections among the alters regardless the value of k. These observations are consistent with the earlier discussion about the density. In Figures 4.14,4.15, and 4.16, we give an example of the evolution of the global clustering coefficient for authors starting publishing on 2004 for $k = 5$ over years 2006, 2009, and 2012 respectively. From these graphics, we can see the decrease of the global clustering coefficient over the years.

  The Wilcoxon test confirms the decreasing global clustering coefficient tendency over years. The test results are given in Table 4.5 for the global clustering coefficient value of personal networks between 2006 and 2008, and between 2008 and 2010 for $k = 2$ to 5. The values $p - value_{decreasing}$ and $p - value_{increasing}$ representing the p-value of the Wilcoxon test whether the alternative hypothesis is "greater" or "less" respectively are given. We can see that $p - value_{increasing}$ equals always 1 which indicates that the increasing trend cannot explain the evolution of the global clustering coefficient, while $p - value_{decreasing}$ that is $<$ 2.2e-16 consolidate the observed decreasing trend as it is significantly lower than 5% (the threshold we use to reject the null hypothesis).

Year	2006-2008	2008-2010
k=2	$V = 71266$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 93353$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$
k=3	$V = 84214$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 134110$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$
k=4	$V = 95899$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 98258$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$
k=5	$V = 102450$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 115400$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$

Table 4.5: Wilcoxon test on the Global Clustering Coefficient, $dataset_1$

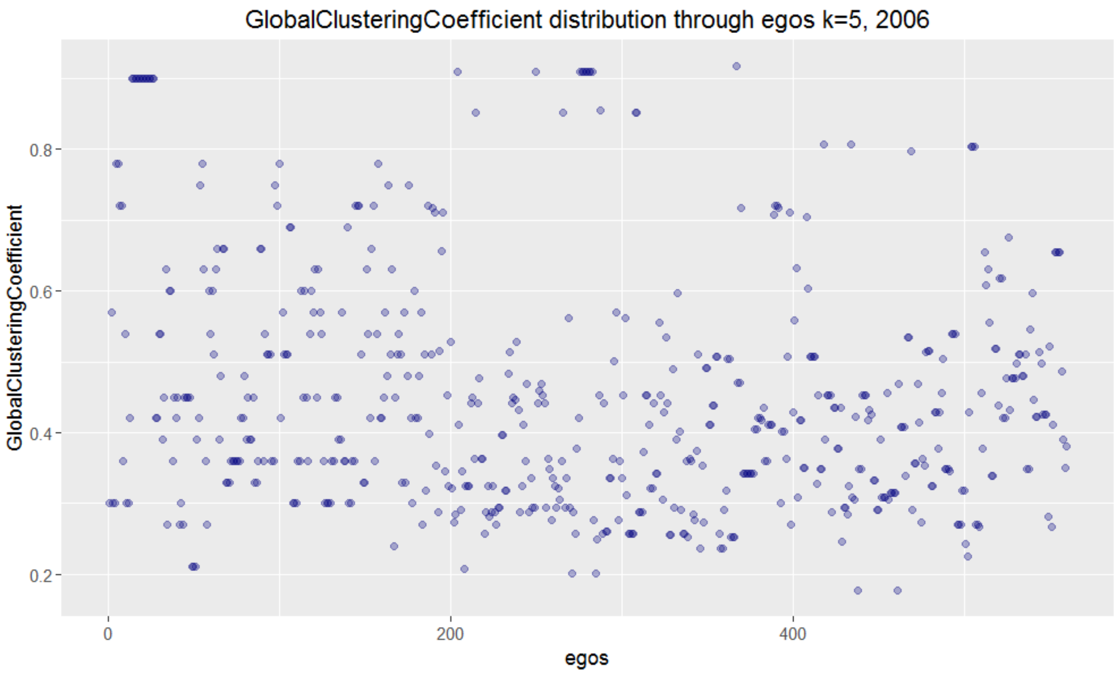


Figure 4.14: Global clustering coefficient values on 2006, $k = 5$ ($dataset_1$).

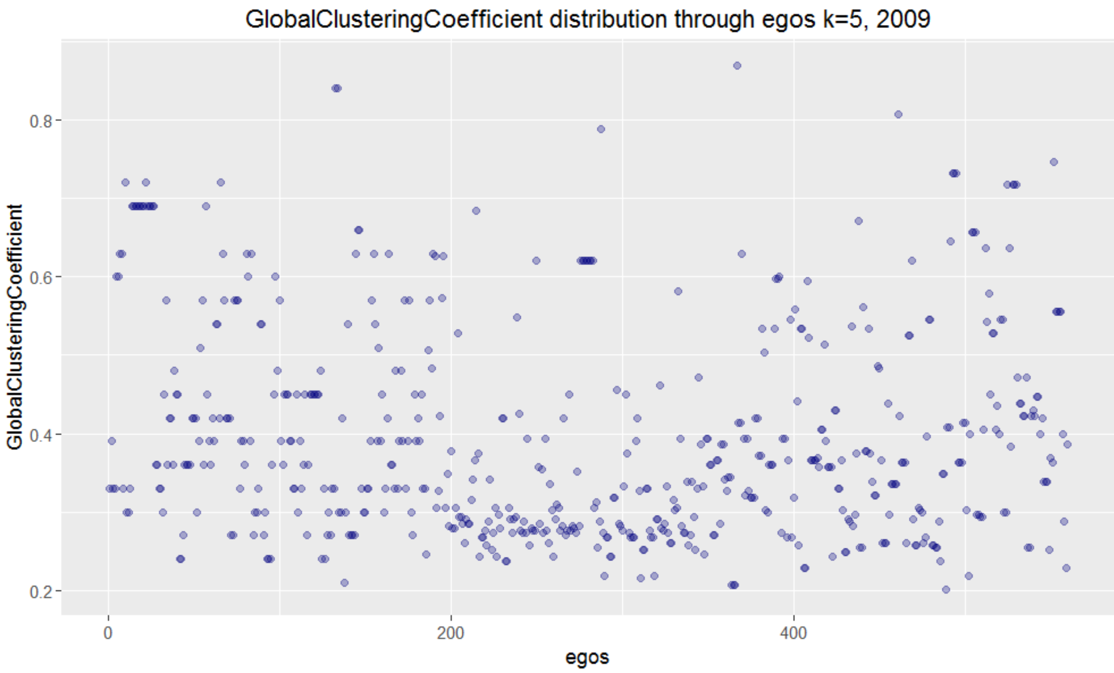


Figure 4.15: Global clustering coefficient values on 2009, $k = 5$ ($dataset_1$).

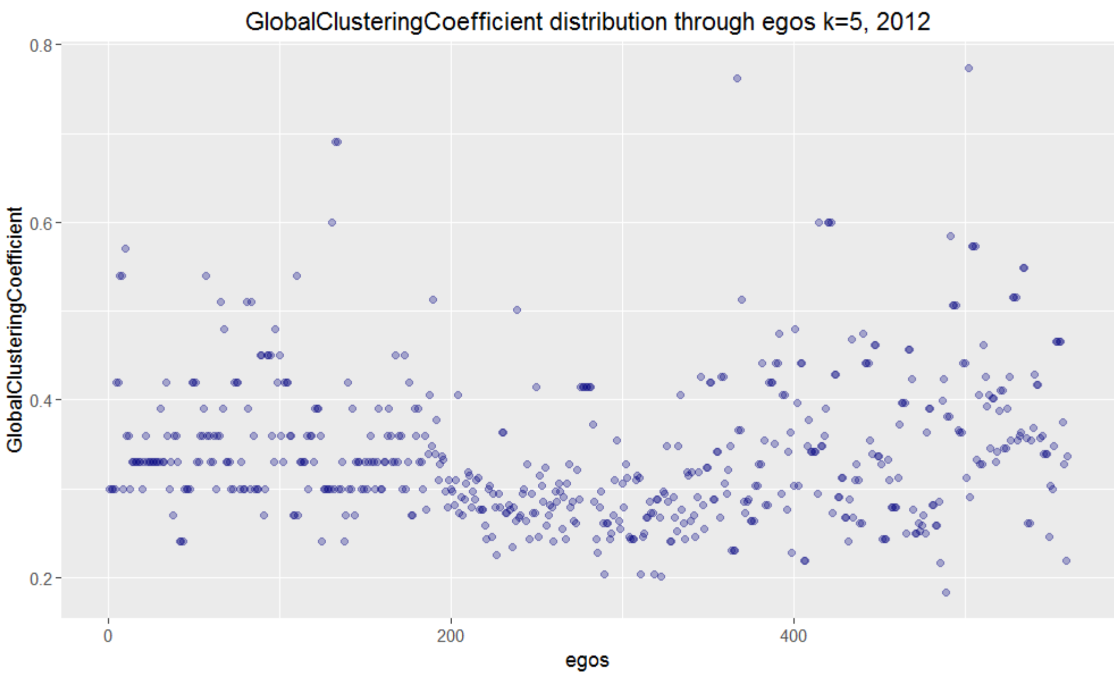


Figure 4.16: Global clustering coefficient values on 2012, $k = 5$ ($dataset_1$).

- **Average clustering coefficient evolution:**
  - **Observations regarding k:** For $k = 1$, and in the early years, a significant set of personal networks hold the maximum average clustering coefficient value (equal to 1), because these personal networks are complete graphs. Similarly, we observed some personal networks with an average clustering coefficient equal to 0 due to the absence of triads around the nodes as we can observe it in Figure 4.17.
  - **Observations over years:** We notice that in general $\forall k = 1..5$, the value of the average clustering coefficient is high with a tendency of increasing through the years. Figures 4.20, 4.21, and 4.22 illustrate this behaviour for authors in $dataset_1$, for $k = 5$ over years 2006, 2009, and 2012. The same tendency was observed in previous works for the whole collaboration networks in the Mathematics and Neural Sciences fields in [15].

    The Wilcoxon test confirms the increasing average clustering coefficient tendency over years. The test results are given in Table 4.6 for personal networks between years 2006 and 2008, and between 2008 and 2010 for $k =$ 2 to 5. We can see that $p - value_{decreasing}$ equals always 1 which indicates that the decreasing trend cannot explain the evolution of the average clustering coefficient, while $p - value_{increasing}$ is < 0.05. However, we notice a different behaviour for $k = 5$ from 2008 to 2010 where $p - value_{increasing}$ approaches 1, which indicates that the average clustering coefficient is not increasing while $p - value_{decreasing} < 0.05$. Thus, we conclude that for this case in particular, the average clustering coefficient is decreasing.

Thus, when the personal networks are evolving over time, the global clustering coefficient and the average one exhibit opposite behaviors, since the first one decreases and the second one increases. We can notice that, while the average clustering coefficient is high and gets higher with years, the global one gets lower and lower.

This observation is not compatible with what is usually claimed in the literature since in real networks both clustering coefficients tend to some positive constant as the networks grow as reported in [108], or in [84], where the authors claimed that both metrics represent the clustered architecture of the network, with a small difference on value scaling.

However, we found that a recent work done by Estrada in [48], reported that for certain classes of graphs, global clustering coefficient and the average one diverge.

Year	2006-2008	2008-2010
k=2	$V = 37430$, $p - value_{decreasing} = 1$, $p - value_{increasing} = 1.652e - 05$	$V = 38819$, $p - value_{decreasing} = 0.9945$, $p - value_{increasing} = 0.005497$
k=3	$V = 48898$, $p - value_{decreasing} = 1$, $p - value_{increasing} = 9.93e - 09$	$V = 46040$, $p - value_{decreasing} = 1$, $p - value_{increasing} = 8.583e - 11$
k=4	$V = 42409$, $p - value_{decreasing} = 1$, $p - value_{increasing} < 2.2e - 16$	$V = 61459$, $p - value_{decreasing} = 0.9999$, $p - value_{increasing} = 0.0001041$
k=5	$V = 36651$, $p - value_{decreasing} = 1$, $p - value_{increasing} < 2.2e - 16$	$V = 84175$, $p - value_{decreasing} = 0.03741$, $p - value_{increasing} = 0.9626$

Table 4.6: Wilcoxon test on Average Clustering Coefficient, $dataset_1$

Figure 4.17: Average clustering coefficient for $k = 1$, $year = 2006$, for authors starting publishing on 2004 ($dataset_1$).

More precisely, Estrada was referring to the *windmill graphs* as an example of graphs in which this phenomenon occurs, which is also verified for our personal networks. Hereafter, we give the definition of a windmill graph.

Definition 4.4.3 *A windmill graph $W(k, n)$ is an undirected graph characterized by a central node that is surrounded by n cliques of size k suhc that $k \geq 2$ and $n \geq 2$*

[60]. These cliques are composed of nodes that are completely connected with each other and with the central node but that do not have any connections with the other cliques (or have few connections in real world networks where cliques are overlapping). The central node is shared by all the cliques of the windmill graph.

An example of a windmill graph $W(3, 4)$, is presented in Figure 4.18 where $k = 3$, and $n = 4$.

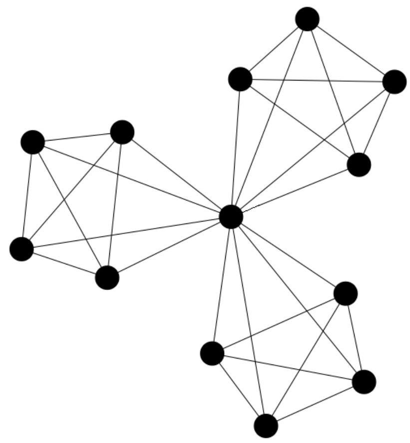


Figure 4.18: Windmill graph example, W = (3, 4).

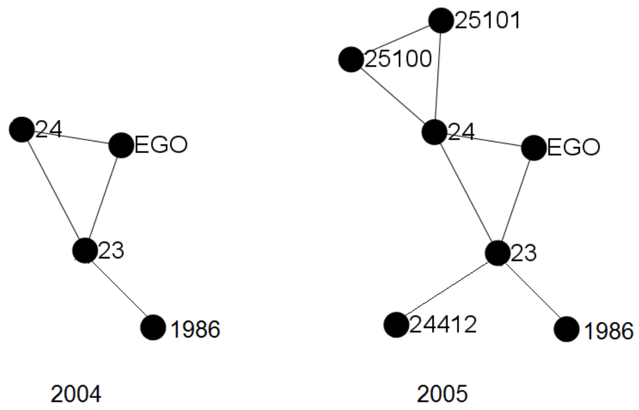
We believe that such structure can arise in personal networks especially when coming from scientific collaboration networks as stated in [48], where the author proved the presence of windmill graph structures inside both collaboration and citation networks, and showed the divergence of the two clustering metrics.

For our case, the underlined observations are explained by the fact that in the personal networks of co-authoring, a publication will involve the creation of a clique between the authors of that publication where each pair of nodes is connected which explains the high average local clustering coefficient around nodes. But, the fact that a given author can over time have publications with new collaborators that will not in the most of cases, concern his/her old co-authors, will imply:

1. From the one hand, the decrease of the global clustering inside the personal network, and

2. From the other hand, the increase of the average clustering coefficient.

An example of such situation is given in Figure 4.19. This could be explained by the fact that scientists during their career are led to change institutions and work places, collaborate with new authors (e.g. PhD students) and even change their research focus. Thus, studying how cliques evolve over time for this type of data is interesting.



On the left, we have the personal network of ego=22 on 2004, and on the right the personal network of ego=22 on 2005. We can see that two new cliques appear: the first clique have a size of 3 and is formed of node "24" with new nodes "25100" and "25101". The second new clique has a size of 2 and is formed between node "23" and the new node "24412".

Figure 4.19: Cliques formation in the personal network of "EGO" between 2004 and 2005, $k = 2$.

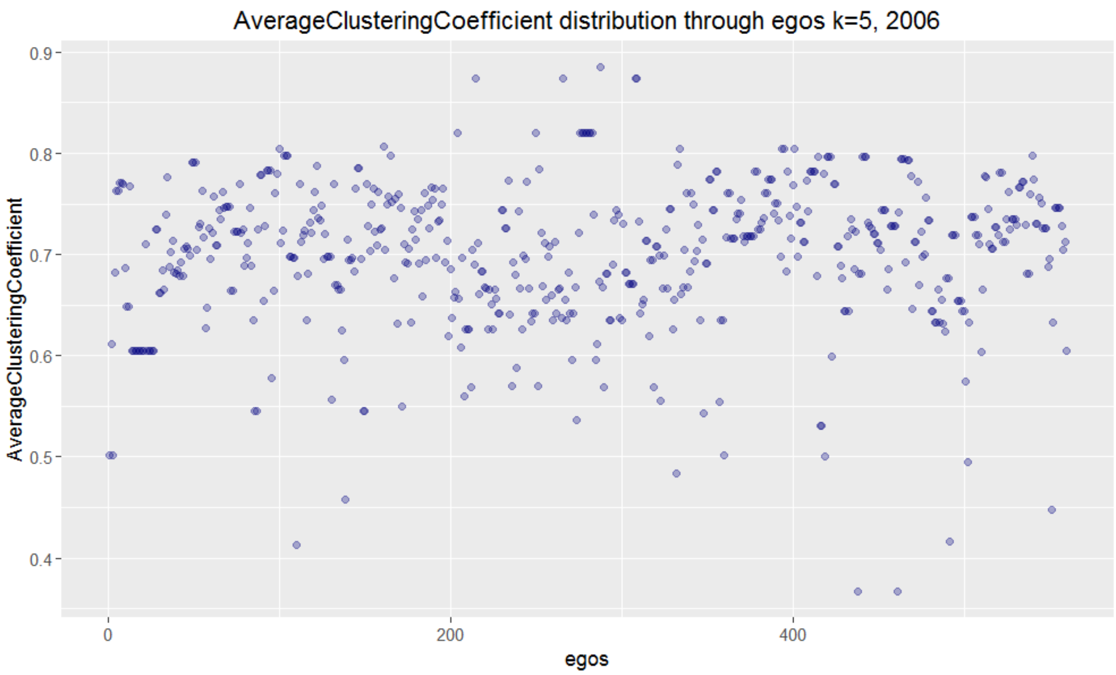


Figure 4.20: Average clustering coefficient values on 2006, $k = 5$ ($dataset_1$).

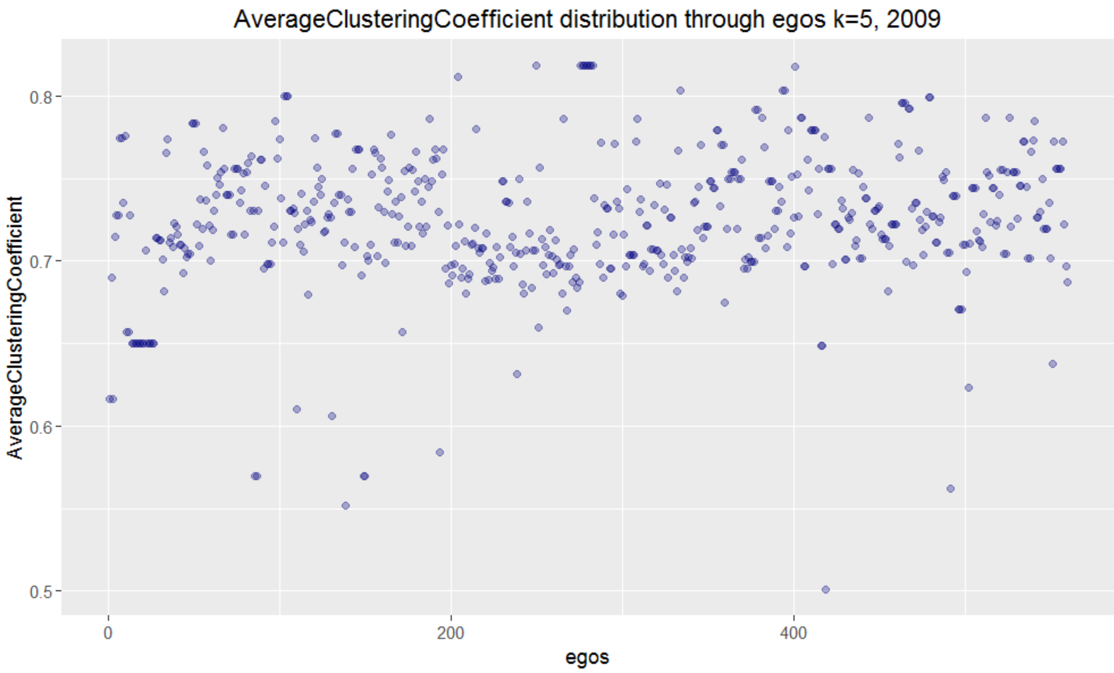


Figure 4.21: Average clustering coefficient values on 2009, $k = 5$ ($dataset_1$).

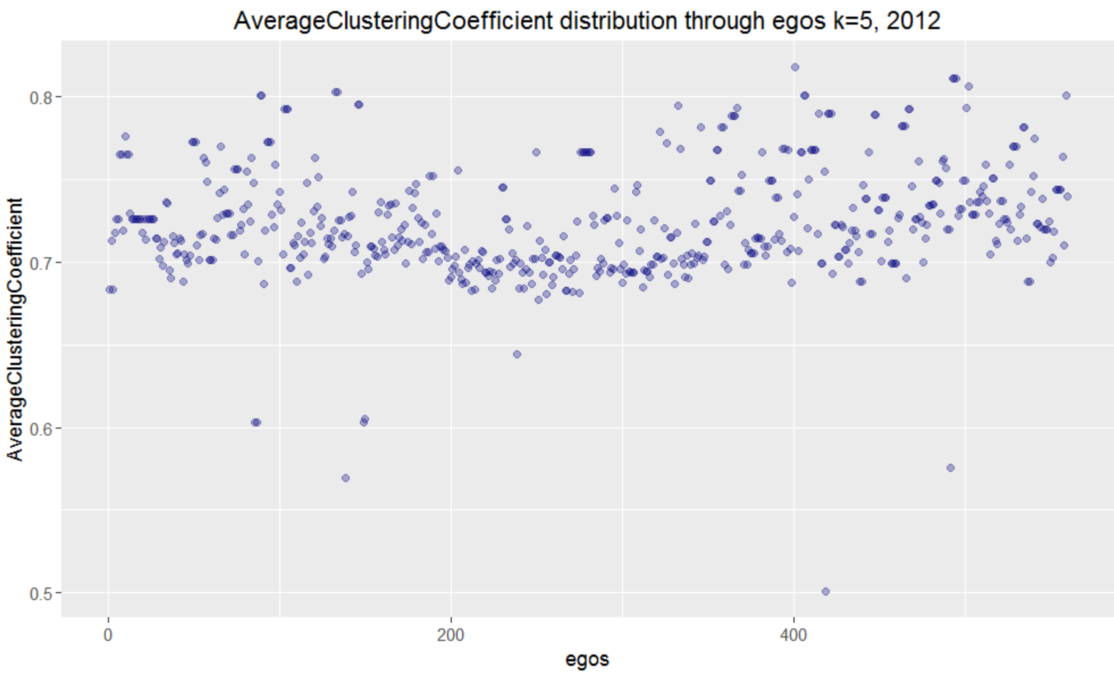


Figure 4.22: Average clustering coefficient values on 2012, $k = 5$ ($dataset_1$).

Power law distribution. In order to check the existence of a power law distribution over the degrees of nodes part of co-authorship personal networks, we have used the approach proposed by Clauset et al. in [34] who presented a statistical framework for discerning and quantifying power-law behavior in empirical data.

This approach combines maximum-likelihood fitting methods with goodness-of-fit tests based on the *Kolmogorov–Smirnov (KS)* statistic and likelihood ratios. More precisely, the approach works as follow: the algorithm proposed in [34] will generate a large number of synthetic data like the input data for which we want to test the existence of a power law distribution and fit a model to it. Then a *Kolmogorov-Smirnov (KS)* goodness-of-fit test is performed between the generated data and the fitted model to estimate the $xMin$ parameter (the data point for which the resulting distribution has the smallest KS statistic to the data). The exponent α of the power law is thus estimated using a maximum likelihood estimator for that $xMin$ value. Then, in order to conclude on whether the power law model is a proper model, we compute the $significance$. A high $significance$ value means that the power law is a good fit. Clauset and co-authors [34] suggest that for a p-value below 0.01 the power law hypothesis should be rejected. If the resulting p-value is greater than 0.1, the power law is a plausible hypothesis for the data, otherwise it is rejected.

Thus, the application of this approach for our case will translate in providing as input a vector of discrete values representing the degrees of nodes of the personal network for which we want to perform the test, and we get as output the answer of the test (*true* if the personal network follows a power law distribution, *false* if not) depending on the $significance$ parameter computed by the algorithm, and the estimated parameter α of the power law distribution function.

It is important to note that, the fact that the degrees follow a power law distribution is translated as in the following in the evolution process:

When new nodes join the personal network, they will tend to connect to nodes having a high degree, which will lead to the appearance of few highly connected nodes in the personal networks and many weakly connected nodes (*preferential attachment*).

- **Observations regarding k:** Figure 4.24 represents, for each k from 2 to 5 and along the period 2006-2013, the proportion of personal networks that follows (in blue) or not (in red) a power law distribution in their degrees. Computing the power law distribution test for personal networks of $k = 1$ was not possible because the networks were too small and the computation did not make any sense since as mentioned in [34], the approach is suitable for a data observation values above 50 (translated in our case into personal network with more than 50 nodes). While, the personal networks at $k = 1$ have a number of nodes below 50, and this overall the years. Thus, we concentrate on the evolution of the power law test from $k = 2$ to $k = 5$.

We notice that the power law distribution of personal networks can be considered as verified in the cases of $k = 2, 3, 4$ even if the false proportion gets increasing when k grows and more significantly starting with $k = 4$. At $k = 5$, the tendency is completely inverted from 2011 since the proportion of personal networks having the power law test to false becomes larger. 84.73 80,20

- **Observations regarding the years:** The power law distribution for $k = 2, 3$ is followed and this for all years as we can observe in Figure 4.24 with a proportion of 84.73% on average for $k = 2$, and of 80.2% on average for $k = 3$. For $k = 4$, it is also followed, because the true proportion is higher with an average of 65,61% over the years. For $k = 5$, the results are not so conclusive and the tendency is inverted starting with 2011 when the proportion of personal networks having the power law test to false overpasses true.

  This later observation is interesting since we expected that, larger personal networks as it is the case when $k = 5$, to confirm the properties that were observed for global networks as the power law distribution of networks' nodes degrees as proved in [15] for the case of scientific collaboration networks.

  We also observed that the proportion of personal networks that returned false was increasing with k (for $k = 4$ and $k = 5$) and over the years. We were unable to verify the main reason why this happens and this behavior is the opposite than the expected one. One possible explanation could be that the evolution of personal networks connects previously existing cliques (see the clique between the 15 nodes labeled from 1589 to 1603 in the example Figure 4.23). In a clique, each node is highly connected locally so the merging of two cliques into one creates a new, bigger and more connected clique; thus we have more nodes with higher degrees and subsequently the power law distribution tests fail for all these networks. This behavior is compatible with what happens in many cases in co-authorship networks, where we frequently see collaborations among groups and a new publication has as authors all (or almost all) the members of both groups. This explanation will be verified later in this chapter with cliques analysis.

If we consider the co-authors network as a whole, the degrees' distribution follows a power law (as observed in [15]), this comes from the fact that at a large scale the proportion of such highly connected authors becomes negligible compared to the majority of authors that have a small amount of collaborations.

As a conclusion, the power law distribution is describing the distribution of the degrees of our personal networks for $k = 2, 3$ but less for $k = 4, 5$. This fact is important since it does not go with the majority of real networks studied in the literature, including personal networks, where the power law distribution was valid and governed the proposed evolution models that are all based on the *preferential attachment* principle. Thus, such model would not be suitable for the evolution of co-authorship personal networks we are studying.

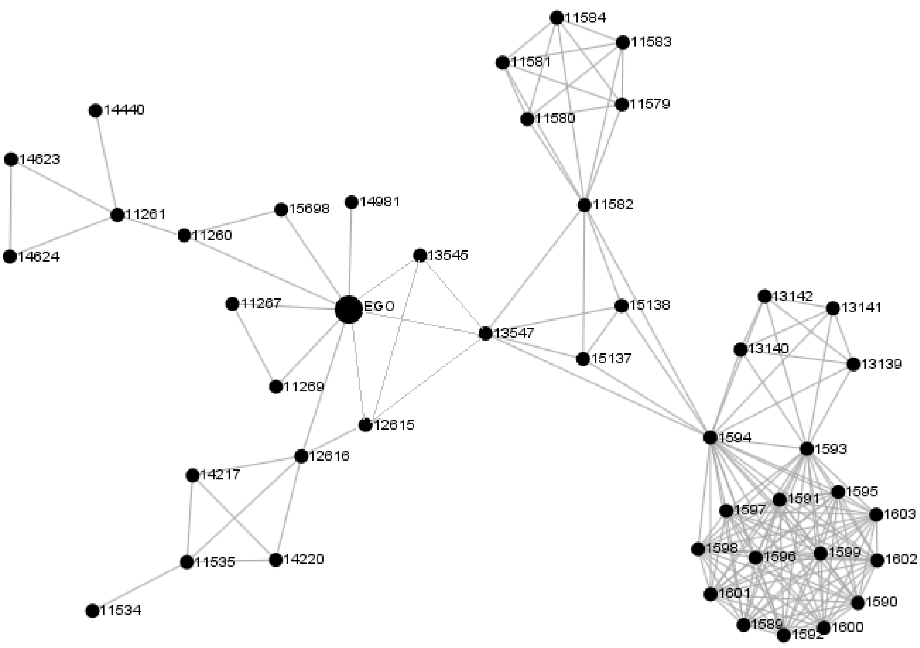


Figure 4.23: Example of a clique formation.

Checking Poisson distribution

Given the obtained results we discussed in the previous section, it is obvious that the power law distribution does not characterize the studied personal networks in particular for $k = 4, 5$. Thus, we decided to check if the Poisson distribution explains our degree distribution as it is the case for random graphs of Erdős and Rényi [47]. To do so, we computed on $dataset_1$ over years from 2006 to 2013 and over k from 1 to 5, the fraction of personal networks following a Poisson distribution at 1%, 5%, and 10% loss. The obtained results are represented in the plots Figure 4.25. The

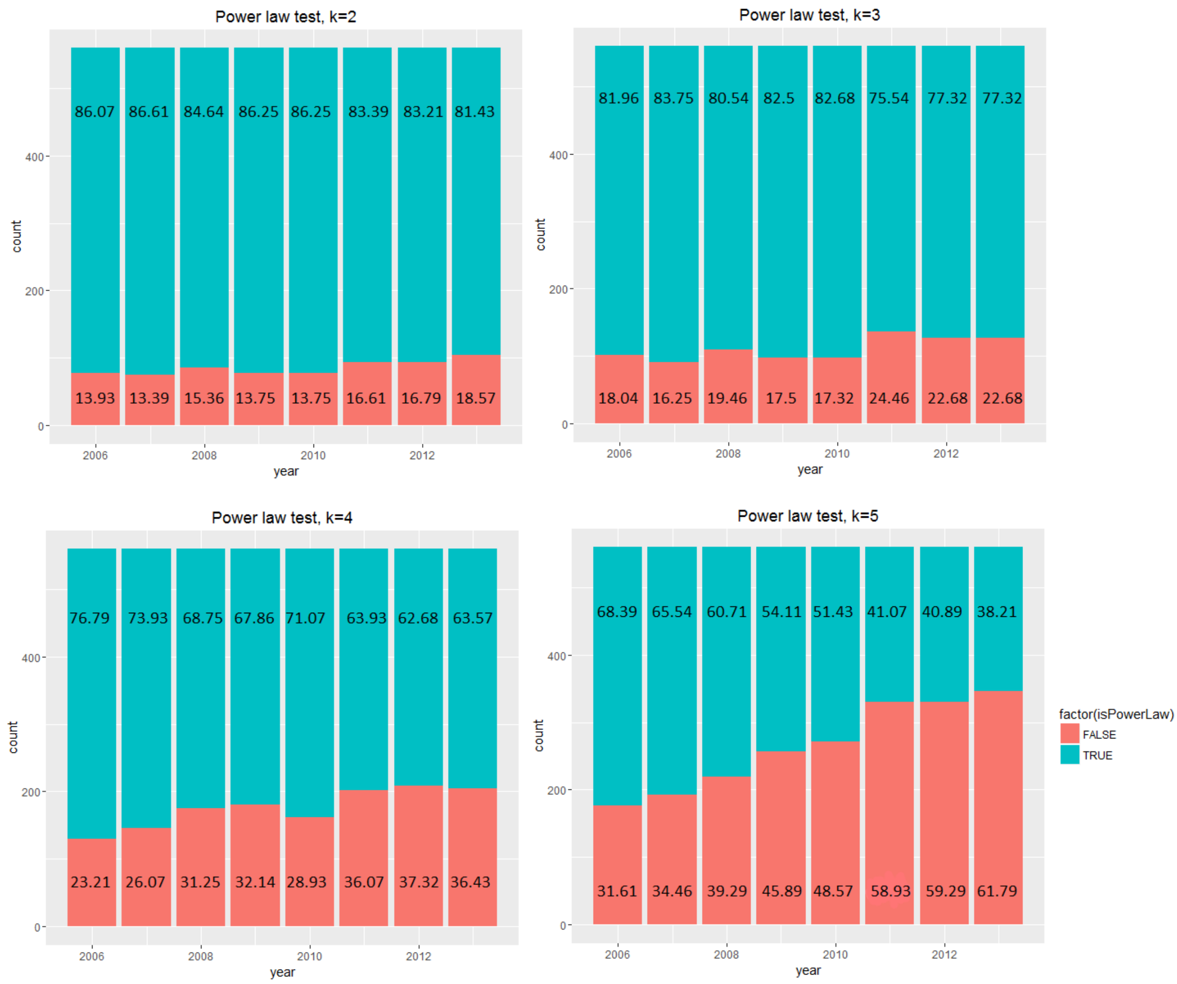
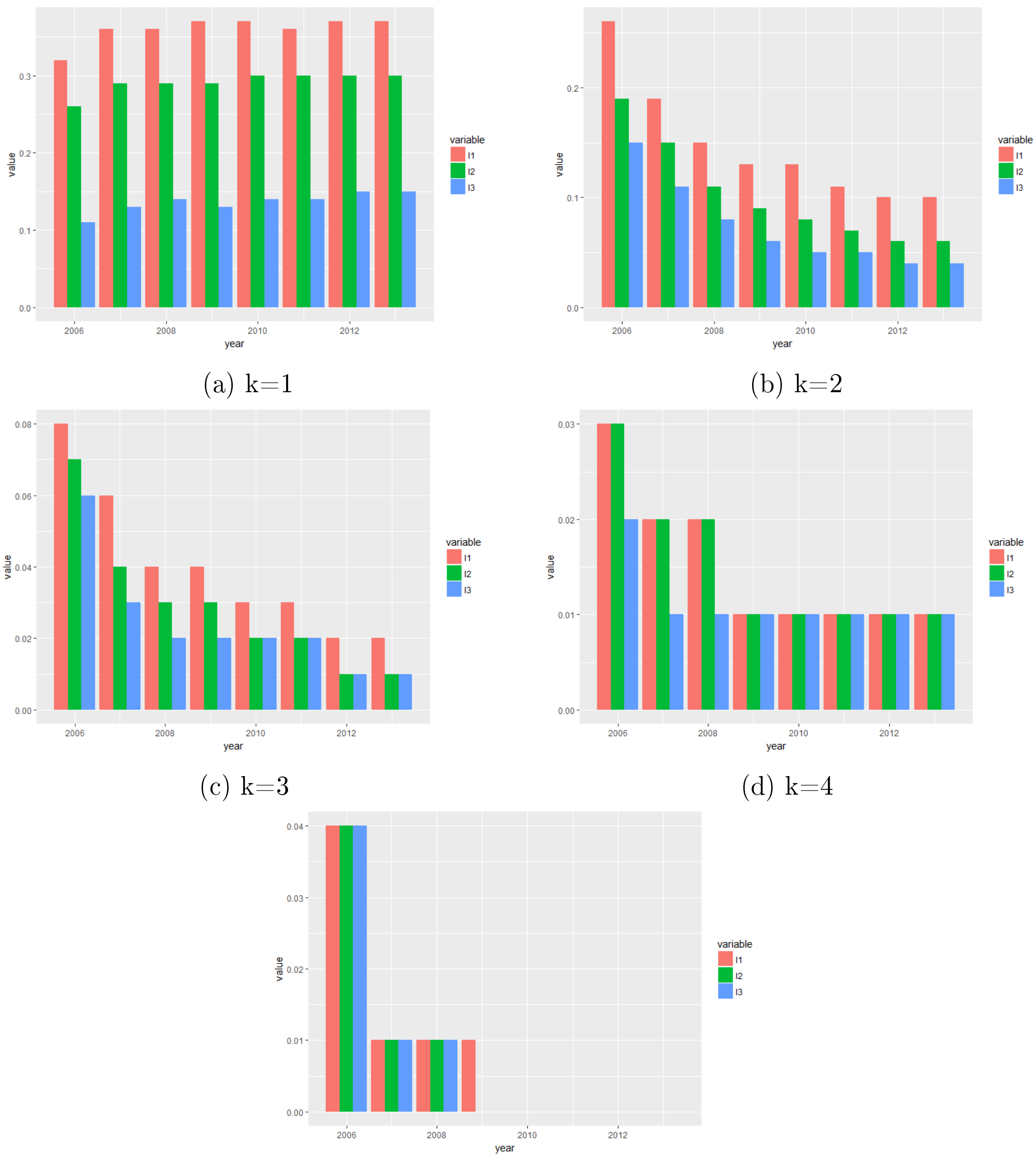


Figure 4.24: Power law distribution test for $k = 1$ to 5, from 2006 to 2013 ($dataset_1$).

variables $l1$, $l2$, and $l3$ in the plots corresponds to 1%, 5%, and 10% loss respectively. For personal networks with $k = 1$, we observe a tendency of increasing over the years of the percentage of personal networks following a Poisson distribution reaching the highest values (more than 35%). We will further see that these personal networks are a particular case comparing to the other k values. Indeed, the percentages are then decreasing over the years and over k with a very low amount of personal networks fitting to the Poisson distribution at k=5, becoming equal to 0 after 2009. The outlined behaviours are similar among the three loss values. Notice that the percentages are always higher at 1% loss if not equal to the ones at 5 and 10% loss. Followed by the percentages at 5% loss which is higher if not equal to the percentages at 10% loss.

Poisson vs Power law distribution. In this part, we are trying to discover if the same networks are following both power law and Poisson distributions or if there are



(a) k=1 (b) k=2

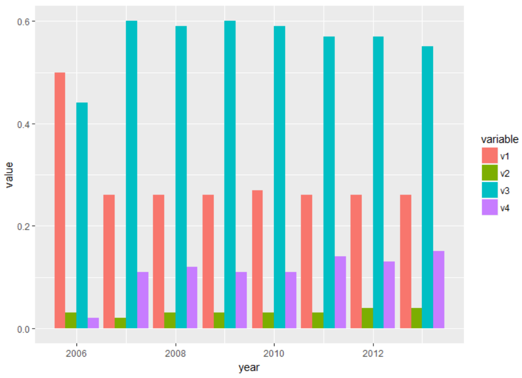
(c) k=3 (d) k=4

(e) k=5

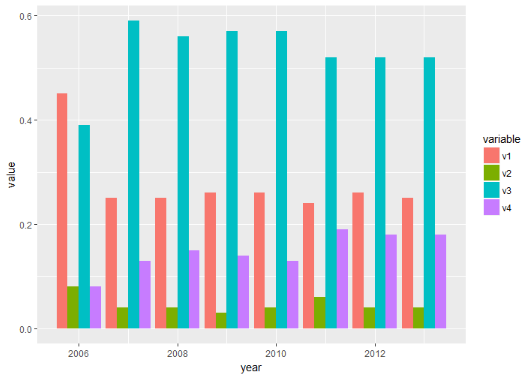
Figure 4.25: Poisson degree distribution test for $k = 1 \ldots 5$, and years 2006 to 2013, for $dataset_1$.

particular personal networks that are governed by one of the distributions, or more, if in some cases none of these distributions is valid. Thus, we computed the following four values: $v1$, $v2$, $v3$, and $v4$ representing respectively the proportions of personal networks following Poisson and Power law distributions, Poisson distribution only, Power law distribution only, and nor Poisson nor Power law distributions. The corresponding plots are given in Figure 4.26. The first prominent observation whatever the value of k is that the proportion of co-authorship personal networks following only a power law distribution in their degrees (given by $v3$) constitute the largest tendency except in two situations: (1) at the starting year of observation, 2006, where it is exceeded by the proportion of networks following both distributions in their degrees, and (2) when k equals 5, we see that, on 2010, the proportion given by $v3$ becomes equal to the one given by $v4$ which represents the set of networks not following nor a power law nor a Poisson distribution in their degrees. This last proportion is then leading starting from 2011. We notice that $v2$ proportion (networks following Poisson distribution only) is less represented in our data.

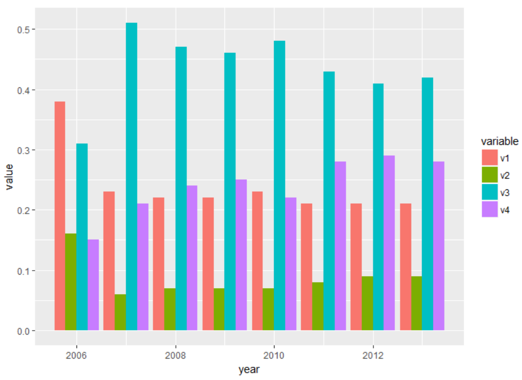
Thus, following the observations above, it is clear that the Poisson distribution is not explaining the cases where the power law test failed in explaining the data. The power law remains a better choice so far to describe the degree distribution except for $k = 1$, where Poisson is a plausible distribution. A particular case however stands out concerning co-authorship personal networks at k=5 in the last years starting with 2009, where the degree distribution seems to follow a particular distribution different from the power law and Poisson since $v4$, meaning that nor power law, nor Poisson distributions explain the data, is the largest proportion among the other combinations. Unfortunately, we could not have from our analysis any indication about the statistical distribution that could explain the degree distribution of the personal networks at this stage.



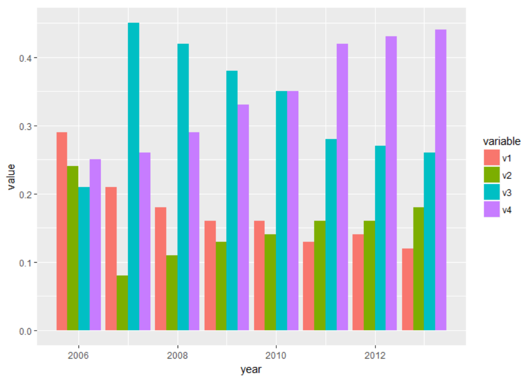
(a) k=2



(b) k=3



(c) k=4



(d) k=5

Figure 4.26: Poisson Distribution vs Power Law Degree Distribution, for $k = 1...5$, and years 2006 to 2013, for $dataset_1$.

Ego's maximum degree of separation (*k-max*).

As presented in Section 4.4.1, we would like to know the maximum shortest path from the ego observed in the whole network. We performed the computation of $k-max$ on our $dataset_1$ for each ego and for each year from 2006 to 2013. The plot in Figure 4.27 gives the distribution of $k-max$ value over all the years. We distinguish two phases: (1) the first around $k-max = 4\ to\ 7$ with very few personal networks having such values, and (2) the second one ranges from $k-max = 13\ to\ 25$. We notice that no personal network has a $k-max$ value between 8 and 12, whatever the year. The evolution of $k-max$ over the years reveals that after reaching a certain

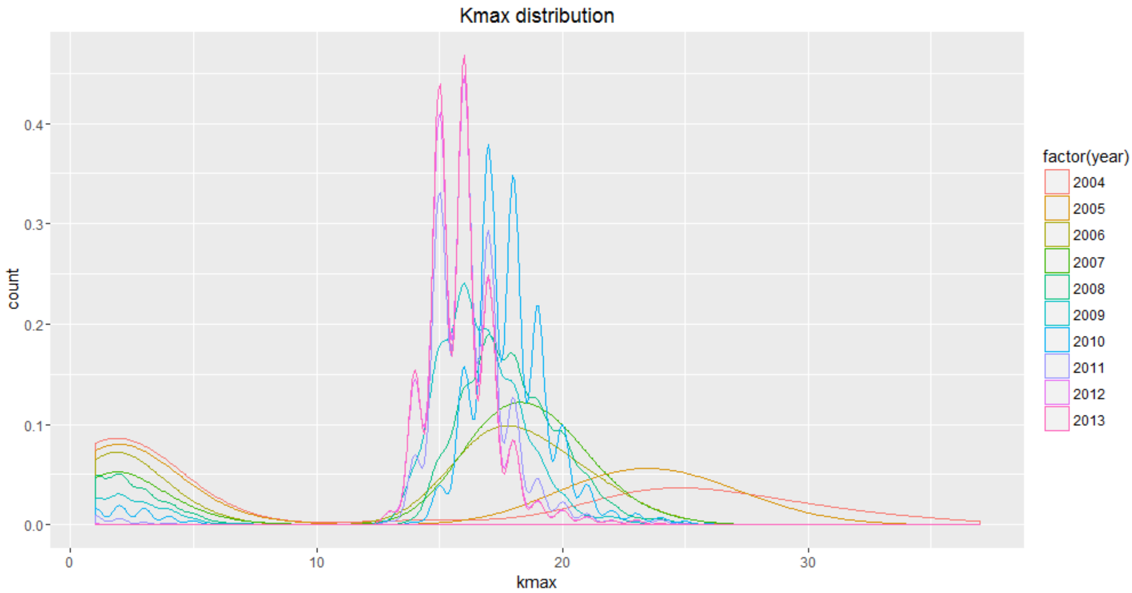


Figure 4.27: $k-max$'s distribution over the years for OPNs in $dataset_1$.

size, the number of nodes composing the personal network will remain stable while $k-max$ can vary (the maximum $k-max$ achieved equals 25). This size corresponds to the whole network giant component size for each year (Figure 4.28). The fact that no $k-max$ value is between 8 and 12 is due to the interconnection among existing personal networks which makes jump $k-max$ from 7 to a value $\geq$ 13. As we

Year	2006	2007	2008	2009	2010	2011	2012	2013
Average $k-max$	18.18	18.39	18.00	16.61	17.82	16.18	15.82	15.81

Table 4.7: Average $k-max$ per year, for OPNs in $dataset_1$.

report in Table 4.7, the average $k-max$ computed for each year from 2006 to 2013

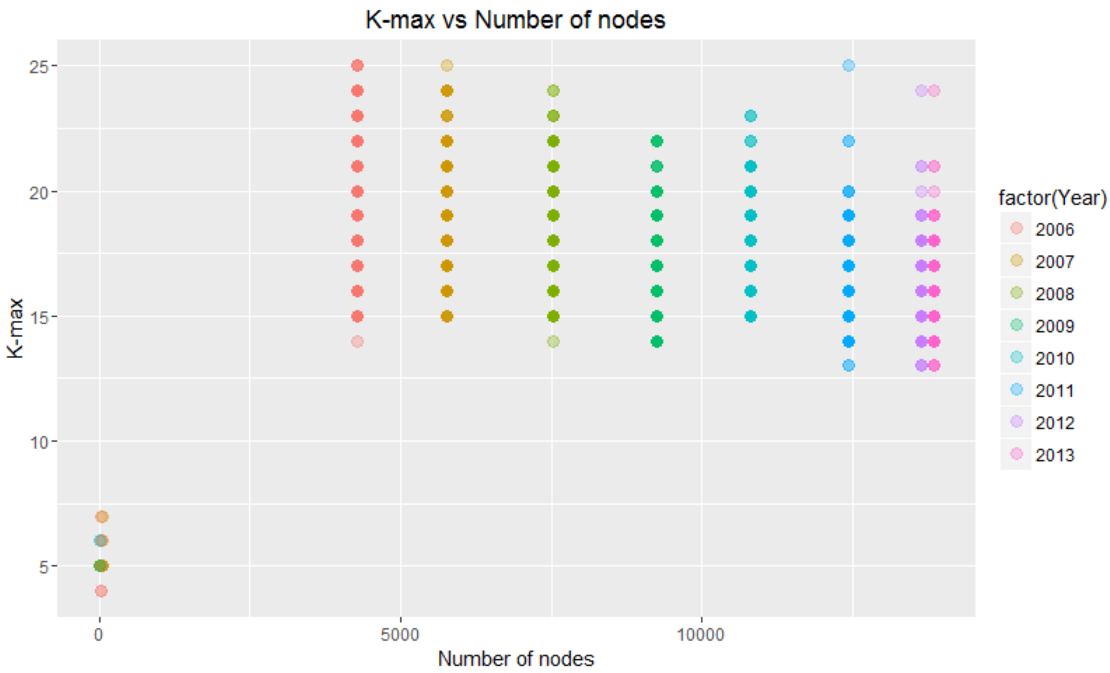


Figure 4.28: $k - max$ vs. the number of nodes over the years, for OPNs in $dataset_1$.

ranges between 15 and 19 approximately, which means that in all cases, the average is higher than 6. We conclude that the *6-degrees of separation* property that characterizes *small world* networks does not hold in our case and thus the collaboration network we are studying does not represent a *small world* network. On the contrary, in [46], the authors found that the 6-degrees of separation phenomenon is valid when they studied the collaboration network of scientists who publish in the Database area.

Ego degree centrality and personal networks average degree. After computing the degree centrality over personal networks's nodes to check the existence of a power law distribution, we now present the observations on the evolution of both the degrees of the ego, and the average degree for all personal network's nodes.

- Degree of the egos: we observed that in general only few egos are very highly connected while most of them have low degrees (Figure 4.29). The distribution remains almost the same even when years pass (Figure 4.30).

- Average degree: for the average degree over all the nodes composing the personal networks, we would like to check if it increases over time when considering personal networks instead of full networks as it was found by Barabasi in [15]

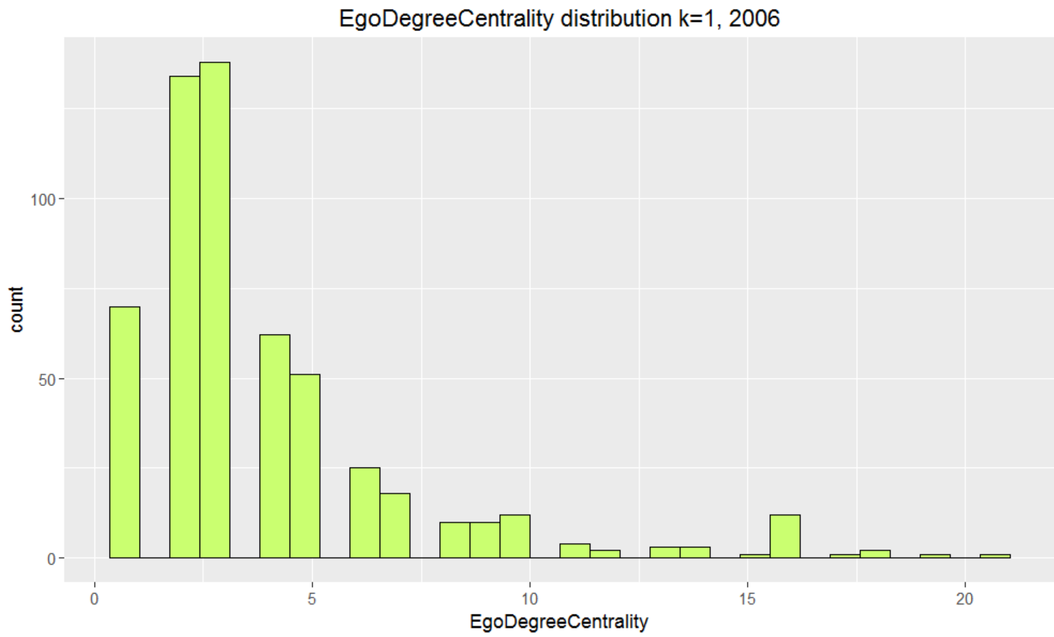


Figure 4.29: Distribution of ego degree centrality on 2006, for $dataset_1$.

when studying the evolution of the whole network of scientific collaboration in both Mathematics and Neural Sciences fields (in both networks the average degree increases over time).

- **Observations regarding k:** We found that for $k = 1$, the average degree remains low (important number of values around 1 and 2) even when the personal network grows as we can see in Figure 4.31. Then, from $k = 2$, we have seen that there are fewer values around 1 and 2 as we can observe it in Figure 4.32 for $k = 3$.

- **Observations regarding the years:** From $k = 2$, the average degree increases to concentrate around a value of 4 - 5 (as shown in Figure 4.32 for $k = 3$).

  The Wilcoxon test confirms that the average degree increases over years. The test results are given in Table 4.8 for the average degree in personal networks between 2006 and 2008, and between 2008 and 2010 for $k = 2$ to 5. The values $p - value_{decreasing}$ and $p - value_{increasing}$ representing the p-value of the Wilcoxon test whether the alternative hypothesis is "greater" or "less" respectively are given. We can see that $p - value_{decreasing}$ equals always 1 which indicates that the decreasing trend cannot explain the evolution of the average degree, while $p - value_{increasing}$ that is $<$ 2.2e-16

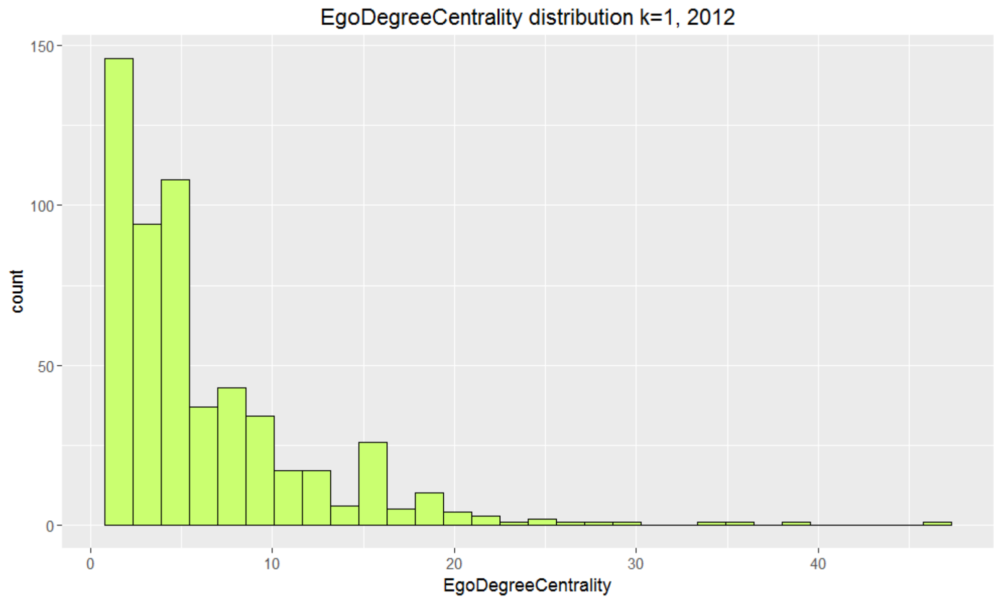


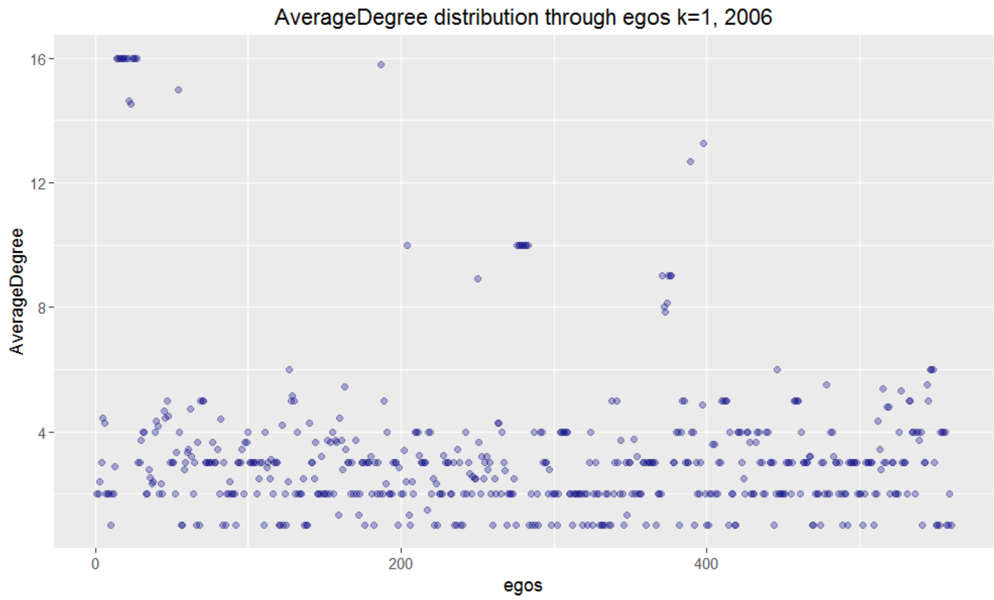
Figure 4.30: Distribution of ego degree centrality on 2012, for $dataset_1$.

consolidates the observed increasing trend as it is significantly lower than 0.05.

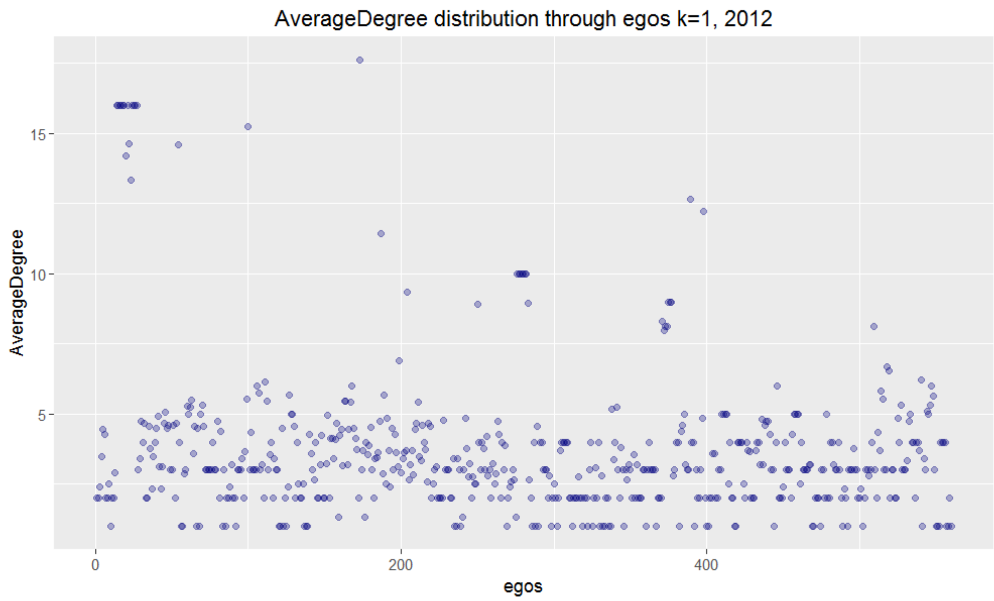
Our results are consistent with what was found by Barabasi in [15]. The increase of the average degree over time is explained by the fact that when a new publication is registered, its authors form necessarily a clique (all coauthors of the same publication are connected between them); then when another publication comes with mainly the same authors and some new ones, these new ones are added to the clique. This results in increasing the average degree among the alters of a personal network with the same k.

Year	2006-2008	2008-2010
k=2	$V = 26058$, $p-value_{decreasing} = 1$, $p-value_{increasing} < 2.2e-16$	$V = 21594$, $p-value_{decreasing} = 1$, $p-value_{increasing} =< 2.2e-16$
k=3	$V = 37386$, $p-value_{decreasing} = 1$, $p-value_{increasing} < 2.2e-16$	$V = 27746$, $p-value_{decreasing} = 1$, $p-value_{increasing} < 2.2e-16$
k=4	$V = 38206$, $p-value_{decreasing} = 1$, $p-value_{increasing} < 2.2e-16$	$V = 24680$, $p-value_{decreasing} = 1$, $p-value_{increasing} < 2.2e-16$
k=5	$V = 36402$, $p-value_{decreasing} = 1$, $p-value_{increasing} < 2.2e-16$	$V = 23385$, $p-value_{decreasing} = 1$, $p-value_{increasing} < 2.2e-16$

Table 4.8: Wilcoxon test on the Average Degree, $dataset_1$

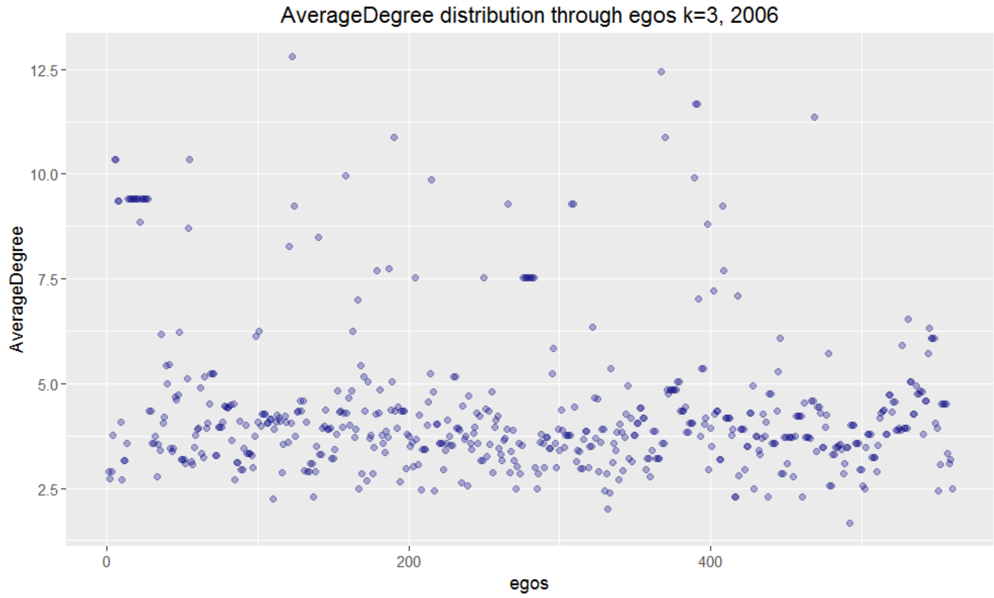


(a) k=1, 2006

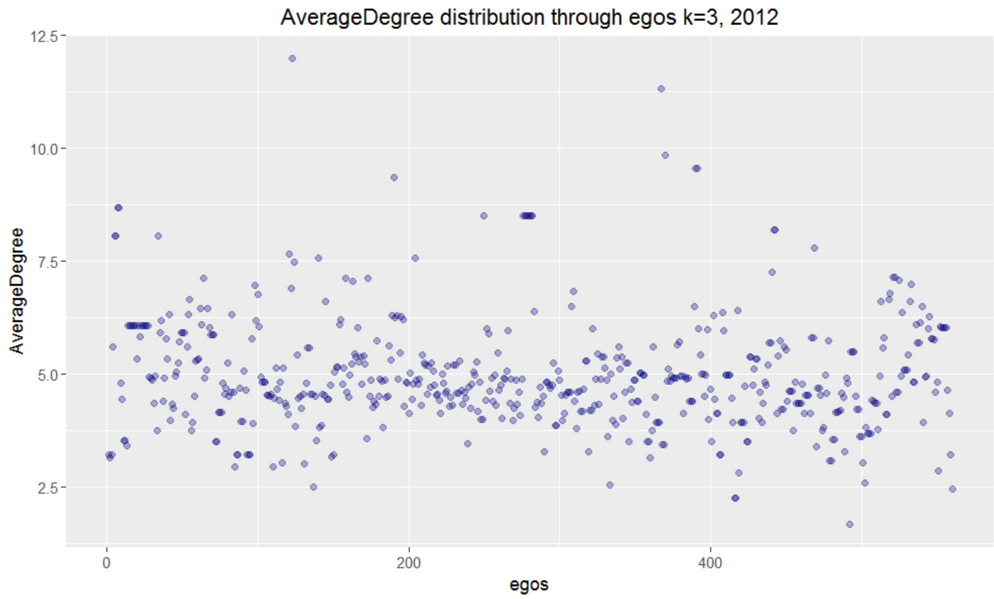


(b) k=1, 2012

Figure 4.31: Average degree of egos for $k = 1$ on $year = 2006$ and $year = 2012$, for $dataset_1$.



(a) k=3, 2006



(b) k=3, 2012

Figure 4.32: Average degree of egos for $k = 3$ on $year = 2006$ and $year = 2012$, for $dataset_1$.

Betweenness centrality. We evaluate the ego betweenness centrality over time in order to observe how its importance is affected when the personal network is evolving.

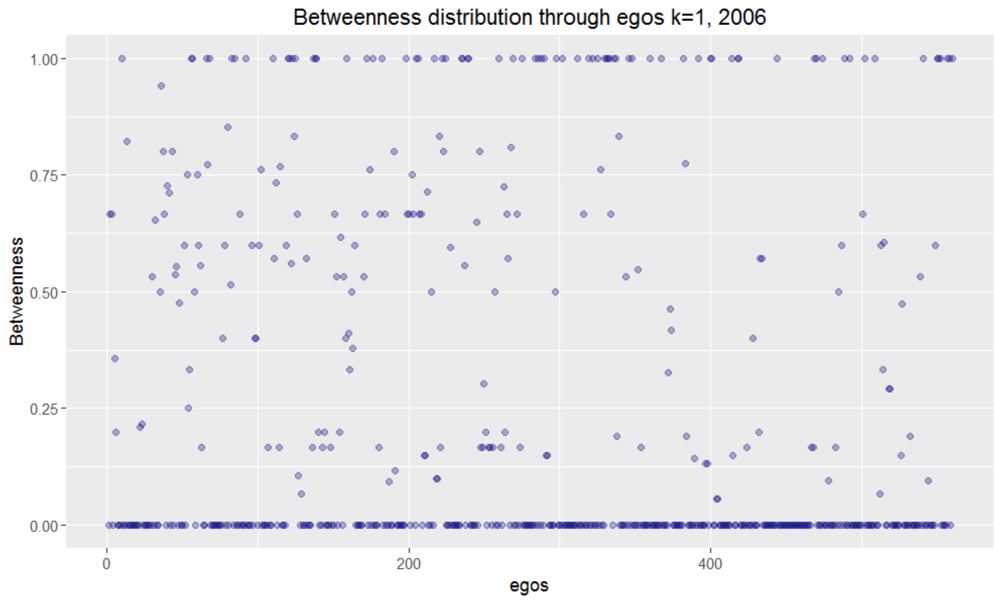
- **Observations regarding k:** For $k = 1$, we observed that a significant number of egos have a betweenness of 1 in their personal networks as we can observe in Figure 4.33, which reflects the case of star networks where the ego is the unique intermediate between its alters.

  Then, from k =2 to 5, this trend is not present anymore, as more nodes are included in the personal networks. We also observed, that many egos have a betweenness of 0 regardless of k, which denotes that no shortest path between any pair of nodes is passing through the ego as presented in Figure 4.34 for $k = 2$ either on 2006 and 2012. In this situation, the ego is considered as not important since its alters can reach each other without passing through it. This is an important conclusion for the information diffusion in personal networks because it shows that many times the personal network evolves without the active participation of the initiating node and that the ego node does not influence significantly after a point in time the other members of the network.

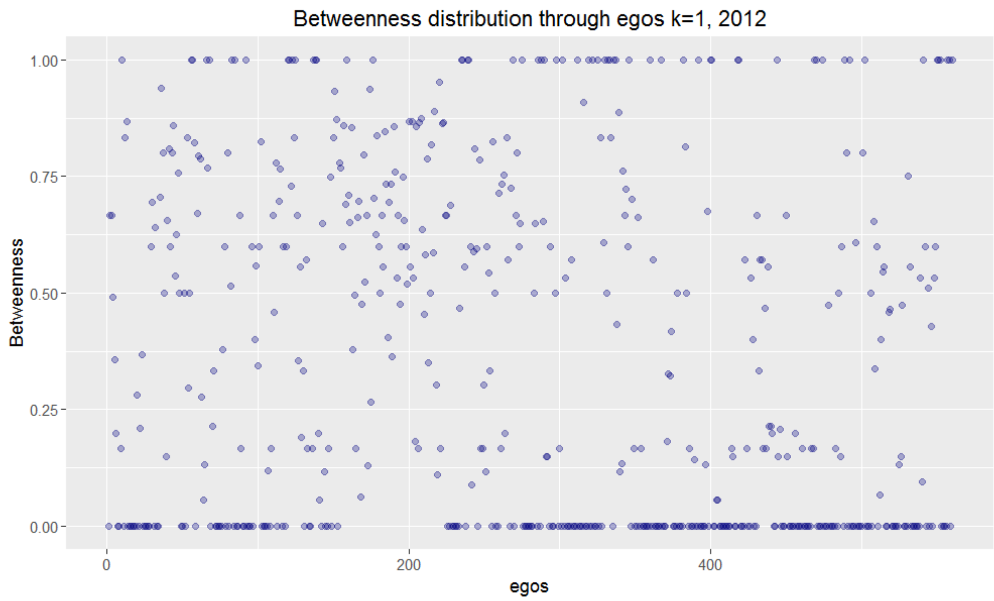
- **Observations regarding the years:** The ego betweenness evolution over time seems from the graphics increasing for $k = 2$ and $k = 3$ and decreasing for $k = 4$ and $k = 5$.

  We performed a Wilcoxon test to verify the observed trends for the betweenness evolution. The test results are given in Table 4.9 for the betweenness value of personal networks between 2006 and 2008, and between 2008 and 2010 for $k =$ 2 to 5. The values $p - value_{decreasing}$ and $p - value_{increasing}$ representing the p-value of the Wilcoxon test whether the alternative hypothesis is "greater" or "less" respectively are given. We conclude the following:

  - k=2 : $p - value_{decreasing}$ equals 1, while $p - value_{decreasing}$ is less then 0.05 which confirms the increasing trend of the betweenness for $k = 2$.
  - k=3: from 2006 to 2008, $p - value_{decreasing}$ equals 1, while $p - value_{decreasing}$ is less then 0.05. This confirms the observed increasing tendency of the betweenness. However, from 2008 to 2010 the null hypothesis can not be discarded and the alternative ones cannot be retained as $p - value_{increasing}$ and $p - value_{decreasing}$ are higher than 0.05. This could mean that there is not a significant change of the betweenness between 2008 and 2010.



(a) k=1, 2006



(b) k=1, 2012

Figure 4.33: Ego betweenness centrality for $k = 1$ and $year$ =2006 and 2012, for $dataset_1$.

– k=4: from 2006 to 2008, $p - value_{decreasing}$ approaches 1 which indicates that the betweenness is not decreasing as we could observe from the graphics. It is not neither increasing as $p - value_{increasing}$=0.0663 $>$ 0.05, but $p - value_{increasing}$ remains close to 0.05 which indicates a willingness to decreasing. Indeed, from 2008 to 2010, the betwenness decreases since

$p - value_{increasing}$=1, while $p - value_{decreasing}$ is significantly lower than 0.05.

– k=5: the betweenness is decreasing since $p - value_{increasing}$ is approaching (0.9793 from 2006 to 2008) or is equal to 1 (from 2008 to 2010), while $p - value_{decreasing}$ is lower than 0.05.

Thus, we can conclude that the Wilcoxon test allowed us to confirm the observations from the graphics with more precision as in the case of $k = 3$ and $k = 4$ the evolution trend was not obvious at a given period in time.

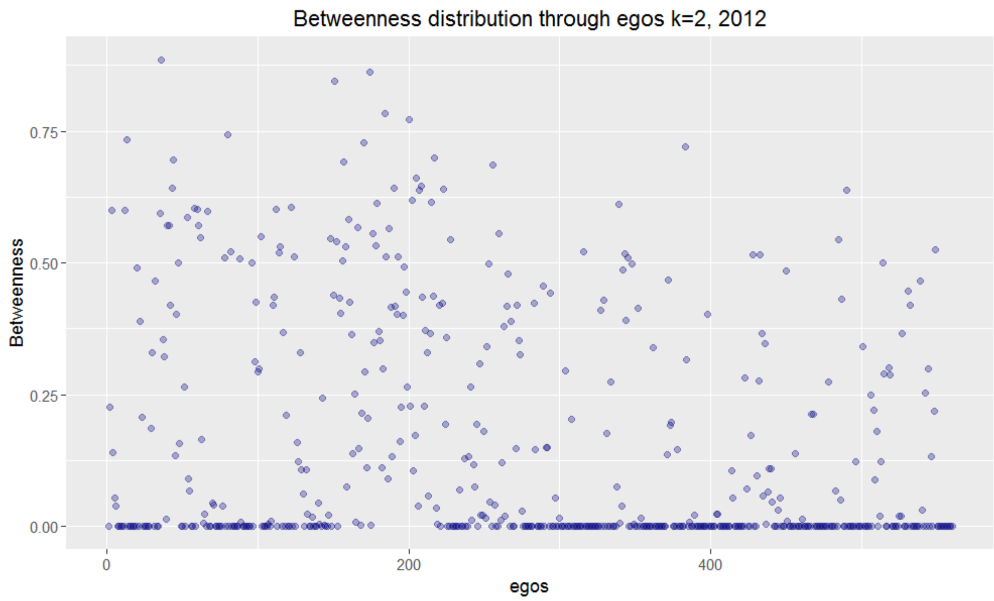
We explain the fact that it increases for $k = 2$ and $k = 3$ by the addition of new nodes that connect to ego's alters but, in order for these new nodes to reach the other alters of the ego, they have to go by the ego and thus ego's importance is increasing. We give in Figure 4.34, an example for the increasing of the betweenness for $k = 2$.

Year	2006-2008	2008-2010
k=2	$V = 5053.5$, $p - value_{decreasing} = 1$, $p - value_{increasing} = 1.28e - 10$	$V = 10784$, $p - value_{decreasing} = 0.9998$, $p - value_{increasing} = 0.000163$
k=3	$V = 7137.5$, $p - value_{decreasing} = 1$, $p - value_{increasing} = 8.859e - 06$	$V = 16490$, $p - value_{decreasing} = 0.2419$, $p - value_{increasing} = 0.7584$
k=4	$V = 9000$, $p - value_{decreasing} = 0.9339$, $p - value_{increasing} = 0.0663$	$V = 20700$, $p - value_{decreasing} = 2.125e - 10$, $p - value_{increasing} = 1$
k=5	$V = 11274$, $p - value_{decreasing} = 0.0208$, $p - value_{increasing} = 0.9793$	$V = 20588$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$

Table 4.9: Wilcoxon test on the Betweenness, $dataset_1$

In the context of scientific collaborations, this situation is taking place when, for example, one of ego's co-authors makes a new collaboration with a new set of authors. Then, the path between one of the new collaborators and an old ego's co-author includes necessarily the ego. For $k = 4$ or 5, one plausible explanation is that since (as discussed earlier) cliques have already started forming then in this case of personal networks, new collaborations happen between cliques in a more complete way (two groups collaborating) and thus the evolved network contains a new bigger clique. Having local cliques decreases the importance of the ego.

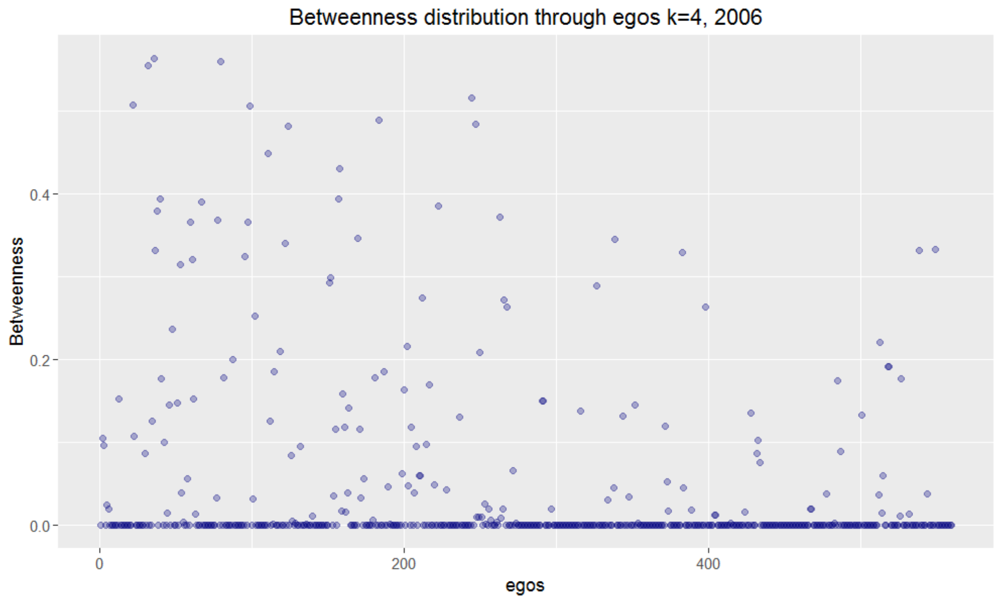
(a) k=2, 2006



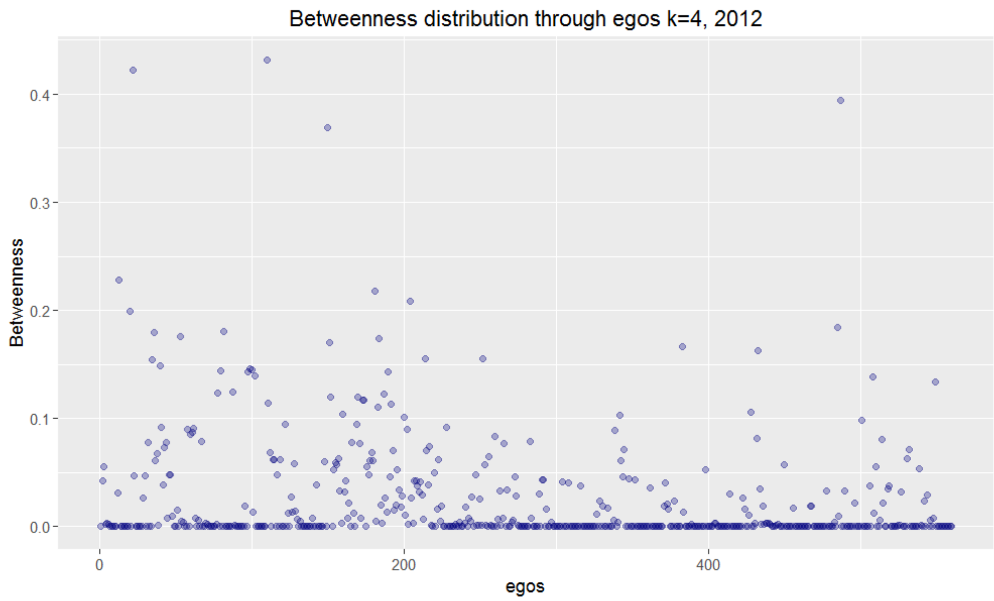
(b) k=2, 2012

Figure 4.34: Ego betweenness centrality for $k = 2$ and $year$ =2006 and 2012, for $dataset_1$.

Another plausible explanation could be that the ego is not getting new connections (collaborations) as the ego in the network in Figure 4.36, or are making few ones, where one (or more) of his alters (co-authors) are leading the main dynamics inside its personal network (as the alter "383" inside the circle). We give in Figure 4.35, an example for the decreasing of the betweenness for $k = 4$.



(a) k=4, 2006



(b) k=4, 2012

Figure 4.35: Ego betweenness centrality for $k = 4$ and $year$ =2006 and 2012, for $dataset_1$.

Local clustering coefficient & Efficiency. We group the discussion about the results of the local clustering coefficient defined in Definition 2.2.5, and the efficiency defined in Definition 2.3.12 because both metrics are computed at the ego level and concern only 1-personal networks ($k = 1$).

We have observed that both the local clustering coefficient and the efficiency

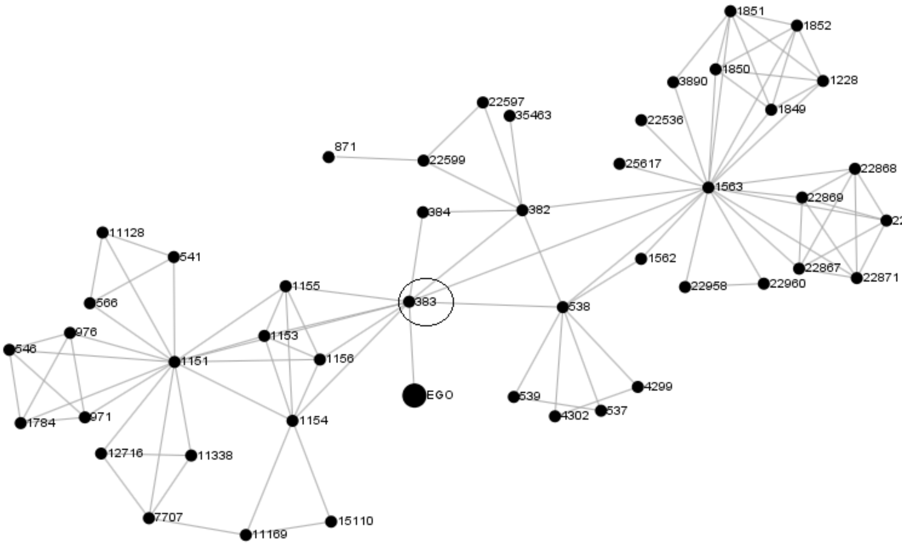


Figure 4.36: Personal network of ego "EGO", for $k = 4$, year=2008 .
Example for low ego betweenness.

decrease over the years, when new nodes integrate the 1-personal networks. This indicates that ego's alters are less connected with each other over time; and it means that new collaborations for 1-level personal networks happen mainly with the ego and usually do not involve many of the other existing collaborators. This explains also the fact that the average clustering coefficient in a 1-level personal network is increasing; the new connections are usually "complete" either with the ego or with the ego and few of the existing collaborators so we have all the possible triangles materialized (the local clustering coefficient of the alters equals 1 which increases the average clustering coefficient of the 1-personal network). One the other hand, for the ego this means that it has a lot of incomplete triangles since the new nodes will not participate in materializing the majority of the possible triangles with the rest of the alters.

4.4.4 Results summary and discussion

In Table 4.10, we summarize the main observations that we made in the previous section, regarding the evolution of the metrics in the co-authorship personal networks. We report in the second column the observations regarding the parameter k, while in the third column we present the observed evolution trends for each metric over the different time steps (years).

In the following subsections, we firstly discuss the metrics' evolution of co-authors personal networks for the special case when k=1 and then we discuss the more general case of $k > 1$.

Metric	Observations regarding k	Observations over years
Ego Degree Centrality (deg_e) (only for $k = 1$)	few egos with a high degree, low degree (2 or 3) frequent	OPN keeps the same tendency while evolving
Local Clustering Coefficient (C_e) (only for $k = 1$)	many networks have $C_e = 1$ some have $C_e = 0$	decreasess
Efficiency (only for $k = 1$)	-	decreases
Density (D)	$\forall k \in \{1, 2\}$, D is high	$\forall k$, decreases
Global Clustering Coefficient (GCC)	$k = 1$, high GCC	$\forall k$, decreases
Average Clustering Coefficient ($< C >$)	$\forall k$, high $< C >$	$\forall k$, increases
Number of nodes and edges	-	$\forall k$, both increase
Average Degree	for $k = 1$, around 1, 2 for $k \geq 2$, around 4, 5	$\forall k \in \{2, 3, 4, 5\}$, increases
Ego Betweenness Centrality (B_e)	$k = 1$, for many networks $B_e = 1$ $\forall k$, for many networks $B_e = 0$	$\forall k \in \{2, 3\}$, increases $\forall k \in \{4, 5\}$, decreases
Power law distribution	$\forall k \in \{1, 2\}$, verified $\forall k \in \{3, 4, 5\}$, less verified	$\forall k \in \{1, 2\}, \forall year$, verified in more than 80% of cases $\forall k \in \{3, 4\}$, less verified especially for $year \geq 2011$ for $k = 5, \forall year \geq 2011$, not verified
$K - max$	-	$\forall year \geq 2011$, $k - max > 13$ for all the personal networks

Table 4.10: Summary of the observations of metrics' evolution.

4.4.4.1 Observations for the evolution of 1-personal networks

From our observations, 1-personal networks constitute a particular case. Indeed, a significant proportion of 1-level personal networks are small star networks with few nodes characterized with an ego betweenness centrality of $B_e = 1$ along with a weak ego degree. For example, if an ego has 2 connections as in Figure 4.37 for ego "5" connected to nodes "3" and "8", these two connections correspond to two independent collaborations, the ego in this case plays a key role given its important position. Furthermore, in other cases we observe the exact opposite phenomenon: a considerable set of 1-personal networks is consisting of complete graphs, characterized by an ego betweenness of $B_e = 0$, because of the co-authorship network's specificity where when we have many coauthors in a paper, these coauthors are connected in a complete subgraph. In Table 4.11, we give the percentages of egos having a betweenness equal to 0 and a betweenness equal to 1 over the years (from 2006 to 2013), when $k = 1$.

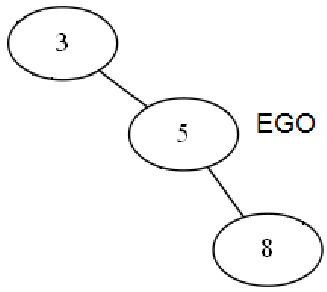


Figure 4.37: 1-Personal network of an ego with 2 alters.

For both the egos' degrees and egos' betweenness, when the network evolves over time, the situation remains the same. More precisely, the evolution affects mostly the connections among alters who were characterized at the beginning with an important number of connections and then become less connected over time as new alters join the 1-level personal network (the density, the local clustering and the efficiency decrease). Moreover, the addition of nodes (alters) keeps the average degree of the 1-level personal network low.

4.4.4.2 Observations for the evolution of personal networks with $k > 1$

The personal networks with $k = 2, 3, 4, 5$ are characterized by both a decreasing density and global clustering coefficient over the years. The nodes joining the personal network over the years create fewer edges compared to the possible number of edges that can be created, and thus the density gets lower, as well as the transitivity at the personal network level.

At the local level, the average clustering coefficient was observed to be high and gets higher when the personal networks are growing which indicates a high local clustering around the nodes. We can explain that by the fact that co-authorships are made generally between 3, 4 or 5 authors and the emergence of a new collaboration implies the creation of a highly clustered local structure since all the authors that made this collaboration are going to be already connected with each other. The trends we discussed so far concerns all personal networks with k from 2 to 5. Nevertheless, ego's betweenness centrality behaves differently: it was increasing for $k = 2, 3$, but decreases for $k = 4, 5$. On the one side, the increase for $k = 2, 3$ is due to the addition of nodes around the ego that form disconnected groups of alters (as the windmill

Betweenness	2006	2007	2008	2009	2010	2011	2012	2013
0	58.57	53.21	49.64	46.07	42.86	41.07	39.64	39.64
1	13.39	12.14	11.25	10.71	10.71	10.71	10.71	10.71

Table 4.11: Ego betweenness extreme values percentages over years, $k = 1$.

graph example) and the ego plays the role of a linker between them. On the other side, when we consider a $k = 4$ or 5, as discussed above, existing cliques might merge into bigger and more connected ones and thus the ego's betweenness decreases. This hypothesis was not validated experimentally yet.

Additionally, the power law distribution of degrees is better satisfied when $k = 2, 3$ than $k = 4, 5$. In addition, for $k = 5$, the networks are not following a power law degree distribution starting from 2011.

4.5 Analysis of co-authorship personal networks evolution via cliques

In the previous chapter, the metrics' evolution put forward the fact that the co-authorship personal networks are strongly locally clustered and less clustered globally; these conclusions were outlined from the evolution of the average and the global clustering coefficients but also from the evolution of the ego's betweeness. The evolution of both clustering coefficients allowed us also to understand that the personal networks that we are studying have the form of windmill graphs, and that can be explained by the nature of the networks, because they are co-authorship networks. Indeed, in a co-authorship network, a new publication joining the network implies the creation of a clique: all the coauthors of the publication are connected with each other, whether if they are new members of the network or already part of.

Given these conclusions made in the previous analysis, in this second analysis we are interested in discovering the patterns governing co-authorship personal networks when considering maximal cliques, i.e. the largest cliques. Our aim is to verify if the hypothesis with which we explained the behaviour of the metrics previously are valid, but also to find out new characteristics of the networks through the composition and its evolution of networks in cliques. In the following, we present the set of analysis done to cover these elements.

We would like to remind that the computations were performed on the same data, more precisely on the set of 560 personal networks of the authors who started publishing on 2004 evolving from 2006 until 2013. The graphics that will be presented in this section focus on the evolution between 2006 and 2012. The choice of not analyzing the complete period (from 2004 to 2013) was governed by the fact that we noticed that during the 2 first years the networks were not evolving a lot and this can be explained by the fact that during the first years the researches tempt to publish generally with the same persons. We also noticed that the evolution from 2012 to 2013 is quite limited. It is also important to note that we study maximal cliques composed of minimum 3 nodes.

The notion of cliques and maximal cliques was introduced in Chapter 2 as a form of cohesion in social networks. In the following, we use *cliques* to designate *maximal cliques*.

4.5.1 Number of cliques

In order to understand if the networks are more locally clustered, we will observe the number of the cliques per OPN. If this number is important, we can conclude that the networks are locally clustered. To observe the number of the cliques per OPN and its evolution, we plot the number of OPNs having a specific number of cliques in 2006 and in 2012 for $k = 1...5$. The result is given in Figure 4.38.

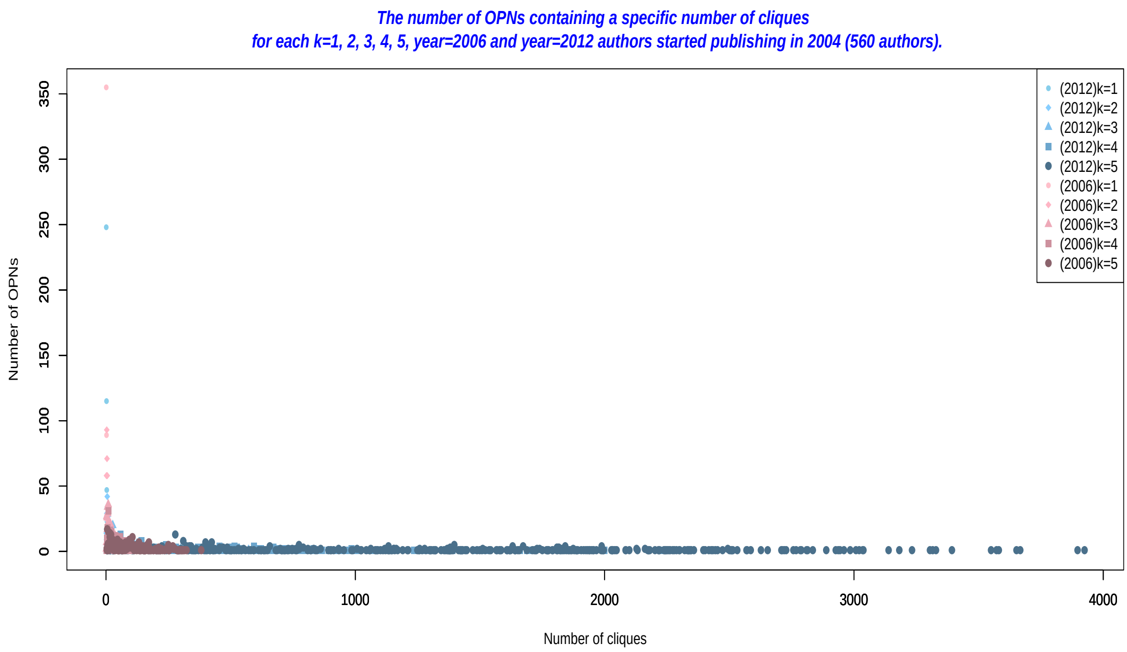


Figure 4.38: Number of OPNs having a specific number of maximal cliques for year 2006 (pink) and year 2012 (blue) and for k=1, 2, 3, 4, 5.

This figure, difficult to analyze given the density of the points, allows us nevertheless to make the following conclusions:

1. in an OPN, we can have from *1* till *4000 maximal cliques* and that these values are distributed quite uniform for both years in the tail of the curve;
2. the OPNs that have between 3000 and 4000 cliques are from 2012 and they have $k = 5$. The maximum number of cliques in 2006 is much less than 1000, but this is not clear from the current graphics and it needs further investigation;
3. more than 100 OPNs with $k = 1$ has a specific small number of cliques: in 2006, they are more than 350 OPNs, and in 2012 more than 250 OPNs. This observation should be further investigated in order to understand why these networks are not locally clustered;
4. for more than 3000 cliques, the density of OPNs decreases, while before 1000 cliques the density of OPNs increases. This latter part should be further investigated;
5. we should also further investigate the cases where the OPNs have a small number of cliques (close to 0) in order to understand, as said previously, why these networks are not locally clustered.

In the following, we will focus on some parts of the graphics in Figure 4.38 in order to better understand the phenomenons described above.

Number of cliques on 2006. In this part, we will try to conclude on the 2nd point above related to the maximal number of cliques that an OPN in 2006 can have. From Figure 4.39, showing only the OPNs in 2006, we can observe that the maximum number of cliques that we can have in an OPN in 2006 is a bit higher than 300, and only OPNs with $k = 5$ have more than 200 cliques. Nevertheless, we can consider the OPNs in the tail of this graphic as locally clustered.

Maximum of 100 cliques. To answer the questions on the points 3 and 5 in the above conclusions concerning of having networks not locally clustered, we focus in Figure 4.40 on the OPNs having a maximum number of cliques of 100 because in Figure 4.39 we can see that between 1 and 100 we have a high number of networks but also more dynamics on the number of networks.

Figure 4.40 allows us to conclude on several subjects:

- first of all, we can indeed observe with the focus on the graphic of Figure 4.39, given in Figure 4.40, more change than in the tail of the graphic.

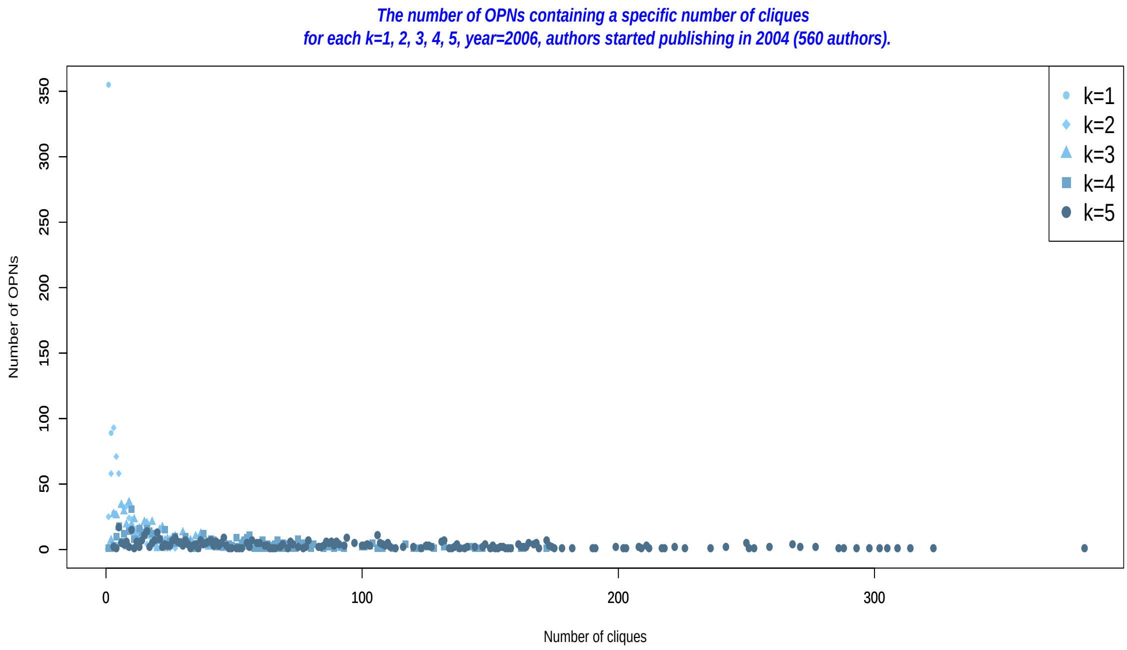


Figure 4.39: Number of OPNs vs number of maximal cliques for year=2006 and k=1-5.

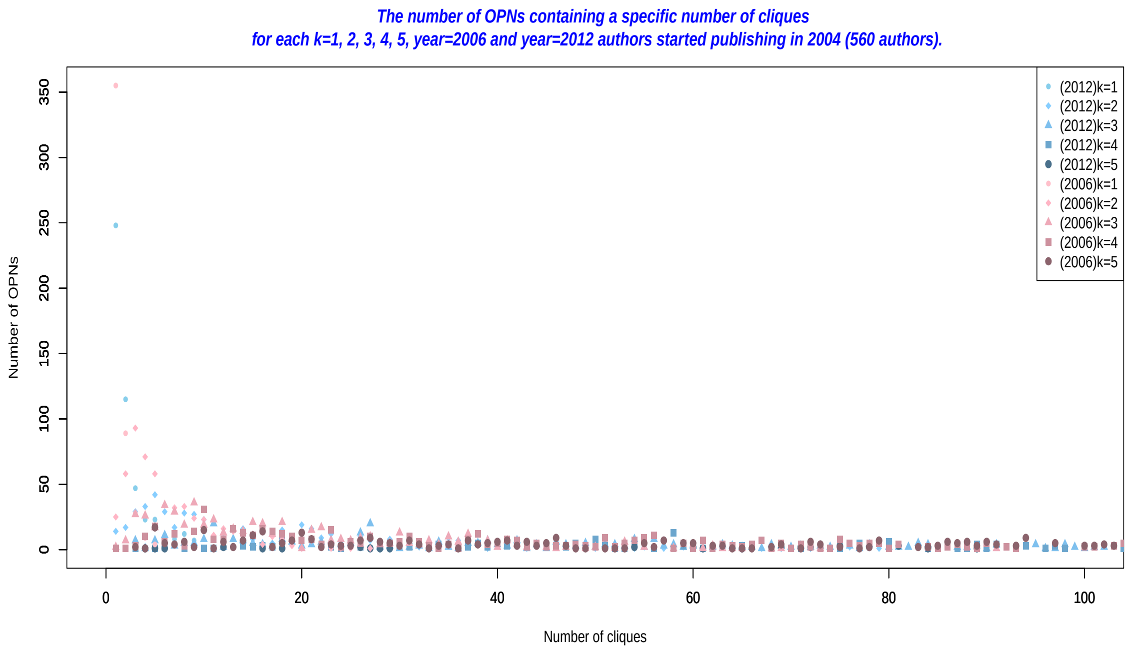


Figure 4.40: Number of OPNs vs number of maximal cliques for year=2006 and year=2012 limited to 100 maximal cliques.

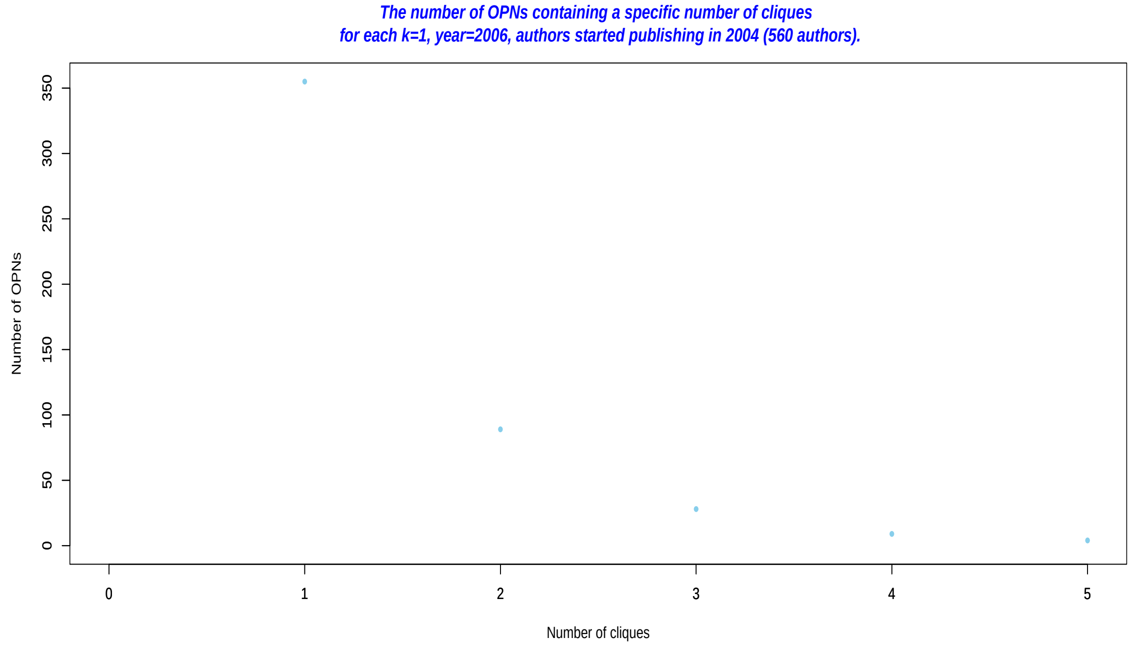
- last, we can see that there is one or two small values of the number of maximal cliques contained in OPNs corresponding to a high number of OPNs (more than 50 OPNs, till 350 OPNs). This can contradict the conclusion from the previous chapter stating that the studied networks are locally clustered. In addition, we can also outline that the OPNs containing this number of maximal cliques have $k = 1$.

Maximum 5/16 cliques and k=1. Given the previous conclusion, it is interesting now to understand how many cliques exist in OPNs when $k = 1$ for each year. Figure 4.41 represents the number of OPNs vs the number of cliques only when $k = 1$ for each year. From this graphic, we can outline the fact that:

- in 2006, OPNs have 5 cliques maximum, and in 2012, OPNs have 16 cliques maximum, when $k = 1$;
- moreover, over the 560 OPNs, we can conclude that, in 2006, more than 350 OPNs have 1 maximal clique, and in 2012, a bit less than 250 OPNs have 1 maximal clique. This corresponds to the conclusions that we had in the previous section regarding the fact that the OPNs with $k = 1$ are very often complete networks. This allows us to outline that OPNs with $k = 1$ have a different structure than the OPNs with a higher k, and thus they can have a different behaviour.
- we can also note that the number of OPNs decreases with the increase of the number of maximal cliques. This outlines that OPNs with $k = 1$ and a high number of cliques are not very numerous.

Maximum 100 cliques and maximum 65 OPNs. To have a clear understanding on the evolution of the number of maximal cliques, we propose to focus on the OPNs having maximum 100 cliques, but also on a maximum of 65 OPNs because we already concluded that the numerouse OPNs having a small number of cliques have $k = 1$. This graphic is given in Figure 4.42, and we can make several observations.

One observation concerns the fact that between 15 and 65 OPNs with $k = 1, 2, 3$ contains between 1 and 20 cliques which is quite reasonable because OPNs with $k = 1, 2, 3$ are smaller than the OPNs with $k = 4, 5$ and can form less cliques. In order to understand which number of cliques gathers an important number of OPNs, we limit the number of cliques to 10. The result is given in Figure 4.43, and it allows us to understand that:



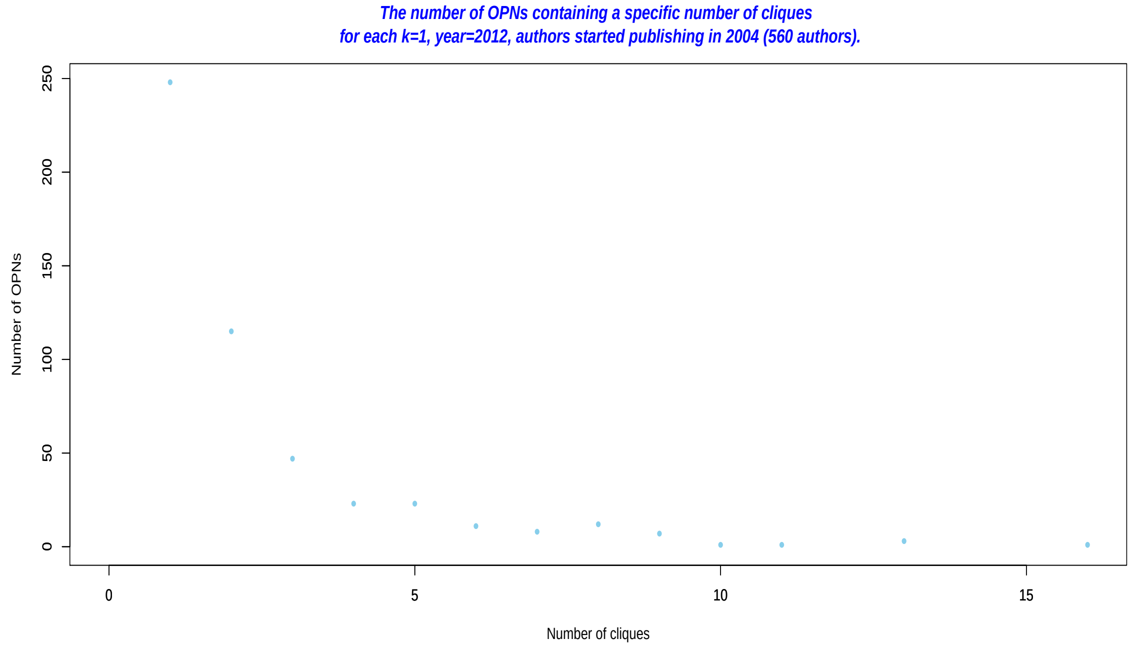


Figure 4.41: Number of OPNs vs number of cliques for year=2006 (top) and year=2012 (bottom) for $k = 1$.

- in 2006 and when $k = 1$, around 350 OPNs have only one clique, and around 120 OPNs have two cliques; this confirms our previous conclusions.
- in 2012 and when $k = 1$, around 250 OPNs have only one clique, and around 90 OPNs have two cliques; this also confirms our previous conclusions.

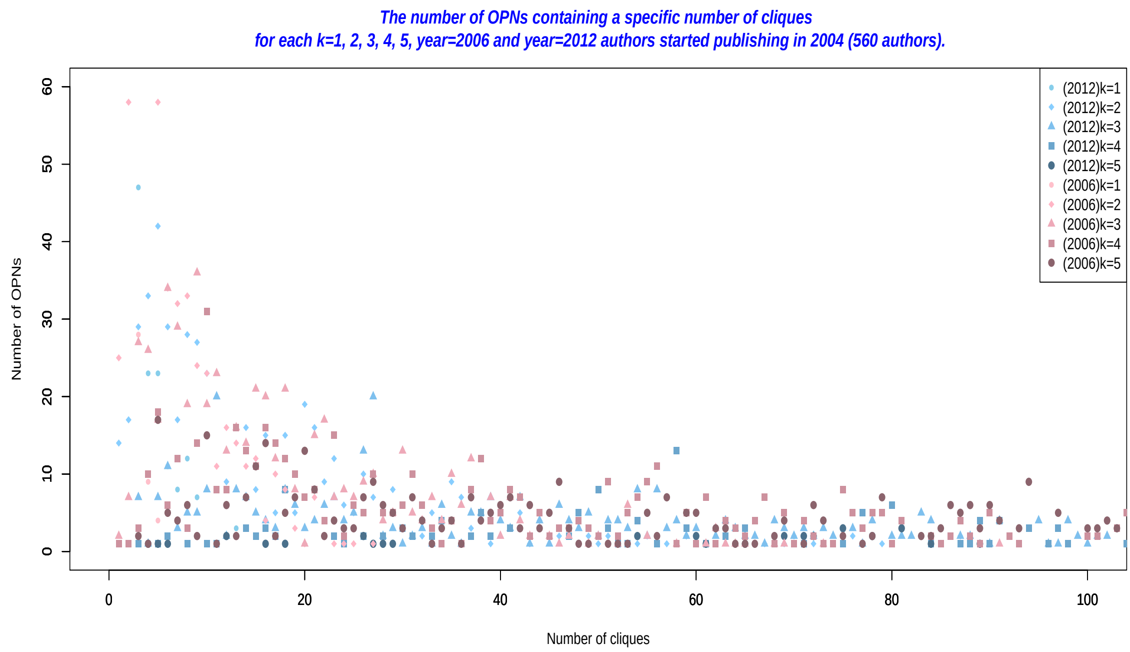


Figure 4.42: Number of OPNs vs number of maximal cliques for year=2006 and year=2012 limited to 65 OPNs and 100 maximal cliques.

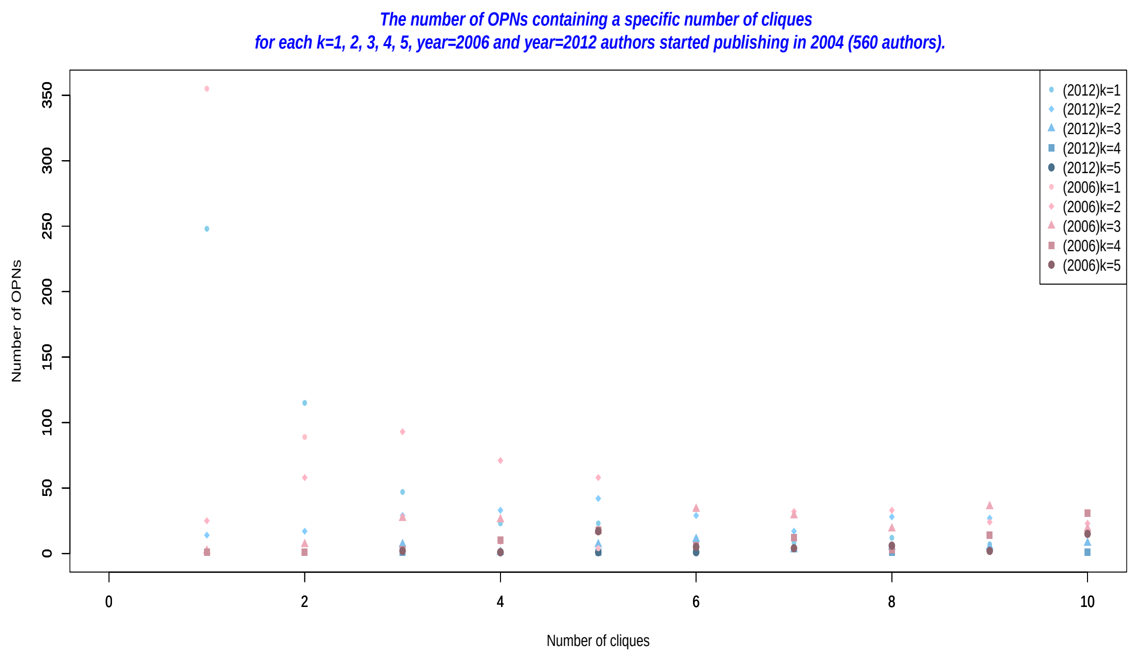


Figure 4.43: Number of OPNs vs number of cliques for year=2006 and year=2012 limited to 10 cliques.

- OPNs with a high k are on the bottom of the graphic, which means that very few OPNs with a high k have a small number of cliques. Nevertheless, we could also point out that a set of OPNs (around 30) with $k = 4$ (represented by the rectangle) have also a small number of cliques, less than 10; this case will the treated below.

The first conclusion stating that for $k = 1$, on the 560 OPNs, in 2006, 470 OPNs, and, in 2012, 340 OPNs have only one or two cliques is very important since it validates some of metrics evolution analysis results discussed in the previous section. Indeed, we previously mentioned that when $k = 1$, we had a considerable set of personal networks for which the value of certain metrics was extreme as it was the case for the local clustering coefficient where many networks had $C_e = 1$, and for the betweenness centrality where $B_e = 1$ for multiple OPNs. The OPNs exhibiting this behaviour are complete networks and so made of a unique clique. Thus, we proved via the analysis of the cliques that OPNs made of one clique are frequent when $k = 1$.

High k, low number of cliques. To study more closer the networks with $k = 4$ having few cliques, in Figure 4.44, we give an example of the personal network of the ego 27967 with $k = 4$ on 2012 having only 6 cliques which is the smallest number of cliques in an OPN with $k = 4$ in 2012. This example shows that OPNs can be of a small size even when $k = 4$ and this explains why they hold fewer cliques.

We can also notice the way cliques are connected: indeed, the connection between cliques is a bit particular since each pair of cliques shares only one node except the two 3-cliques to which the ego belongs which share two nodes (the ego and his alter '756'). Thus, we can confirm that, even if the network with $k = 4$ has a small number of maximal cliques, the network is still locally clustered. For this particular ego 27967, we observed the evolution for k=1 to 5 from 2006 to 2012. We found that the OPN does not evolve when considering k=1,2,3, and 4 along the years. The only evolution happens with k=5 where one single node and one single edge are added on 2007.

Number of cliques reported to number of nodes and edges. We analyzed the relationship between the number of cliques and the number of nodes composing the personal networks for the different k values and the different years. We discovered that there is a linear relationship between the number of nodes of a personal network and the number of cliques it contains, and this whatever k, and whatever the year. We give in Figure 4.45 the graphics illustrating this linear relationship. The relationship between the number of cliques and the number of edges is linear as well for all k values and whatever the year as we can see in Figure 4.47 for $k = 4$ on 2010.

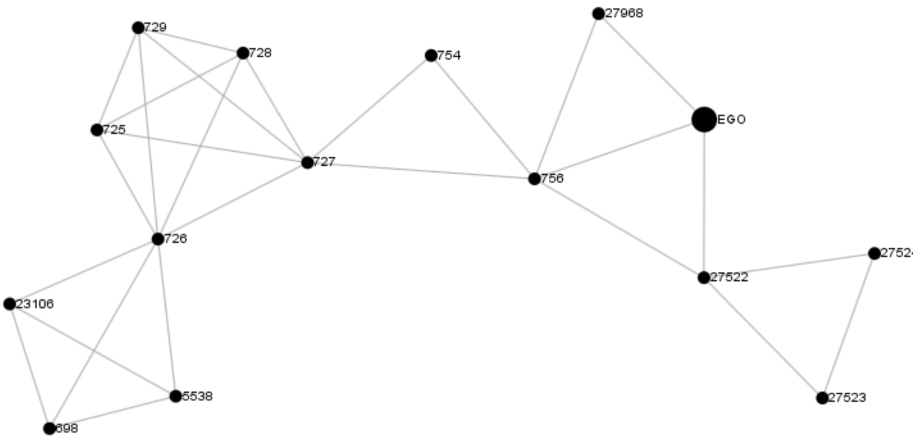


Figure 4.44: Personal network of the ego 27967 with 6 cliques, $k = 4$, 2012.

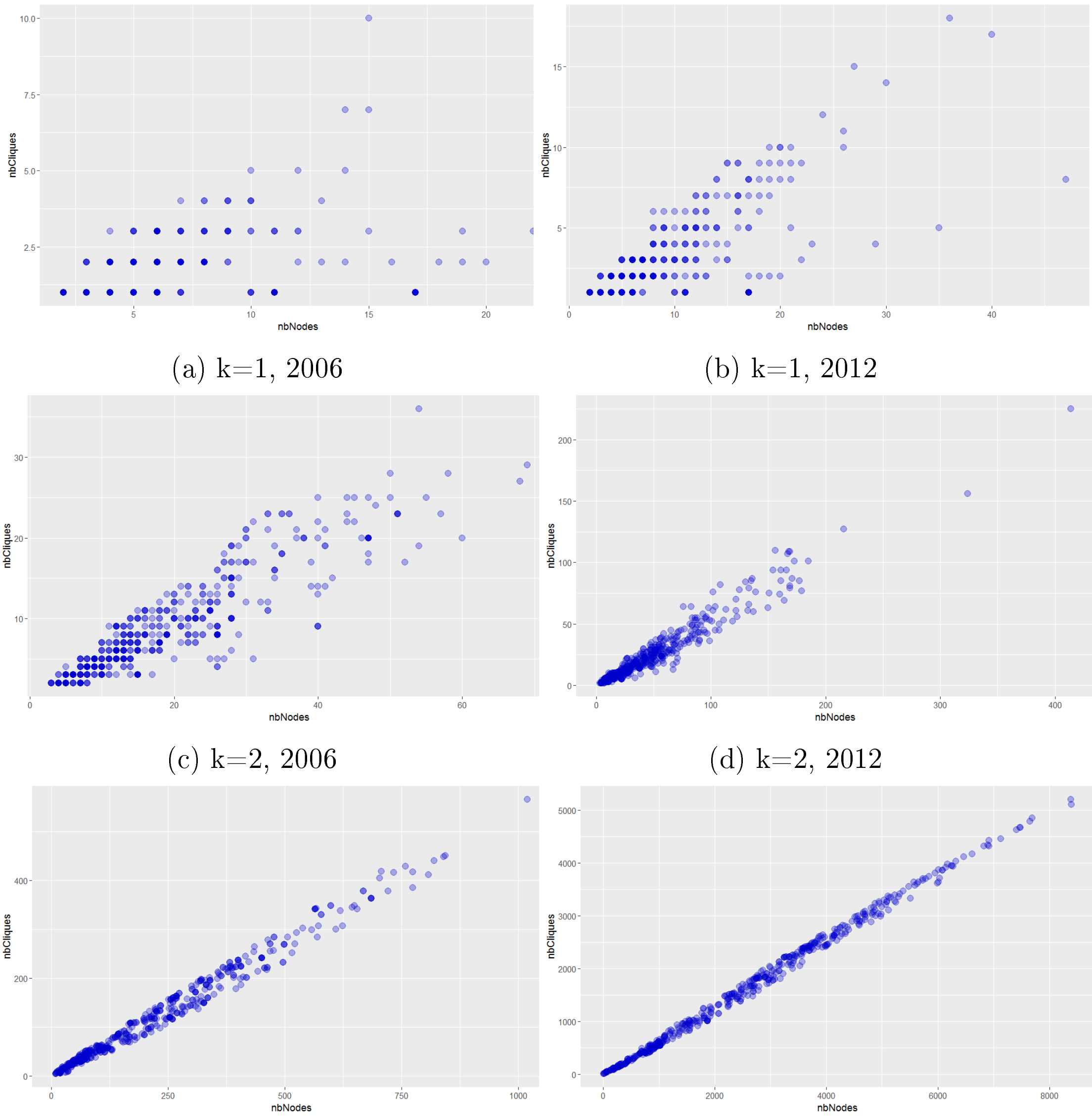
4.5.2 Cliques' size

In this section, we propose to understand if the OPNs behave the same when taking into account their composition in cliques. To this end, we produced in Figure 4.48 a graphics that represents the number of cliques of a given size in each OPN.

By observing the graphics for $k = 1$, $k = 3$ and $k = 5$, we can conclude that the OPNs contain in general more small cliques than big cliques. When $k = 1$, we can note that OPNs contain maximum 5 cliques of size 3, when $k = 3$, the OPNs can contain 40 cliques of size 3, and when $k = 5$ the number of 3-cliques in some OPN can be close to 200.

Besides the increase of the number of cliques per OPN with k, we can also note that the increase is not linear. Indeed, when $k = 3$, for the cliques of size 3, there are holes between the 30 and 40, which means that there are no OPNs containing for example 33 of 3-cliques.

For the same OPNs, in 2012, the situation is comparable to the one in 2006; the number of cliques with a given size in OPNs is decreasing with the size of the clique, and we can note important holes in the increase of the number of cliques per size of clique, as presented in Figure 4.49.



(a) k=1, 2006

(b) k=1, 2012

(c) k=2, 2006

(d) k=2, 2012

(e) k=5, 2006

(f) k=5, 2012

Figure 4.45: Number of cliques vs number of nodes per OPN.

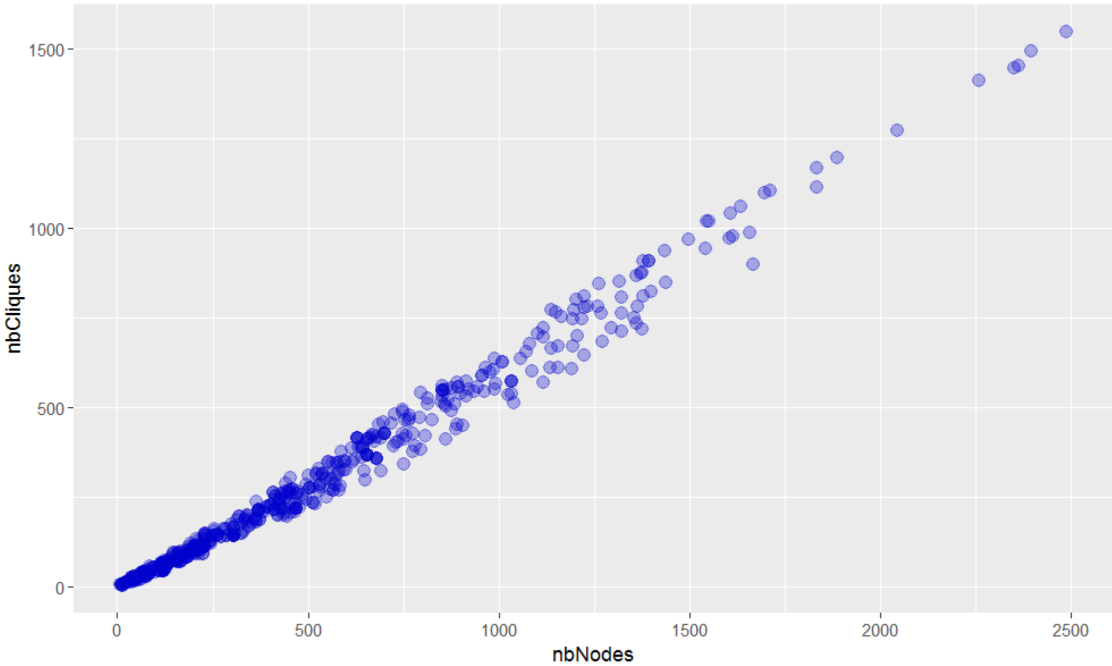


Figure 4.46: Number of OPNs' nodes vs number of cliques for $k = 4$ and year=2010.

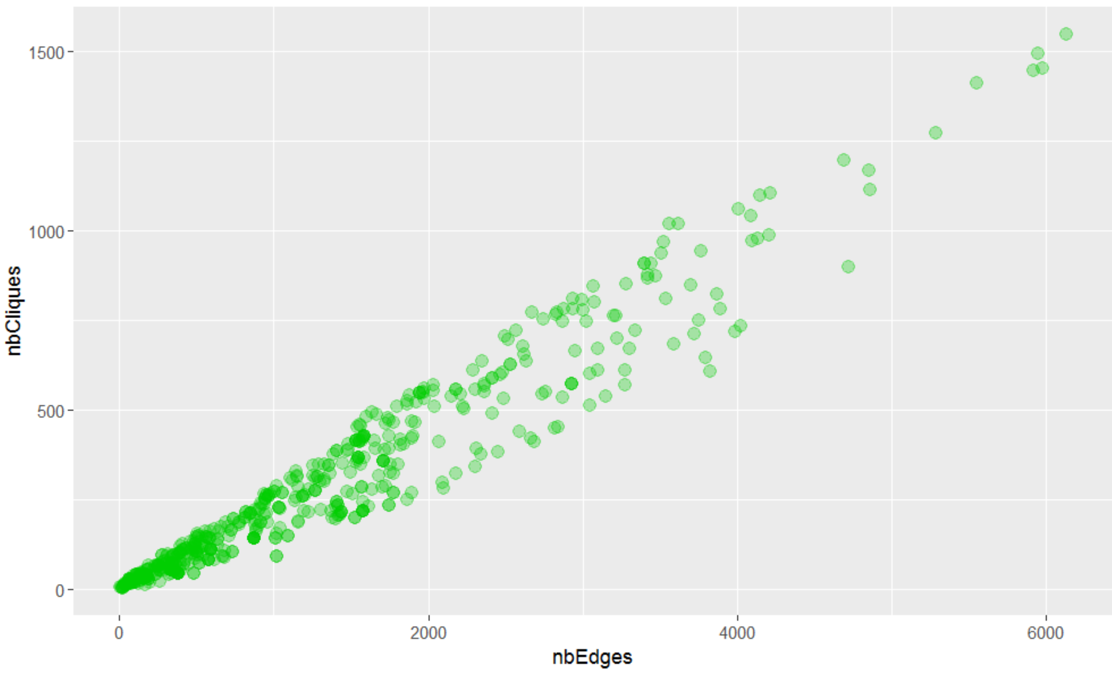


Figure 4.47: Number of OPNs' edges vs number of cliques for $k = 4$ and year=2010.

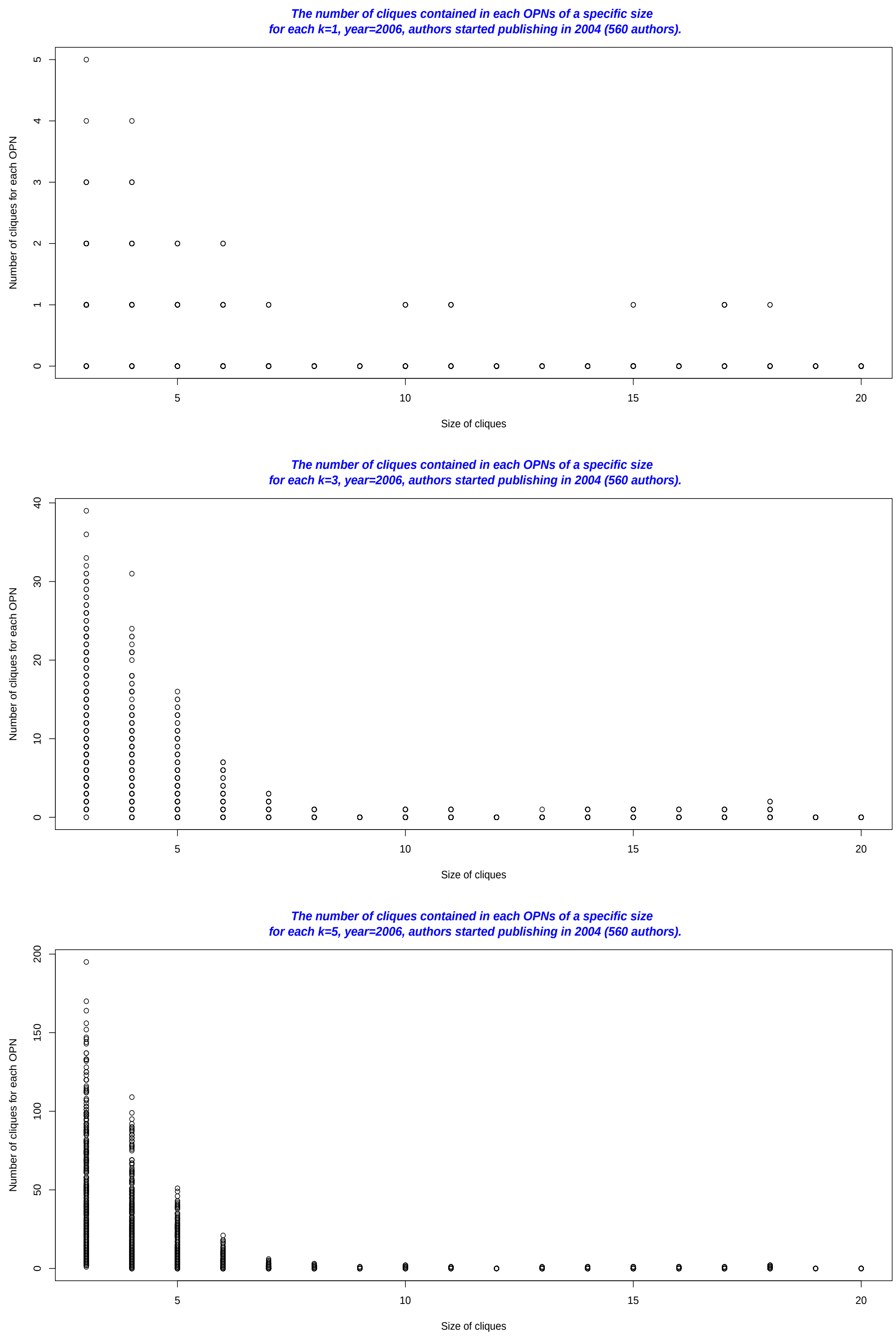


Figure 4.48: The number of cliques of a given size in each OPN in 2006 when $k = 1, 2, 3$ (top, middle and bottom).

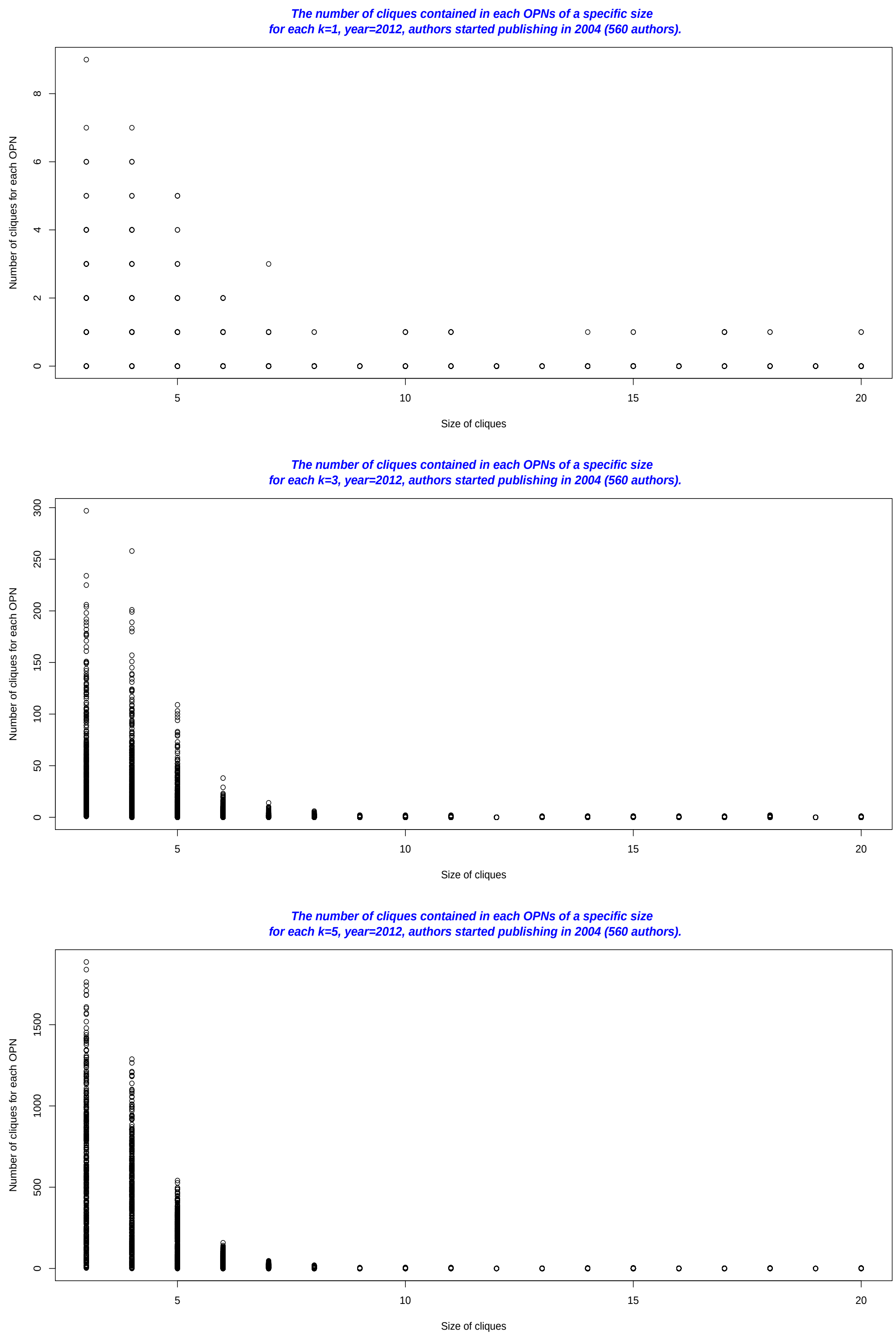


Figure 4.49: The number of cliques of a given size in each OPN in 2012 when $k = 1, 2, 3$ (top, middle and bottom).

4.5.3 Clique size power law

In Figure 4.50, we give the distribution of the size of the cliques for $k = 3$ on 2012. This distribution is characterized by the important presence of small sized cliques while big sized cliques are less and less frequent (the same distribution was observed for other k values and other years). Thus, we wanted to check both power law and Poisson distributions to verify which of these theoretic distributions represents better the distribution of clique size in our data. The test was performed for all k values and all the years over $dataset_1$.

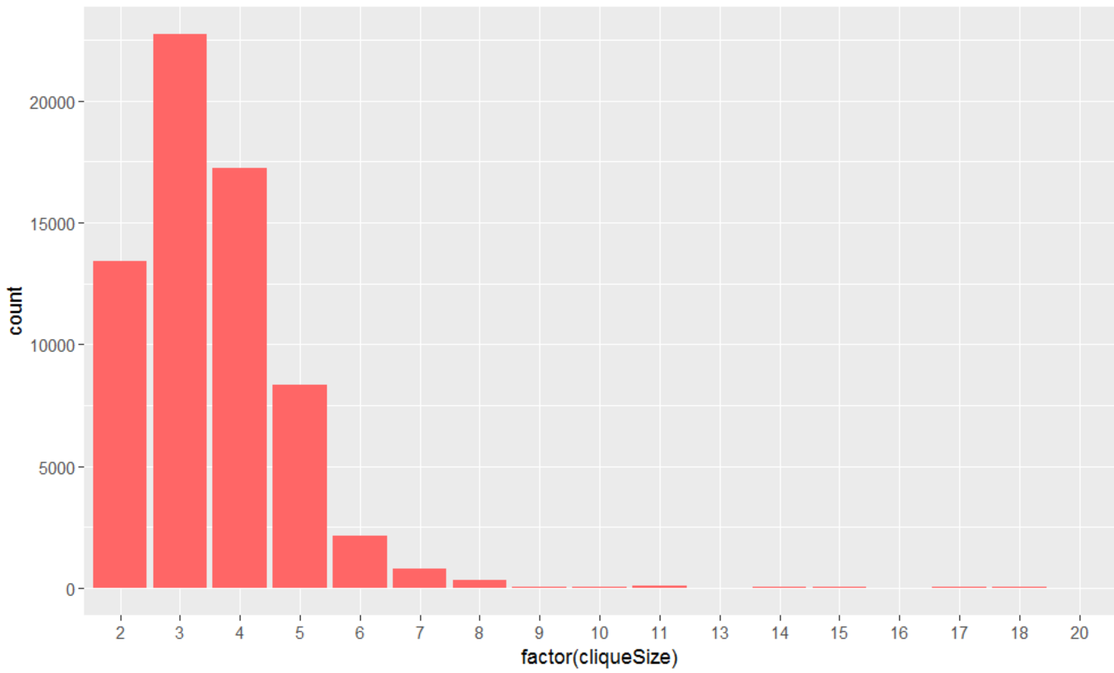


Figure 4.50: Clique size distribution, $k = 3$, 2012, $dataset_1$.

The result of the test are represented in the graphics in Figure 4.51 and Figure 4.52 for the power law and Poisson distributions respectively, and are presented hereafter:

- Power law distribution: the power law distribution is well verified when k equals 1, 2, and 3 and less for $k = 4$ and 5, and this whatever the year.

- Poisson distribution: the Poisson distribution is acting as complementary to the power law distribution given that it better represents the data when k equals 4, and 5 than when it equal 2 and 3. While we can see that it does not at all represent the data for $k = 1$ from 2008 to 2012.

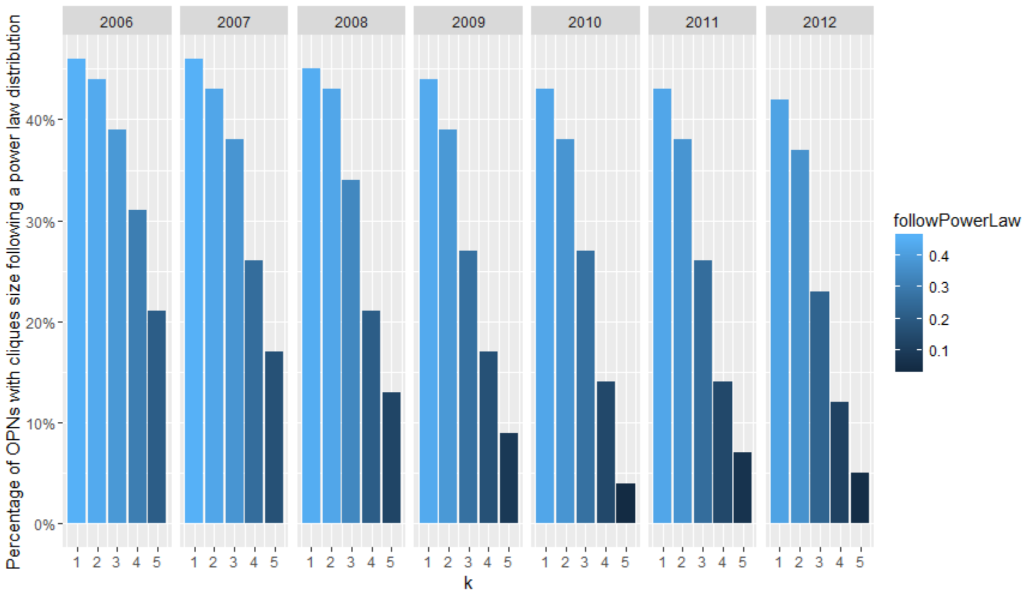


Figure 4.51: Power law distribution for the clique size, $dataset_1$.

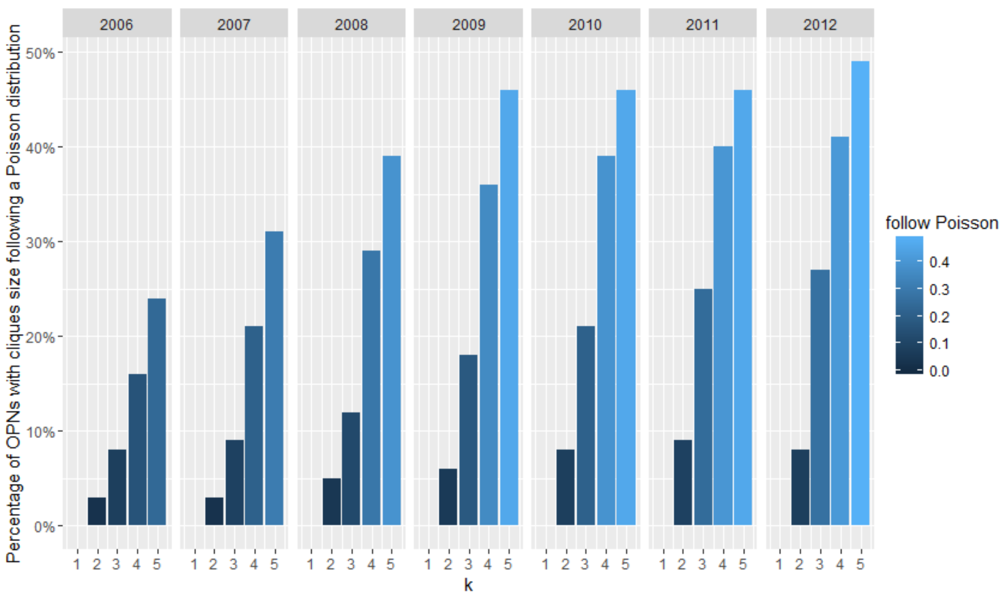


Figure 4.52: Poisson distribution for the clique size, $dataset_1$.

As a conclusion, the power law distribution can be used for reproducing the clique size distribution of the data when we consider $k = 1, 2, 3$, while for $k = 4$, and $k = 5$ the Poisson distribution is better representative.

4.5.4 Evolution of the size of cliques

In Figure 4.53 we provide a graphics of the number of OPNs vs the size of their cliques. A clique is used in the computation if it appears at least one time in the OPN, so no redundant information is used in the graphics. The observations in this graphics are very important:

- we can note that, per year, when k increases, the number of OPNs progresses to the right and to the top. This means that with k, the size of cliques is growing, but also the number of OPNs containing the cliques is growing.
- we can also outline that, comparing the OPNs in 2006 (blue) and 2012 (pink), the pink color is progressing to the left-top corner, which means that the more the year increases, the more the size of cliques increases and the number of OPNs containing the cliques increases.
- we can also point out, that the progression is not linear. A set of sizes of cliques is not very present in the OPNs, such as: the 9-cliques, the 12-cliques, the 13-cliques, the 17-cliques, and the 19-cliques, so it is maybe interesting to study sections of the data, like for example the one having OPNs with cliques of size from 3 to 8.

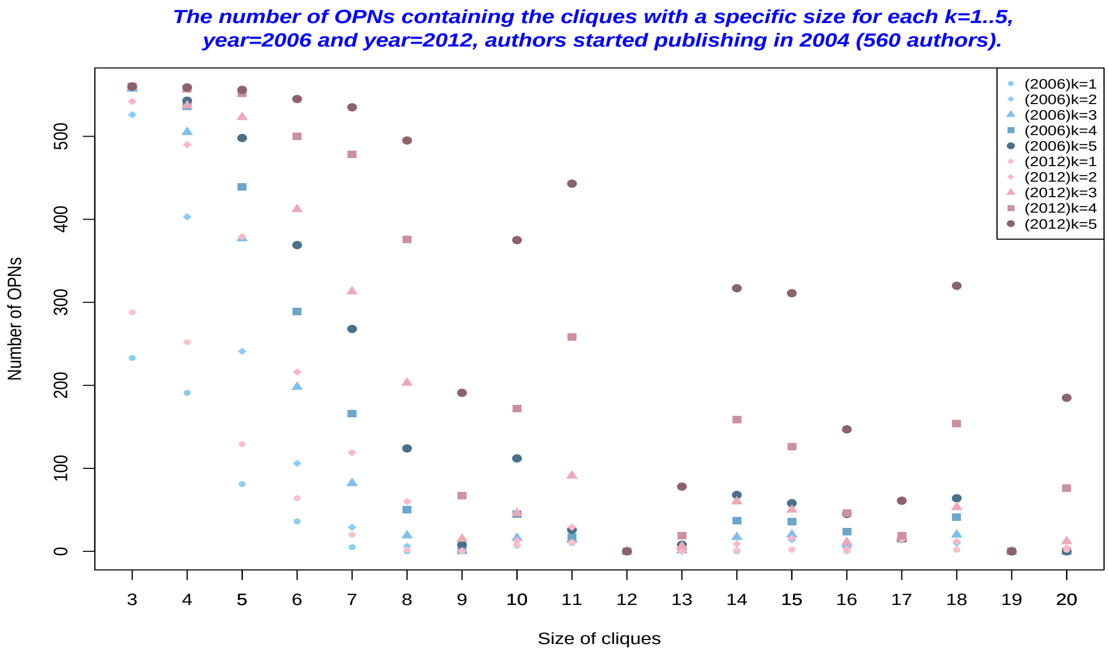


Figure 4.53: Number of OPNs vs size of cliques in 2006 and 2012, all k.

4.5.5 Minimum cliques size

For the minimum size of cliques, we plot the number of OPNs vs the minimum size of cliques; to compute the minimum size of cliques we computed, for all the OPNs, the minimum among the size of the cliques composing an OPN. The graphics is given in Figure 4.54 and it allow us to outline the following observations:

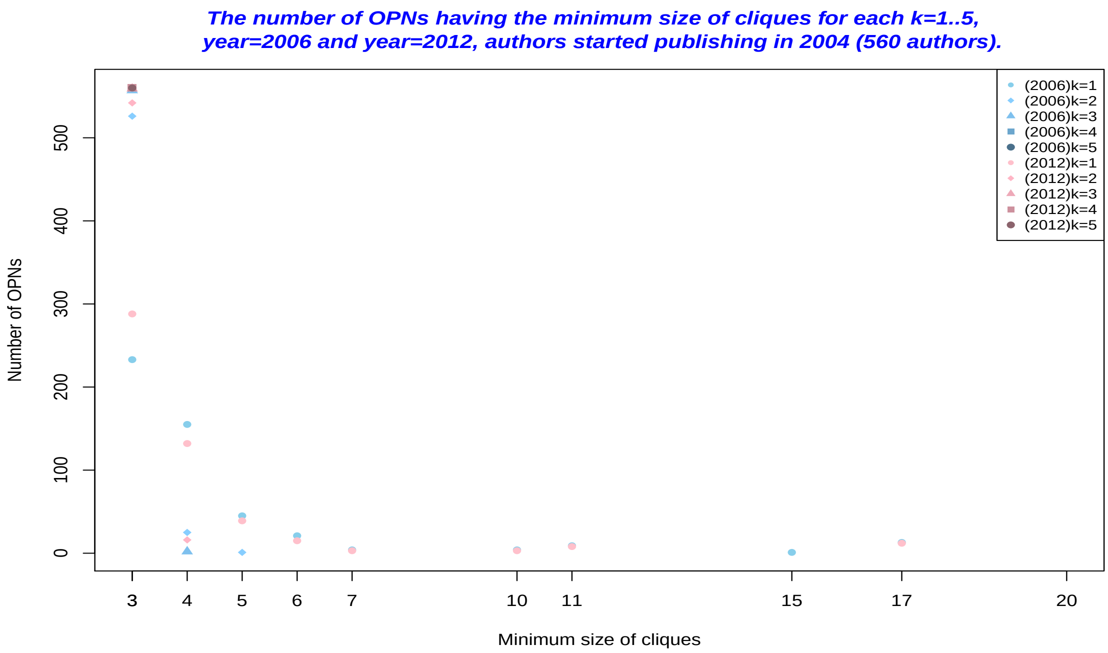


Figure 4.54: Number of OPNs vs minimum size of cliques in 2006 and 2012, all k.

- first of all, for a high number of OPNs the minimum size of its cliques is 3 (more than 500); this is valid whatever k. We recall that we discarded cliques of size 2 made between solely two nodes (which are actually edges).
- second, we can notice 4 important values of minimum value for the size of cliques: 10, 11, 15 and 17. The four of them appear in OPNs in 2006 when $k = 1$, and also in 2012 (except 15) in OPNs when $k = 1$. It is surprising to have cliques composed with 17 nodes which consists in the ego and its *direct* alters from the beginning as the authors start publishing and thus building their own personal networks. For these particular cases, the authors in question integrates the entire collaboration networks by doing a publication which brings together a large number of co-authors (10, 11, 15 or 17). The fact that such clique (among cliques of size 10, 11, and 17) remains the one holding the minimum

size even after some time as years pass and so as the network evolves means that no other new node joined the personal network and connects to the ego as we are considering the personal networks at $k = 1$. Which is not the case of the clique of size 15 which is not anymore the one of minimum size of any personal network on 2012 because new connections to the ego created new cliques (at least one) of smaller size.

- last, we can also outline the fact that the OPNs in both years have a comparable behaviours when coming to the minimum size of their cliques.

4.5.6 Maximum cliques size

As for the minimum size of cliques, we do the same for the maximum cliques size by plotting the number of OPNs vs the maximum size of cliques; computed for all the OPNs, as the maximum among the size of the cliques composing an OPN. From the graphic given in Figure 4.55, we outline the following observations:

- the set of around 170 OPNs in 2006 with $k = 1$ have a maximum size of their cliques of 3. This means that these OPNs should be composed only by cliques of size 3 (one or more) or have in addition some edges (cliques of size 2) but no clique of size bigger than 3.

- a high number of OPNs contains cliques with a maximum size of 18 or 20 (around 160 and respectively 190 OPNs). As all the authors we are evaluating belong to the giant component these large cliques are probably the same appearing in many personal networks.

- the previous situation is more surprising when the year is 2006 (given by the blue dots) as we can observe it for $k = 2$, 3, 4, and 5 where cliques of size 18 for example are there as the personal network is created. We explain the presence of such big cliques at the first phases of the personal network construction by the collaborations gathering a large set of authors for a unique paper so the formation of these cliques is not a consequence of merging of some smaller cliques but join to the network directly bigger as it is.

- however, large cliques of size 20 were not present in any personal network on 2006. Thus, we can conclude that these cliques of size 20 are a consequence of the merging of preexisting cliques.

- finally, we can notice that large cliques are more frequent when k equals 4 and 5 (more than 150 OPNs have cliques of size 18 when k=5).

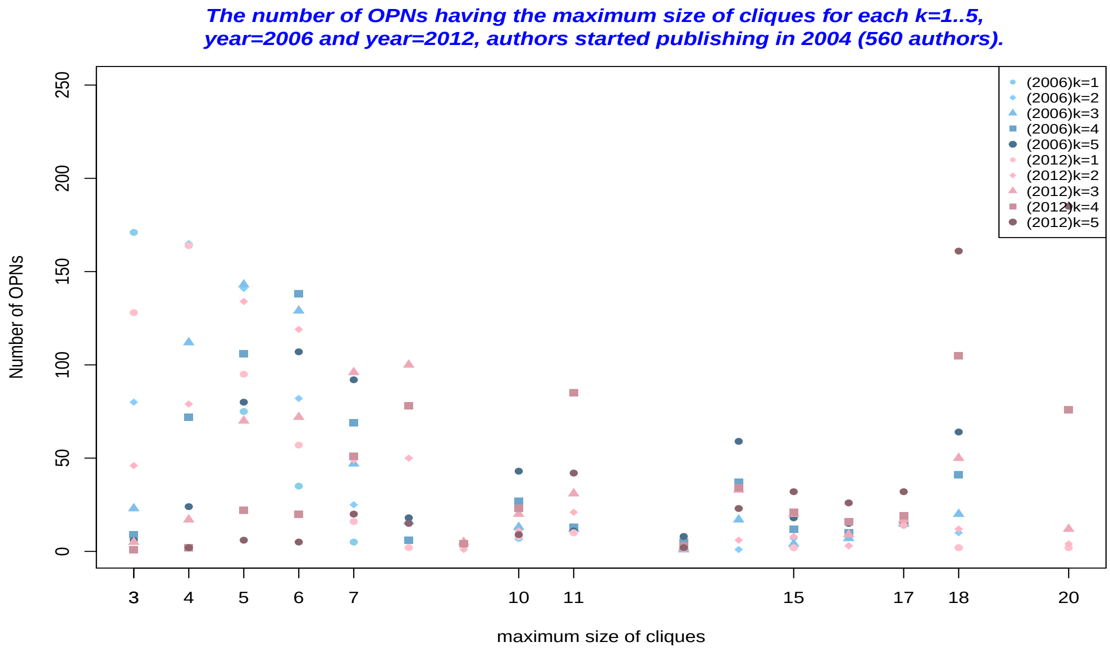


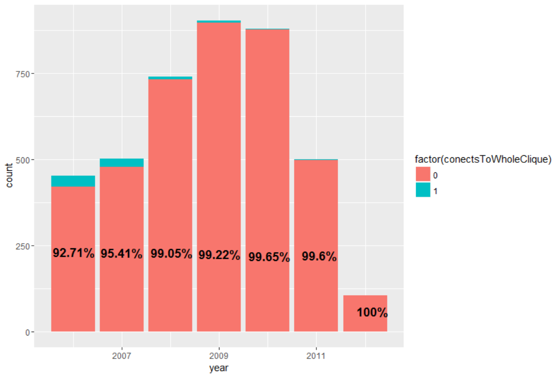
Figure 4.55: Number of OPNs vs maximum size of cliques in 2006 and 2012, all k.

We particularly find very interesting the last point. Indeed, the fact that OPNs with k=4 and 5 hold an important set of large cliques confirms the hypothesis we did previously (in the section of metrics evolution analysis) to explain the changing behaviour of OPNs at k=4 and 5 regarding two metrics: the betweenness centrality and the power law distribution of the degrees. We recall that for the betweenness centrality we found that it increases when k=2 and 3 but decreases when k=4 and 5, and for the power law distribution of OPNs' nodes degree is considerably less verified when k=4 and 5 than when k=1, 2 or 3. So, the changing evolution tendency of these two metrics is effectively explained by the creation of large cliques.

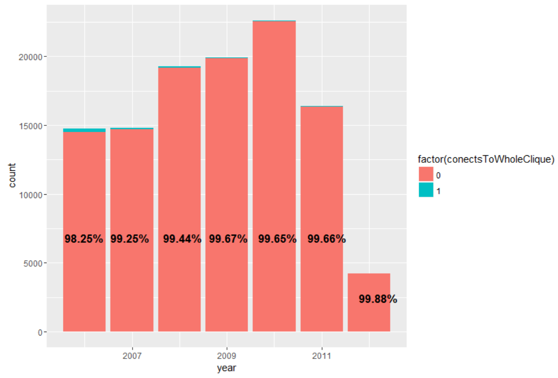
4.5.7 Connection between new nodes and (maximal) cliques

In order to understand how new nodes connect to the existing cliques of the personal network, we performed a set of experiments that we describe in the following.

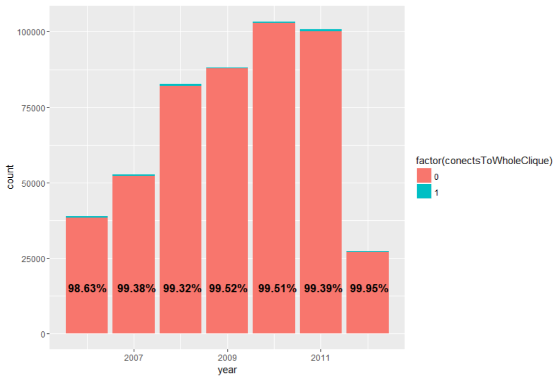
First of all, we tested whether new nodes connect to all the nodes composing an existing clique (thus, we say that the node connects to the whole clique) or to only few of them (thus, we say that the node connects to a sub-clique). The results of this test are given in Figure 4.56. A graphic in the Figure 4.56 presents, over the total number of new nodes on the discussed OPNs, the percentage of new nodes connecting to whole cliques (blue) and the percentage of new nodes connecting to sub-cliques. This graphics allows us to conclude that new nodes tend to connect to sub-cliques rather than to whole cliques.



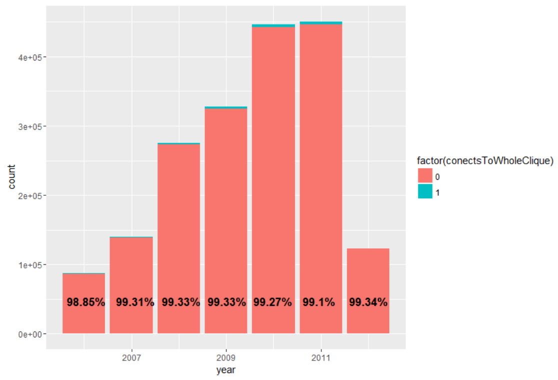
(a) k=1



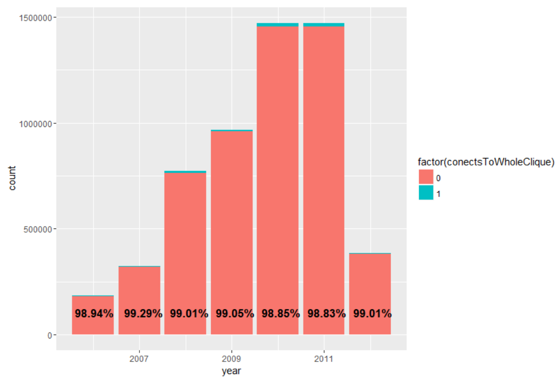
(b) k=2



(c) k=3



(d) k=4



(e) k=5

Figure 4.56: New nodes connection to a whole clique (blue) or to a sub-clique (red), for $k = 1...5$, and years 2006 to 2012.

Second, we computed the percentage of the nodes in the existing cliques to which new nodes connect to. We found that new nodes connect more often to around 30%-50% of the clique nodes, and rarely above 50% of them. Figure 4.57 gives the corresponding graphic for $k = 4$, and $year = 2009$. A similar tendency is observed for the other k values and years.

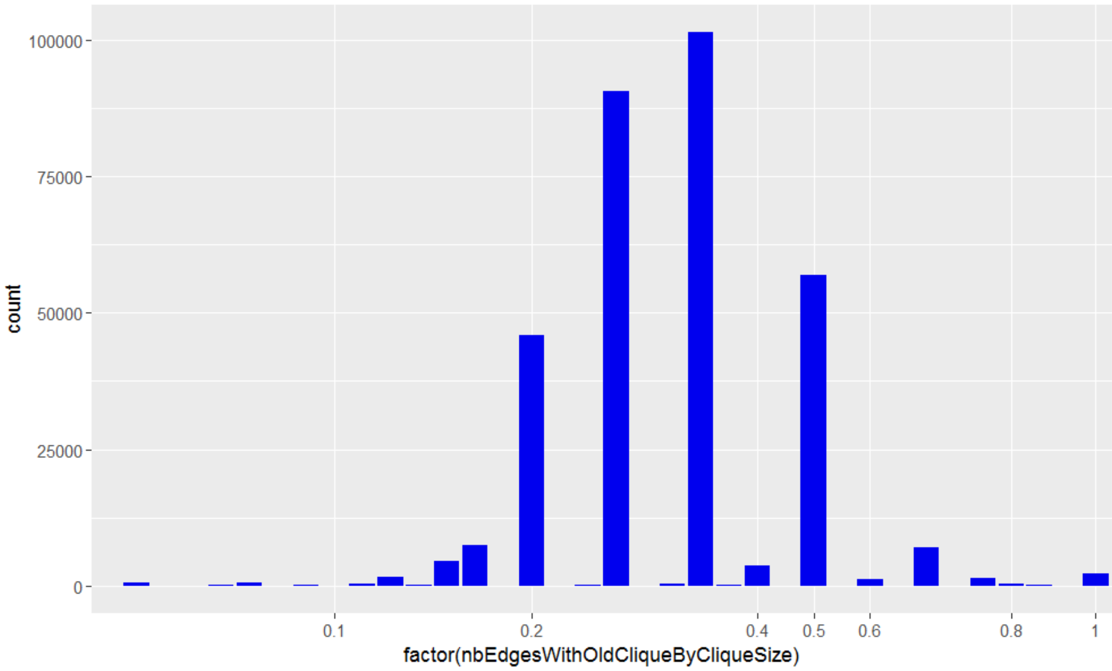


Figure 4.57: Percentage of clique nodes to whom new nodes connect to, for $k = 4$, year$=$ 2009.

Last, we computed the number of cliques to which new nodes connect to. We found that new nodes are more tended to connect to few cliques (1, 2, or 3) that to a large number of cliques, as we can observe on Figure 4.58 for $k = 3$, on 2010 (the same tendency was observed for the other k values and the other years).

4.5.8 Results summary and discussion

From the cliques analysis, we could confirm the findings from the metric analysis of the co-authorship personal networks and discover new interesting patterns.

First, we confirmed the fact that OPNs with $k = 1$ are particular and have a different structure since they are characterized by a few number of cliques composing them and we found that the size of cliques in 1-personal networks does not exceed 3. In addition, a large number of them are composed of only one single clique which proves that these networks are complete and which explains the fact that from the metrics analysis we had a considerable set of personal networks for which the value

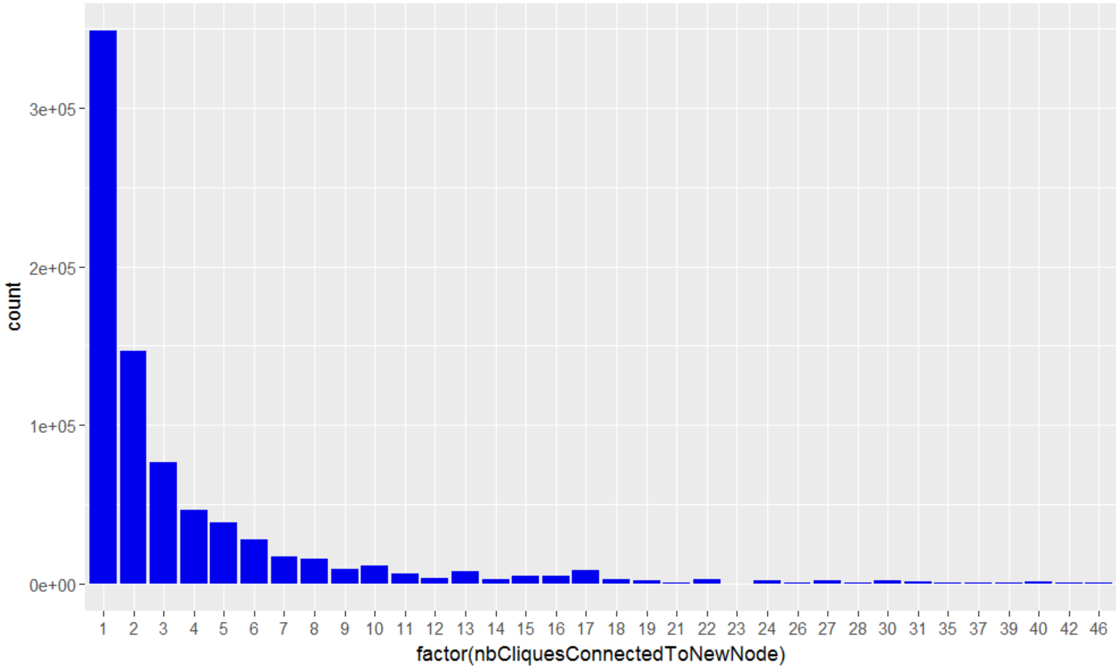


Figure 4.58: Number of cliques to whom new nodes connect to, $k = 3$, $year = 2010$.

of certain metrics was extreme as it was the case for the local clustering coefficient - many networks had $Ce = 1$, and for the betweenness centrality - many networks with $Be = 1$. Thus, we proved via the analysis of the cliques that OPNs made of one clique are frequent when $k = 1$.

Second, for higher k values ($k = 2, 3, 4, 5$), we discovered a linear relationship between the number of cliques and the number of nodes as well as with the number of edges. We also found that, in general, co-authorship OPNs have more small cliques than big ones, and that the number of cliques of a given size is increasing with k but the increase is not linear. Moreover, with the year, the size of cliques increases and the number of OPNs containing the cliques increases. In addition, we saw that for a high number of OPNs the minimum size of its cliques is 3 (more than 500); this is valid whatever k, while a high number of OPNs contains cliques with a maximum size of 18 or 20 which constitute the same cliques appearing in different OPNS as they belong to the giant component.

An important finding concerns the discovering of a different behaviour between OPNs with $k = 2, 3$ and OPNs with $k = 4, 5$ as it was the case with metrics analysis concerning the betweenness centrality (increasing when $k = 2, 3$, and decreasing for $k = 4, 5$) and the degree distribution of nodes composing the OPNs (the power law holds for $k = 2, 3$ but not holds for $k = 4, 5$). Indeed, we found that clique's size is following a power law distribution when $k = 2, 3$ but a Poisson distribution when $k = 4, 5$. Moreover, we confirmed that OPNs with $k = 4, 5$ are often characterized by the presence of large cliques. So, the changing evolution tendency of the betweenness centrality and the power law distribution of the degrees is effectively explained by the creation of large cliques. Thus the discovered patterns when considering cliques is a plausible explanation of the changing behaviour of the betweenness and the power law distribution of the degrees we obtained via metrics analysis.

Finally, by analyzing the connections between new nodes and cliques, we could identify that:

- new nodes tend to connect to sub-cliques rather than to whole cliques.
- new nodes connect more often to around 30%-50% of the clique nodes, and rarely above 50% of them.
- new nodes are more tended to connect to few cliques (1, 2, or 3) than to a large number of cliques.

We will see in the next chapter that such patterns are useful when proposing models for OPNs evolution based on cliques.

4.6 Conclusion

Along this chapter, the first obvious observation that we notice is that 1-personal networks constitute a particular case and have a different structure and thus evolution trends for the following reasons:

1. A significant proportion of 1-level personal networks (around 11%) are small star networks with few nodes. These networks describe independent ego's publications with distinct alters (co-authors). Such personal networks are characterized with an ego betweenness centrality of $B_e = 1$ along with a weak ego degree. The ego in this case plays a key role given its important position.
2. A more significant set of 1-personal networks is consisting of complete subgraphs (around 46%), which translates in one single publication grouping the ego and his/her alters (co-authors). These networks are characterized by an ego betweenness of $B_e = 0$ and a local clustering coefficient $Ce = 1$. In addition, we proved via the analysis of the cliques that these complete networks corresponds indeed to OPNs made of one single clique.

3. The evolution of 1-personal networks affects mostly the connections among alters who were characterized at the beginning with an important number of connections and then become less connected over time as new alters join the 1-level personal network (the density, the local clustering and the efficiency decrease), while egos' degrees and egos' betweenness are not affected. Moreover, the addition of nodes (alters) keeps the average degree of the 1-level personal network low.

4. Finally, the analysis of new links among alters in 1-level co-authorship personal networks shows that most of the new links are formed among new coming nodes which reveals that nodes join the network under the form of cliques. Thus, a new collaboration involves most of the time new authors that were not in the network at time t with inevitably the ego author (because we consider 1-level personal networks) and eventually his alters (old nodes) which explains the second important proportion of edges made between new and old nodes. However, we found that it is unlikely for two old authors not connected at time t to connect at time $t+1$ given the very small proportion of links made between old alters.

The metrics analysis (published in [40]), and cliques analysis of personal networks with $k = 2, 3, 4, 5$, allowed us to discover interesting findings concerning the topology and the evolution of co-authorship personal networks. These finding are summarized in the following:

1. The density and global clustering coefficient are both decreasing over the years. The nodes joining the personal network over the years create fewer edges compared to the possible number of edges that can be created, and thus the density gets lower, as well as the transitivity at the personal network level.

2. At the local level, the average clustering coefficient was observed to be high and gets higher when the personal networks are growing which indicates a high local clustering around the nodes. We can explain that by the fact that co-authorships are made generally between 3, 4 or 5 authors and the emergence of a new collaboration implies the creation of a highly clustered local structure since all the authors that made this collaboration are going to be already connected with each other. We also found that the average degree of nodes are increasing over years.

3. We saw that for a high number of OPNs the minimum size of its cliques is 3, while a high number of OPNs contains cliques with a maximum size of 18 or 20 which constitute the same cliques appearing in different OPNS as they belong to the giant component. Furthermore, we discovered a linear relationship between the number of cliques and the number of nodes as well as with the number of edges. We also found that, in general, co-authorship OPNs have more small cliques than big ones, and that the number of cliques of a given size is increasing with k and with the years.

4. Ego's betweenness centrality behaves differently depending on k since it was increasing for $k = 2, 3$, but decreases for $k = 4, 5$. On the one side, the increase for $k = 2, 3$ is due to the addition of nodes around the ego that form disconnected groups of alters (as the windmill graph example) and the ego plays a key role among them. On the other side, when we consider a $k = 4$ or 5. This behaviour is justified by the fact that OPNs with $k = 4, 5$ are often characterized by the presence of large cliques as discovered with cliques analysis.

5. As for the betweenness, the degree power law distribution test results are different depending on k as it is better satisfied when $k = 2, 3$ than for $k = 4, 5$. In addition, for $k = 5$, the networks are not anymore following a power law degree distribution starting from 2011. We explained this changing behaviour when we analyzed the composition of the OPNs when considering cliques, as we found that OPNs with $k = 4, 5$ are often characterized by the presence of large cliques.

6. Another different pattern between $k = 2, 3$ and $k = 4, 5$ was discovered about the clique's size, which is following a power law distribution when $k = 2, 3$ but a Poisson distribution when $k = 4, 5$.

7. Finally, by analyzing the connections between new nodes and cliques, we could identify that new nodes tend to connect to sub-cliques rather than to whole cliques and they connect more often to around 30%-50% of the clique nodes, and rarely above 50% of them. At last, new nodes are more tended to connect to few cliques (up to 3) than to a large number of cliques.

All the findings mentioned above contribute to bringing a better understanding around the structure and evolution of personal networks. Such a deep investigation has not been carried out before. In addition, the discovered patterns constitute a sufficient baggage to approach modeling the evolution of the studied networks. In particular, we distinguish that the organization of nodes and edges in the personal networks in the form of cliques is a dominant characteristic during the evolution of these networks given the nature of the data. Thereby, our target when thinking about an evolution model for personal networks has to take mainly into account the clique dimension in addition to the other discovered properties. Thus, we propose in the next chapter a new evolution model for personal networks based on cliques.

Chapter 5

PERSON*EM* - a New Evolution Model

5.1 Introduction

In the previous chapter, we concluded that cliques dominate co-authorship personal networks' structure and play a key role in their evolution. This lead us to investigate whether previous works on networks evolution have considered this dimension when proposing models. A limited number of works are based on cliques, and the reason might come from the fact that the computational problem of finding cliques in a graph is NP-complete and might constitute a real constraint when considering large social network graphs. In our case, we are dealing with personal networks that are less large. Thus, working towards an evolution model based on cliques is less problematic. In that perspective, the work done by Yan et al. in [134] is the only one proposing a model based on cliques. In addition, this work was interesting for us because they consider the co-authorship network. However, the proposed model is a generative model designed for generating artificially social networks graphs, while in our case, the objective is different since we would like to predict the evolution of a personal network, in particular, from a given point in time t to a next point in time $t+1$.

Indeed, the model we would like to build can be summarized as follows: at time t , an OPN has n_t nodes and m_t edges; at times $t+1$, $\delta n = n_{t+1} - n_t$ new nodes are entering the personal network and the evolution model should predict how these nodes will connect to the network and what other links are created in the network. In addition, if at time t, the maximum distance from the ego is equal to k, then at time $t+1$, it may be equal to $k+i$, as new nodes connect to existing nodes that are at k distance from the ego. Thus, the model should account of that in order to maintain the same k as at time t.

In this chapter, we will present in the first place the clique-superposition model proposed by Yan et al. in [134], and discuss in details why this model cannot work for predicting the evolution of personal networks. Then, based on this model, we will show how starting from the generative model of Yan et al., we build a new predictive model we named PERSON*EM* for the evolution of personal networks that we presented in [41]. Later in this chapter, we present the results of application of the new clique-based predictive model dedicated to OPNs, and finally discuss its efficiency and its performance.

5.2 Clique-superposition evolution model [134]

The clique-superposition model, proposed in [134], is a generative model for undirected weighted networks that reproduces real social networks properties. This model proposes to construct a network with N nodes by using the clique-superposition concept that we describe in the following section.

5.2.1 Clique-superposition concept

In 1981, Feld [51] studied the organization of ties within a social network. He pointed out that a social link is more likely to be formed between two individuals having some intersection. This fact comes from the theory claiming that the relevant aspects of the social environment are like foci around which individuals organize their relationships where a focus is defined as a social, psychological, legal, or physical entity around which joint activities are organized [51] (eg. workplaces, families, etc.). Consequently, a group of individuals whose activities are organized around the same focus will tend to become tied to each other and form a **cluster** hence the concept of *focused organization* titling the paper of Feld in [51].

In addition, Feld distinguished between the *simply focused* situation where each individual is related to one single focus and thus, each focus is represented by a cluster and links among clusters are just *random*, and the situation that is more complex but most common where an individual acquires over time multiple focus at the same time. In this last situation, clusters form a complex interpenetration or *superposition* structure. This is from where came the idea of providing a generative model that simulates the clique-superposition evolution behaviors of social networks.

To illustrate the superposition of cliques, we give the following example. Let us assume that we have a clique and a new node comes to join that clique, then the new node links to all clique members and altogether form a bigger clique. If we suppose that the new node belongs to other cliques through its former activity, thus the cliques are superposed at that new node. For example, in Figure 5.1, cliques C_1 and C_2 are superposed in node 3. In the next section, we present the clique-superposition model and discuss in details how this model works.

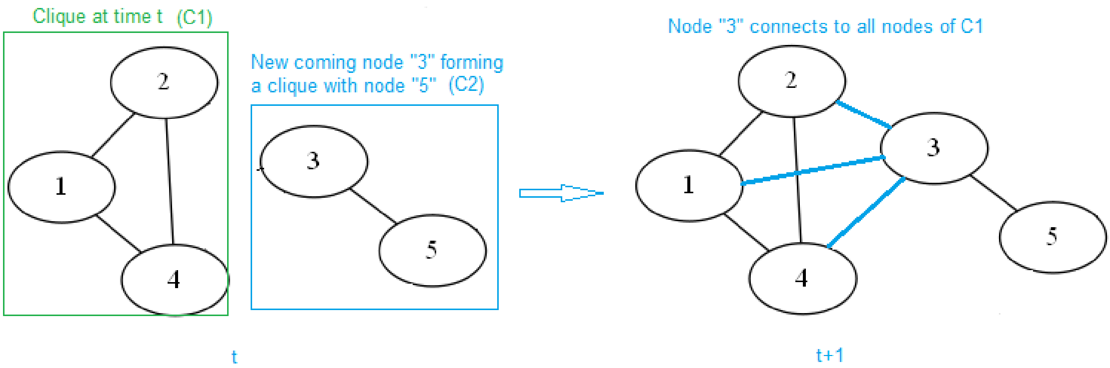


Figure 5.1: Clique-superposition example.

5.2.2 Model's description

As stated previously, the clique-superposition model is a generative model and is based on three assertions. The first assertion assumes that social links are driven by social intersections, an example to illustrate this assertion is the friendship link where a new friendship is formed because two individuals share one common friend or share a common social environment as the workplace; this assertion was translated by the social intersection. The second assertion concerns the fact that social networks evolve in the form of superposition of cliques. This assertion is related to the first one since from a friendship for instance between two individuals, we can move to a triangular relationship that includes the common friend. Another example, presented in [134] as the extreme example for illustrating the clique-superposition, is the co-authoring networks where authors of the same paper are present in the co-authoring network as a clique and thus the co-authoring network is the result of superposed clique of different sizes. Finally, the third assertion is the fact that the co-relationship network is a kind of social network as it was demonstrated in the literature for the scientific collaboration network [99] or the film co-acting network [68] that have the same structural properties (high clustering coefficient, power law degree distribution, and small diameter) as social networks in their usual sens of acquaintance networks.

The clique-superposition model proposed in [134] is given in Algorithm 1, and in the following we will describe the steps of the algorithm.

Algorithm 1: Clique-superposition model

```
   Input  : G_init, k, β, p_node, p_sample, N
   Output: G
1  begin
2  Graph G ←newGraph();
3  while G.getNodes() < N do
4      if Random() < p_node then
5          Node current ← newNode();
6          G.addNode(current);
7      else
8          Node newMember ← G.getRandomNode();
9          k ← GetCliqueSize();
10         Clique C ← FindClique(G,k);
11         if C ≠ null then
12             if Random < p_sample then
13                 Clique C′ ← Sample(C);
14             else
15                 Clique C′ ← C;
16             end
17             for i = 1, . . . , Size(C′) do
18                 G.addEdge(newMember, C′.node[i]);
19                 // if the edge already existing, the edge weight plus 1
20             end
21         end
22     end
23 end
24 return G;
25 end
```

The algorithm gets as an input the personal network at time t (G^{init}), the value k for this network, the parameter of the power law distribution of the cliques' size β, the p_{node} value corresponding to the probability of having a nodes or edges evolution phase, p_{sample} value corresponding to the probability that a new node connects to a cliques or a sub-clique, the number of nodes of the final graph G (N). The output of the algorithm is the graph G, corresponding to the predicted personal network at time $t + 1$ of the initial network.

The initial graph is instantiated via the method *newGraph()*. Then, while the initial graph do not reach N nodes, the following steps are executed (line 3).

A random number is generated and compared to p_{node}. If the random number is less than p_{node}, the **node evolution step** starts (line 4). In a node evolution steps, a new node is created and added to the graph G (and the generated graph size is incremented) (lines 5, 6). If the random number is grater than p_{node}, the **edge**

evolution step starts (line 7).

The evolution step runs as following. A node, from the existing nodes, is randomly selected as *newMember* (line 8) in order to be connected to a clique or a sub-clique. The value q, corresponding to the size of the clique, is extracted from the power law distribution of the clique sizes (line 9). Then, using KEWLS algorithm, a random clique of size q is selected (line 10); KEWLS which is a vertex cover finding algorithm that will be detailed in the next section. KEWLS algorithm has as input three parameters: the complementary graph $\overline{G}$ of the initial graph G, the size of the vertex cover $(|V| - q)$, and the number of maximum steps to be executed in order to find a vertex cover.

In case KEWLS did not succeed in finding a vertex cover (and so a random clique), the edge evolution step ends (lines 12, 13) and the algorithm goes back to line 3. Otherwise, the vertex cover and the q-clique are returned (lines 14, 15). To know if the *newMember* node will connect to the complete clique or a sub-clique of the returned clique, a random number is generated. If the random number is less than p_{sample}, a sub-cliques is sampled; otherwise, the full q-clique is taken as sample clique. The sampling method works as follows: *q-sample* are picked randomly in the range $[1, q-1]$, a smaller clique of size *q-sample* is constructed by selecting *q-sample* nodes from the original clique nodes set and add edges between all the selected nodes to for a new smaller clique.

Finally, edges between sampled clique nodes and *newMember* are added, if no edge already exists among them (line 18), otherwise, the weight of the existing edge is incremented (line 19).

5.2.3 Vertex cover algorithm for random clique finding - KEWLS algorithm

KEWLS is an heuristic proposed [134] and based on the EWLS algorithm proposed in [24] for solving the vertex cover problem. The vertex cover problem is related to the random clique finding problem since finding a random clique of size k in a graph $G(V, E)$ consists in finding a vertex cover of size $|V| - k$ in the complementary graph $\overline{G}$. We describe in the following the different steps of KEWLS, given in 2.

The first phase of KEWLS algorithm is the initialization of the following entities:

- the number of steps is set to 1 (line 2);
- the list of uncovered edges l is initialized with the whole set of edges (line 2);

- all edges weight are set at 1 (line 3);
- all edges age is set at 0;
- the k vertex cover is initialized with a random vertex set of size k (line 4).

Then, a loop is executed until $maxsteps$, which indicates the maximum number of steps for executing the loop, is reached (line 5) with the following steps: constructing the set l of uncovered edges (not covered by the vertex cover). If this set is empty, it means that the initialization of the vertex cover is actually a vertex cover. If not, a list containing the uncovered edges endpoints is defined. Afterwords, the function $chooseExPair$ is called, which finds a pair of nodes (u, v) such that u belongs to the vertex cover and v belongs to the list of uncovered edges endpoints (line 6). Finally, two cases are possibles:

1. If $(u, v)! = (0, 0)$, the vertex u is removed from vertex cover and the vertex v is added (line 7).The edge age of edges of vertex v (that were uncovered) are incremented.
2. If $(u, v) = (0, 0)$, a local optima was found, thus the weight of all edges is incremented (line 11) and a random walk is processed (line 12) which exchanges a random vertex u in C and a random vertex v in $l_{endpoints}$ list.

Algorithm 2: KEWLS(G, k, maxSteps) [134]

Input : G, k, maxSteps
Output: k-vertex cover of G

begin
  step $\leftarrow 0$; $L \leftarrow E$;
  initialize edge weights;
  construct C randomly until$\mid C \mid= k$;
  while *step* $<$ *maxSteps and* $L \neq \emptyset$ **do**
    if *((u,v)* $\leftarrow$ *ChooseExPair(C, L))* $\neq$ *(0,0)* **then**
      $C \leftarrow [C \setminus \{u\}] \cup \{v\}$;
      tabuAdd $\leftarrow u$;
      tabuRemove $\leftarrow v$;
    else
      update edge weights;
      take a random walk;
    end
    step $\leftarrow$ *step* $+ 1$;
  end
  return C;
end

The function *ChooseExchangePair***.** The function *ChooseExchangePair* finds a pair of nodes (u, v) such that u belongs to the vertex cover and v belongs to the list of uncovered edges endpoints, and $score(u, v) > 0$. The $score(u, v)$ is computed as follows:

Given a graph $G = (V, E)$ and a current candidate solution C, for a pair of vertices $u, v \in V$, where $u \in C$ and $v \notin C$,

$score(u, v) = dscore(u) + dscore(v) + w(e(u, v))$ if $e(u, v) \in E$, and

$score(u, v) = dscore(u) + dscore(v)$ otherwise.

Such as: $dscore(v) = cost(G, C) - cost(G, C')$,

where $cost(G, C) = \sum_{e \in E \text{ and e is not covered by C}} w(e)$, $w(e)$ determines the weight of edge e.

In this function, presented in Algorithm 3, the vertex v to put in C is chosen from endpoints of the oldest uncovered edge $e^*(v_1^*, v_2^*) \in L$. If there exists at least one vertex pair $u \in C$ and $v \in (v_1^*, v_2^*)$ such that $score(u, v) > 0$, the function returns one of them randomly (line 3). If not, it goes to check the edges in UL, according to the order from old to young. If for some edge $e(v_1, v_2)$ in UL, the set $S := \{(u, v) | u \in C, v \in \{v_1, v_2\}$ and $score(u, v) > 0$ is not empty, then the function returns one vertex pair $(u, v) \in S$ randomly (line 10). Finally, if the function fails to find such a vertex pair, it returns (0,0) (line 14).

Algorithm 3: function *ChooseExchangePair(C, L, UL)*

Input : current candidate solution C, uncovered edge set L, edge set UL of uncovered edges unchecked in the current local search stage

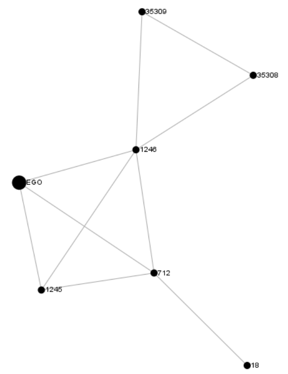
Output: a pair of vertices

begin
if $S := \{(u, v)|u \in C, v \in \{v_1^*, v_2^*\}$ *and* $score(u, v) > 0\} \neq \emptyset$ **then**
    return random(S);
else
    foreach $e(v_1, v_2) \in UL, from\ \ old\ \ to\ \ young$ **do**
        if $S := \{(u, v)|u \in C, v \in \{v_1, v_2\}$ *and* $score(u, v) > 0\} \neq \emptyset$ **then**
            return random(S);
        end
    end
end
Return $(0, 0)$
end

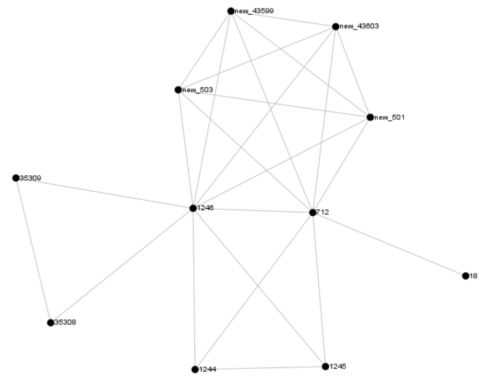
5.2.4 The clique-superposition model applied on co-authorship personal networks

Let us consider the OPN of ego 1244 with $k = 2$ evolving from 2008 to 2009. The number of new nodes added from 2008 to 2009 is 4. Figure 5.2a gives the OPN of ego 1244 in 2008 with $k = 2$, and Figure 5.2b gives the OPN of ego 1244 in 2009 with $k = 2$. The latter should be predicted by the clique-superposition model.

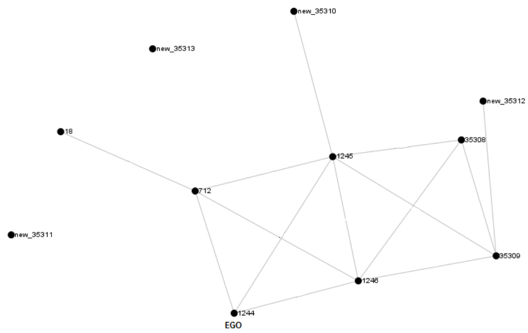
Thus, we applied on the OPN of ego 1244 given in Figure 5.2a the clique superposition model proposed in [134] using the following parameters $pnode = 0.1$, $psample = 0.6$, $beta = 2.5$ (used in the original work). The result is given in Figure 5.2c.



(a) Real OPN on 2008.



(b) Real OPN on 2009.



(c) Predicted OPN on 2009 with clique-superposition model.

Figure 5.2: Example of application of the clique superposition model on the OPN of ego 1244, for $k = 2$ and $year = 2008$.

5.3 PERSON*EM*: a new evolution model for personal co-authoring networks evolution

Let us consider a personal network PN_t at time t. Our objective is to predict PN_{t+1} at time $t+1$ after the addition of a set of new nodes; the number of the new nodes is given. Thus, to model should predict the new edges in the network, and, indirectly, how many edges are added. This way, we are not generating a network but we start from a personal network and a given amount of nodes to add and we have as output the personal network at the next time step.

The clique-superposition model presented in section 5.3 is a generative model of social networks based on the organization of the connections among users in terms of cliques and sub-cliques. In our work, we are interested in modeling the evolution of personal co-authoring networks. Inspired by the work in [134], we built a predictive model taking into account the specificity and the properties of personal and co-authoring networks.

To this end, firstly we tested the clique-superposition model on our data; the results were given above. Hereafter we will provide a set of observations made on the results of the tests performed; these observations concern the limitations of the superposition cliques model when applied on personal co-authoring networks.

5.3.1 Limitations of the clique-superposition model

5.3.1.1 Isolated nodes

The clique-superposition model tolerates having isolated nodes, i.e. nodes that are not connected to any other node in the network (with a degree equal to 0) as we can see it from the example of application of the clique superposition model given in Figure 5.2c. A simple illustration of this fact is to imagine that we have as a last step in the model a node evolution step which means that we add a node but this added node will stand disconnected because there is no edge evolution step afterwords that can eventually connect that node.

In the case of personal networks, isolated nodes are not accepted because by definition all nodes within a personal network are connected to the ego via a path of a given distance depending on k. So, to be applied on personal networks, the clique-superposition model should be adapted to avoid having isolated nodes.

5.3.1.2 The k level of the personal network

The second problem outlined when running the clique-superposition model on personal co-authoring networks is to take into account the parameter k, the maximum geodesic distance that separates the ego node from the rest of nodes in the personal network.

Indeed, the network predicted by the clique-superposition model at time $t + 1$ might have a k value different from the initial k value of the personal network at time t. But, the way we defined the predictive model we would like to have, that we presented in the introduction of this chapter, does not allow to overpass k.

5.3.1.3 KEWLS algorithm's performance

The vertex cover algorithm, KEWLS, used in the clique-superposition model for random clique finding problem has important limitations regarding performance in terms of running time and memory especially when dealing with large networks. Thus, we tested the algorithm KEWLS separately on a sample of the personal networks of authors in $dataset_1$. In order to select the sample, we performed a k-means clustering on the personal networks of $dataset_1$ for $k = 3$ on 2008 (set of 560 OPNs whose number of nodes and number of edges distributions are given in Figure 5.3, and Figure 5.4 respectively) after discarding personal networks that were not evolving from 2007 to 2008 (set of 466 OPNs). We chose the data for $k = 3$ on 2008 because it represents the central point in terms of k and years providing a representative set of networks in terms of size.

In the k-means algorithm, we used the following metrics to characterize an OPN: number of nodes and edges, density, global clustering coefficient, average clustering coefficient, betweenness centrality, and the average degree. We obtained 5 clusters and we took randomly 50% of the OPNS from each cluster and we got at the end 230 representative personal networks. Then we applied KEWLS algorithm to search for cliques of size 5 to 6, and used different values for $maxSteps$ parameter (7, 20, 100, 200).

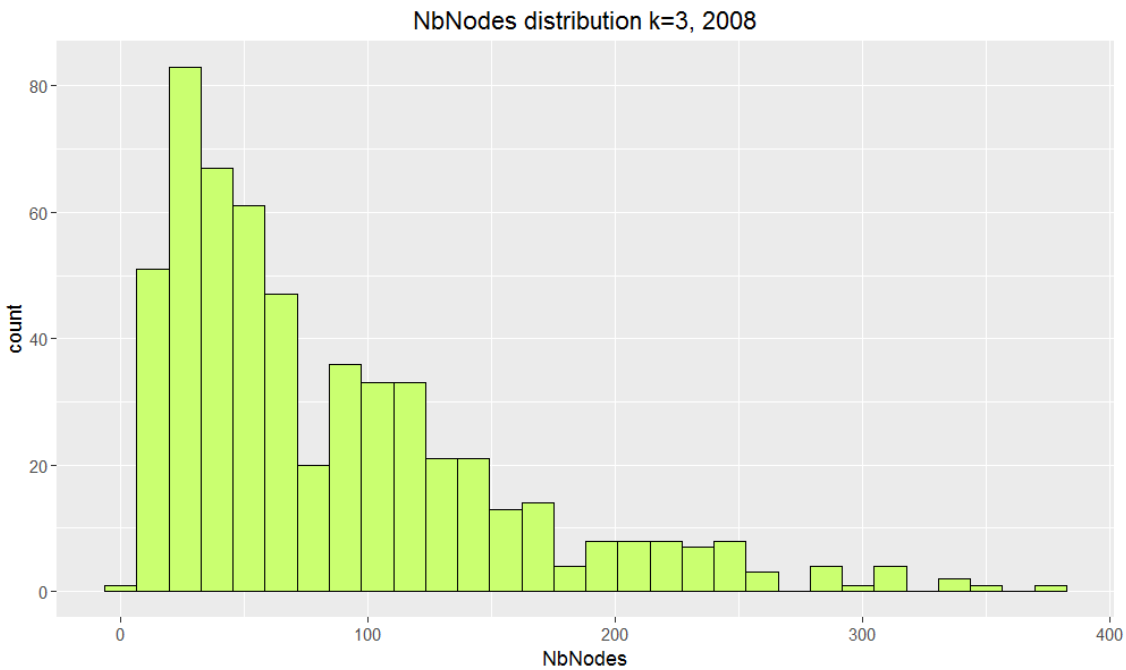


Figure 5.3: Distribution of the number of nodes.

It is important to outline that the personal networks under the scope are not huge and contain a maximum of 349 nodes and 798 edges. From the graphic in Figure 5.5 representing KEWLS performed for searching cliques of size 5 with $maxSteps$ of 200, we can see that the running time increases exponentially with the number of nodes.

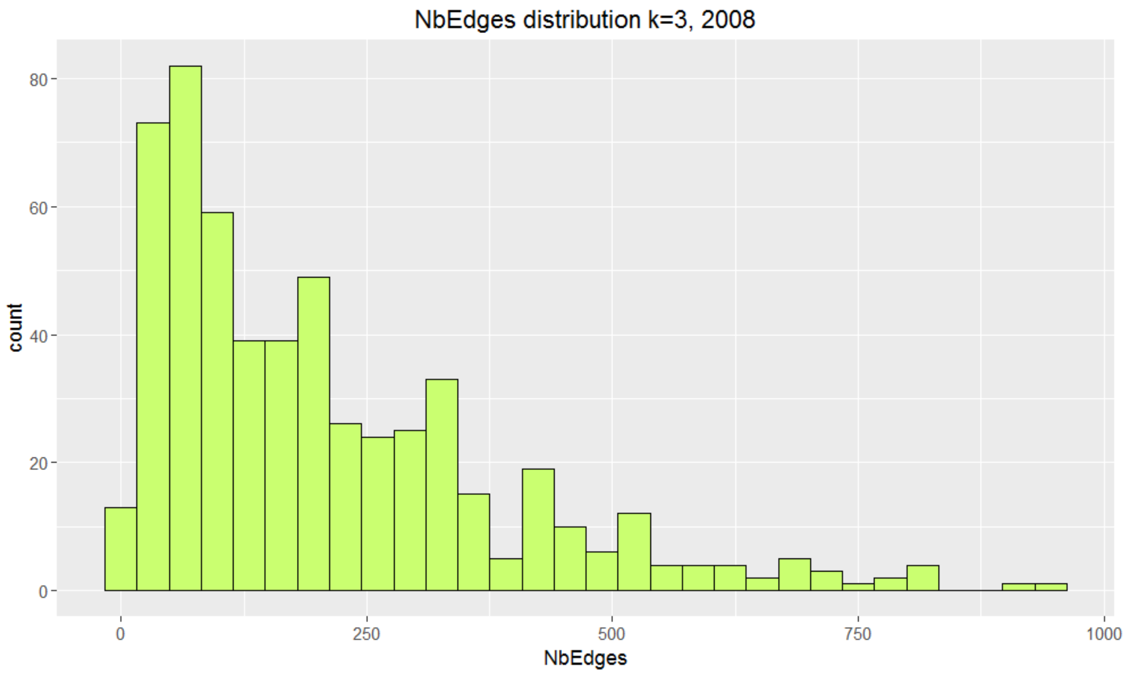


Figure 5.4: Distribution of the number of edges.

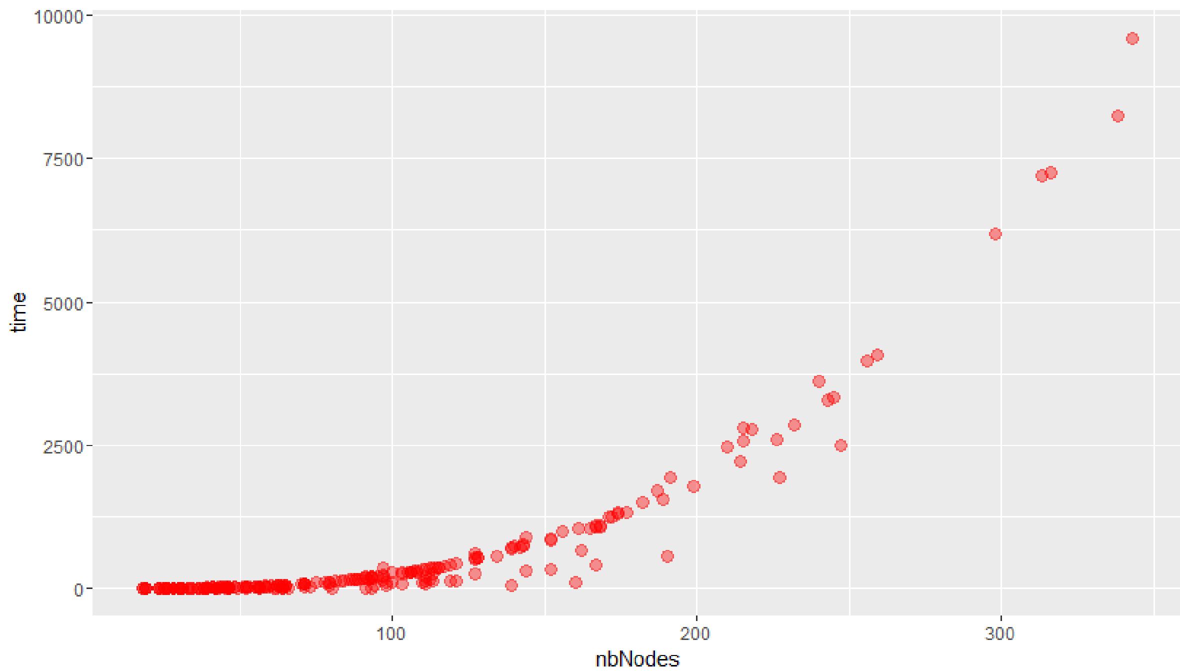


Figure 5.5: KEWLS algorithm running time in order to find a 5-clique with $maxSteps = 200$, when $k = 3$ and $year = 2008$, for $dataset_1$.

While the time is acceptable for small networks (with less than 100 nodes), the fact that performing KEWLS on a network of about 300 nodes takes about two hours is problematic since the networks on which we work could be much bigger in addition to the fact that if we run the clique-superposition model, we might involve the call of KEWLS algorithm many times until we reach the desired network of size. Thus, we need to replace KEWLS algorithm with a more performing alternative for finding a vertex cover.

5.3.1.4 Model's parameters values

It is important to note that the authors propose in [134] to use in the clique-superposition model for the three parameters with possible open values the following values:

- $p_{node} = 0.1$ which indicates that the model performs a node evolution step in 10% of cases and an edges evolution step in 90% of cases;
- $p_{sample} = 0.6$ which indicates that in 60% of cases the model connects the new member to a sub-clique rather than the entire clique;
- $\beta = 2.5$ which indicates the power law parameter of cliques size distribution.

This values should be revised in order to match the personal co-authorship networks.

5.3.2 Solutions to the outlined limitations: towards a predictive evolution model

To fit the personal co-authorship networks particularities, several adjustments have to be made on the clique-superposition model. Next, we discuss how we overcome the limitations described in the previous section.

5.3.2.1 Isolated nodes

Previously, we saw that with the clique-superposition model it was possible to have one or more isolated nodes. These isolated nodes constitute new nodes that remains disconnected.

In the case of personal co-authorship networks, we start with a connected graph (personal network at time t), and we should have as output a connected graph as well (at time $t+1$) without any isolated nodes. Thus, to avoid having isolated nodes after the classic run of the model (lines 3-23 in Algorithm 1), we propose to run the model two times.

The first run will correspond to the classic one presented in the Algorithm 1 in lines 3-23 that will launch both the node and edge evolution steps, thus allowing to add all the nodes to the graph in order to achieve the N number of nodes.

The second run will take place after the line 23, and will have as main goal to connect all the remaining isolated nodes; thus, the run launches only the edge evolution step without the node evolution step, since all new nodes are already added. Moreover, instead of selecting a random node from the whole nodes list, we select a random node only from the isolated nodes list.

The choice of proposing two runs of the algorithm instead of having one single run, comprising the two steps of evolution, and bring in a set of constraints was motivated by the fact that we wanted to modify as less as possible the original clique-superposition model.

5.3.2.2 The k level of the personal network

As already stated, in our work we formalized and analyzed the evolution of personal networks as following: we consider, in time t, a k-personal network with a number of nodes and edges, and this network evolves in time $t+1$ to a k-personal network generally with a different number of nodes and/or edges. It is important to note that both personal networks have the same depth, k. The clique-superposition model does not respect this constraint because, the networks predicted in time $t+1$ have a depth (k value) higher or equal than the network's depth in time t.

To account of this constraint, we propose to add the following constraint in line 10 in Algorithm 1: in the edge evolution step, we consider that a node was selected and has to be connected to a clique or a sub-clique. When a q-clique is selected in line 10, we select a q-clique only if it has at least a node at a distance less than q. If it is not the case, we continue searching for a solution . Then, on the next lines, if we have to select a sub-clique, we select a sub-clique having at least a node at a distance less than q.

This new constraints allow us to ensure that the new added nodes will not cause the over-passing of the depth (k value) of the personal network. We note that for personal networks with $k = 1$, for a clique in a 1-personal network the nodes are necessarily connected to the ego. Thus, we skip dealing with this new constraint when considering OPNs with $k = 1$.

5.3.2.3 KEWLS algorithm's performance

As stated in the previous section, the KEWLS algorithm is lacking of performance especially when applied on networks with more than 100 nodes. New heuristics have been proposed since and had been demonstrated as outperforming KEWLS algorithm (an overview is given in [131]). The NuMVC algorithm [25] was proposed as a more efficient algorithm than KEWLS for the minimum vertex cover problem. Next, we will present the NuMVC algorithm and provide a comparison between NuMVC and KEWLS algorithms performed on our data.

NuMVC description. NuMVC algorithm was introduced in [25] as a more efficient new local search heuristic algorithm for solving the Minimum Vertex Cover (MVC) problem. NuMVC proposes two new strategies to cover two drawbacks of existing in previous MVC solvers approaches as KEWLS.

The first drawback consists in the fact of selecting a pair of vertices to exchange simultaneously as done in KEWLS where a pair of vertices with $score(u, v) > 0$ is selected randomly. In [25], the authors outlined that this technique is time consuming since it leads to a quadratic neighborhood for candidate solutions. In addition, the evaluation of a pair of vertices depends on *dscore* between the two vertices, but also involves the relationship between the two vertices, like "do they belong to a same edge". Thus, evaluating all candidate pairs of vertices is rather time-consuming. To address this issue, the NuMVC algorithm uses a *two-stage exchange* strategy and selects the two vertices for exchanging separately in two stages, i.e. a removing stage and an adding stage. The removing stage consists in selecting a vertex from the current candidate solution and removing it, while the adding stage consists in selecting a vertex in a random uncovered edge and adding it to the current solution. This two stages exchange strategy yields an efficient two-pass move operator for MVC local search, in which the first pass is a linear-time search for the vertex-to-remove, while the second pass is a linear time search for the vertex-to-add.

The second drawback consists in the edge weighting technique; KEWLS increases the weights of uncovered edges only when reaching a local optima. However, KEWLS or other algorithms do not have a mechanism to decrease the weights which could be deficient because the weighting decisions made too long ago may mislead the search as stated in [25]. In the NuMVC algorithm, the authors propose to use an *edge weighting with forgetting* technique by increasing the weights of uncovered edges by one at each step. Then, when the averaged edge weight achieves a threshold (γ), it reduces weights of all edges by multiplying a constant factor ρ ($0 < \rho < 1$) to forget the earlier weighting decisions.

The NuMVC algorithm is given in Algorithm 4 and it is described in the following.

Algorithm 4: NuMVC(G, cutoff)

```
   Input  : G, cutoff time
   Output: vertex cover of G
 1 begin
 2 initialize edge weights and dscores of vertices;
 3 initialize the confChange array as an all-1 array;
 4 construct C greedily until it is a vertex cover;
 5 C* := C;
 6 while elapsedTime < cutoff do
 7     if there is no uncovered edge then
 8         C* := C;
 9         remove a vertex with the highest dscore from C;
10         continue;
11     end
12     choose a vertex u ∈ C with the highest dscore, breaking ties in favor of
         the oldest one;
13     C := C\ {u}, confChange(u) := 0 and confChange(z) := 1 for each
         z ∈ N(u);
14     choose an uncovered edge e randomly;
15     choose a vertex v ∈ e such that confChange(v) = 1 with higher dscore,
         breaking ties in favor of the older one;
16     C := C ∪ {v}, confChange(z) := 1 for each z ∈ N(v);
17     w(e) := w(e) + 1 for each uncovered edge e;
18     if w̄ ≥ γ then
19         w(e) := ⌊ρ · w(e)⌋ for each edge e;
20     end
21 end
22 return C*;
23 end
```

NuMVC starts with an initialization phase by setting all edges weights at 1 and computing *dscores* accordingly. Moreover, NuMVC maintains a boolean array $confChange(v)$ initialized at 1 for each vertex v. The $confChange$ array handles the configuration checking (CC) strategy which avoid cycling problem in local search . Then, NuMVC constructs the current candidate solution C by iteratively adding the vertex with the highest *dscore*, until it becomes a vertex cover and the best solution C^* is initialized as C.

After the initialization, a loop is executed until the cutoff time is reached (lines

6 to 19). The cutoff time is the time over which we stop searching for a solution (a vertex cover). During the search procedure, if there is no uncovered edge (C is a vertex cover), NuMVC updates the best solution C^* as C (line 8). Then it removes one vertex with the highest *dscore* from C (line 9), to go on to search for a vertex cover of size $|C| = |C^*| - 1$. Then, the two-exchange strategy is executed (lines 12 to 16) it first selects a vertex $u \in C$ with the highest *dscore* to remove, then chooses an uncovered edge e uniformly at random, and selects one of e's endpoints to add into C as follows: If there is only one endpoint whose $confChange$ is 1, then that vertex is selected; if the $confChange$ values of both endpoints are 1, then NuMVC selects the vertex with the higher *dscore*. After the two-exchange step $confChange$ array is updated as well as all uncovered edges weights which are increased by 1 (lines 16, 17).

Afterwords, the forgetting mechanism is performed (line 18, 19) for decreasing the weights. At this step, if the averaged weight of all edges reaches a threshold γ, then all edge weights are multiplied by a constant factor ρ $(0 < \rho < 1)$ and rounded to an integer (edge weights are defined as integers in NuMVC).

q-NuMVC algorithm: an adaptation of the NuMVC algorithm for the k-vertex cover problem. The NuMVC algorithm presented above returns the minimum vertex cover a graph G. In the context of the clique-superposition model, we need to find a clique of size k. This task is undertaken by a MVC algorithm (currently the KEWLS algorithm), which allows to extract a vertex cover of size $|V| - k$ in the complimentary graph $\overline{G}$. Thus, in order to replace the KEWLS algorithm with the NuMVC one, we needed to adapt the NuMVC algorithm so that it returns a minimum vertex cover of a given size.

In the Algorithm 5, we give the pseudo-code of the new corresponding version of NuMVC that we call q-NuMVC we are proposing, where k represents the size of the minimum vertex cover we search for.

The first difference between NuMVC and q-NuMVC resides in the initialization of the candidate solution C (line 4) where instead of constructing C greedily as a vertex cover, we initialize it as a random vertex set of size k.

The second change concerns the test on whether there is no uncovered edge at line 7 of NuMVC from which we remove vertices with high *dscore* from C. This test is inefficient in q-NuMVC since arriving to a k-sized set C with no uncovered edges means that we got the desired solution and there is no need to remove vertices otherwise the solution C would be of size less than k.

Algorithm 5: q-NuMVC(G, k, $cutoff$)

Input : G, k, $cutoff$

Output: k vertex cover of G

1 **begin**
2 initialize edge weights and *dscores* of vertices;
3 initialize the *confChange* array as an all-1 array;
4 construct C randomly until $\mid C \mid = k$;
5 C* := C;
6 **while** *elapsedTime* < *cutoff* **do**
7   **if** *there are uncovered edges* **then**
8   choose a vertex $u \in C$ with the highest *dscore*, breaking ties in favor of the oldest one;
9         $C := C \setminus \{u\}$, $confChange(u) := 0$ and $confChange(z) := 1$ for each $z \in N(u)$;
10   choose an uncovered edge e randomly;
11   choose a vertex $v \in e$ such that $confChange(v) = 1$ with higher *dscore*, breaking ties in favor of the older one;
12         $C := C \cup \{v\}$, $confChange(z) := 1$ for each $z \in N(v)$;
13         $w(e) := w(e) + 1$ for each uncovered edge e;
14   **if** $\overline{\mathrm{w}} \geq \gamma$ **then**
15             $w(e) := \lfloor \rho \cdot w(e) \rfloor$ for each edge e;
16   **end**
17   **end**
18 **end**
19 **return** C*;
20 **end**

Then, the remaining steps in q-NuMVC (from line 8 to 15 in Algorithm 5) are the same as in NuMVC (from line 12 to 19 in Algorithm 4) as we exchange two vertices iteratively until C becomes a vertex cover.

Thus, we performed only limited changes on the original version of the NuMVC algorithm and we believe that this change will not affect the performance of the algorithm. In the following, we test the q-NuMVC and KEWLS for the k-vertex cover problem.

q-NuMVC performance against KEWLS. Above, we proposed to use the q-NuMVC algorithm instead of KEWLS for random clique finding in the clique-superposition model because of their performance in terms of running time (this test was performed in [25]). Hereafter, we will test both algorithms on their efficiency for finding a solution.

To do so, we have considered a sample of 230 personal networks from the set of authors in $dataset_1$ evolving from 2008 to 2009 for $k = 3$; the sample is the same produced in Section 5.3.1.3. On these personal networks we applied KEWLS and q-NuMVC algorithms to search for cliques of size 2, 3, 4, and 5. For KEWLS, we used the parameter $maxSteps$ witht the values 7, 20, 100, and 200, while for q-NuMVC, we used for the parameter $cutoff$ the values 10, 100, 200, 1000.

For the KEWLS algorithm, we computed the number of OPNs that contains a given clique size (2, 3, 4, and 5) that KEWLS could find as well as the number of OPNs containing a given clique size but that KEWLS could not find. For the OPNs that belong to the later category, we run KEWLS again with $maxSteps$ equal to 20, 100 and 200 if necessary. The results are given in Table 5.1.

MaxSteps	**Clique found**	**2-clique** 230 networks	**3-clique** 230 networks	**4-clique** 230 networks	**5-clique** 221 networks
7	false	147	202	220	216
	true	83	28	10	5
20	false	76	167	197	203
	true	154	63	33	18
100	false	6	61	121	174
	true	224	169	109	47
200	false	1	26	91	122
	true	229	204	139	99

Table 5.1: Performance results of the KEWLS algorithm.

For q-NuMVC, the same strategy is followed since we computed the number of OPNs that contains a given clique size (2,3,4, and 5) that q-NuMVC could find, as well as the number of OPNs containing a given clique size but that q-NuMVC could not find starting with $cutoff = 10$. For the OPNs that belong to the later category, we run q-NuMVC again with a $cutoff$ equal to 100. The results are given in Table 5.2.

$cutoff$	**Clique found**	**2-clique** 230 networks	**3-clique** 230 networks	**4-clique** 230 networks	**5-clique** 221 networks
10	false	2	0	2	10
	true	228	230	228	211
100	false	0	0	0	0
	true	230	230	230	221

Table 5.2: Performance results of the q-NuMVC algorithm.

It is important to point out that, in both cases, we only search for a clique of a given size that do exist in the personal networks.

For the KEWLS algorithm, we can see that its efficiency improves with the increase of $maxSteps$ parameter. However, as the size of cliques to find increases, KEWLS has more difficulty to find such cliques. We should also note that, increasing $maxSteps$ means also increasing the time needed to find cliques.

For the q-NuMVC, starting with a $cutoff = 10sec$, the results are already very good and with a $cutoff = 100sec$, q-NuMVC is able to find all the cliques in all the networks whatever the size. Thus, we proved based on our data that q-NuMVC is a more efficient algorithm than KEWLS.

5.3.3 The description of the proposed evolution model, PERSON*EM*

In this section, we present the new evolution model for personal co-authoring networks that we propose, called PERSON*EM*(PERSonal cO-authoring social Networks Evolution Model). PERSON*EM* is based on the clique-superposition model and integrates the solutions described previously in Section 5.3.2, including the q-NuMVC algorithm for random k-clique finding that we discussed in Section 5.3.2.3 and that was given in Algorithm 5.

The pseudo-code of the new PERSON*EM* evolution model is given in Algorithm 6. The new model takes as an input a personal network at time t and predicts its structure at time $t+1$, under the form of a personal network with a given number of nodes. Thus, the model predicts to which nodes these new nodes will connect, and, indirectly, the number of new connections.

The PERSON*EM* model takes as input the parameters G_{init} - the personal network at time t, k - the depth of G_{init}, p_{node} probability, $p_{sampleNew}$ and $p_{sampleOld}$ probabilities, N - the number of nodes of the predicted network and β - the parameter of the power law distribution, and it provides as output the personal network graph G with N nodes. The predicted personal network graph G is initialized with the personal network graph at time t, G_{init}, given by $newGraph()$ (line 2).

The PERSON*EM* evolution model, based on the clique-superposition model, follows the two phases described hereafter:

1. the fist phase allows to randomly add a new node to the network, or connect a randomly chosen node to a (sub)-clique. This phase corresponds to the clique-superposition model, with several modifications that we will describe below.

Algorithm 6: PERSON*EM* (PERSonal cO-authoring social Networks Evolution Model) evolution model.

```
Input  : G_init, k, p_node, p_sampleNew, p_sampleOld, N, β
Output: G
begin
Graph G ← G_init;
while G.getNodes() < N do
    if Random() < p_node then
        Node current ← newNode();
        G.addNode(current);
    else
        Node newMember ← G.getRandomNode();
        q ← SamplePowerlaw(β);
        while ! valid_q_clique do
            Clique C ← V/q-NuMVC(Ḡ, |V| - q);
            if newNodes.contains(newMember) then
                p_sample = p_sampleNew;
            else
                p_sample = p_sampleOld;
            end
            if C ≠ null then
                if Random < p_sample then
                    Clique C′ ← SampleValidSubClique(C);
                else
                    Clique C′ ← C;
                end
                for i = 1, . . . ,Size(C') do
                    G.addEdge(newMember, C′.node[i]);
                    //if the edge already existing, the edge weight plus 1
                end
            end
        end
    end
end
connectIsolatedNodes(G);
Return G;
end
```

2. the second phase allows to connect the new nodes that, after the first phase, are isolated, to a (sub)-clique.

As in the clique-superposition model, the first phase is composed of two steps: the node evolution step and the edge evolution step. The execution of one of the two steps is driven by the probability p_{node}, and the steps are performed till the total number of nodes of the network N is reached. The node evolution step (lines 5 and 6) consists in adding a new node the G, as in the clique-superposition model.

In the edge evolution step, we start by selecting randomly a node ($newMember$) from G to connect (line 8) and a random value q from a power law distribution of parameter β (line 9). The value q corresponds to the size of the clique that the model should retrieve next from the network; the goal here is that the node $newMember$ should connect to a clique of size q, or a sub-clique of this clique.

Once the value q generated, PERSON*EM* model searches for a *valid* k-clique in G (line 10). A valid clique is defined as a clique having at least one node at distance less that k from the ego of the personal network. In order to find the q-clique, the PERSON*EM* using the q-NuMVC algorithm that takes as arguments the complementary graph $\bar{G}$, and $|V| - q$ representing the size of the vertex cover to find by q-NuMVC (line 11).

In the next step, the PERSON*EM* model should decide, by using a probability, if the node $newMember$ will get connected to the whole clique retrieved or a sub-clique of it. This probability, p_{sample}, is determined according to whether $newMember$ is a new or an existing node in G at time t (line 12 to 16). We recall that the original clique superposition model integrates one single value for p_{sample} without making a distinction whether $newMember$ is a new or an old node. Depending on the value of p_{sample}, the node $newMember$ will connect either to the entire clique C, or to a sample clique (sub-clique) C' retrieved using the function *SampleValidSubClique(C)* (line 18 to 22). Moreover, the retrieved sub-clique should be valid, which means that it has at least a node at a distance less than k from the ego.

We notice that the constraint of selecting a valid q-clique and to sample a valid sub-clique from that q-clique does not exist in the original clique superposition model. We are adding this constraint in order to not overpass k and maintains a $k-$personal network as discussed in section 5.3.2.

Finally, edges are added between $newMember$ and nodes composing the clique C or sub-clique C' (line 23 to 26).

At the end of the first phase of the PERSON*EM* model, we might have nodes in G that are isolated. In order to connect them, we add the second phase with the function *connectIsolatedNodes(G)* (line 31). This step consists in repeating edge evolution steps until all new nodes are connected. This difference between the edge evolution step from the first phase (from line 8 to 28 in Algorithm 6) and the one in the second phase consists in the selection of the $newMember$ node. In the second phase, we select randomly a node from the isolated set of nodes. Thus, the line:

Node $newMember \leftarrow G.getRandomNode()$;

from the first phase, is replaced with:

Node $newMember \leftarrow isolatedNewNodesList.getRandomNode()$;

in the second phase, in Algorithm 6. This step using the function *connectIsolatedNodes(G)* contributes to solving the problem of disconnected nodes observed with the original clique superposition model discussed in section 5.3.2 since personal networks cannot have disconnected nodes.

During this work, an important study was carried in order to find the right values for the parameters of the model. In the next section, a large discussion on this subject will be carried. This discussion, will be followed by the presentation of the evaluation of the results for the prediction of real personal co-authoring networks by the PERSON*EM* model.

5.3.4 PERSON*EM* model's parameters

The PERSON*EM* model that we proposed and discussed previously uses a set of parameters. In the following, we present the chosen values for each of these parameters; generally, the values are computed from the data that the model is applied on, in our case the DBLP personal co-authoring personal networks that we already described.

q-NuMVC algorithm's parameters. Below, we will discuss the values that we use in the algorithm q-NuMVC. Three parameters are concerned, and our decisions are detailed hereafter:

- $cutoff$ time: when comparing KEWLS and q-NuMVC algorithms, we discussed the effect of the variation of $cutoff$ time parameter in q-NuMVC algorithm. Indeed, we presented in Table 5.2 the results of application of the q-NuMVC algorithm to search for cliques of size 2, 3, 4, and 5 in a set of a set of sampled personal co-authoring networks from the authors in $dataset_1$ evolving from 2008 to 2009 for $k = 3$ (230 personal networks). We used for the parameter $cutoff$ the values 10, 100, 200, 1000, and we concluded that for $cutoff = 10sec$ the results are already very good, and for $cutoff = 100sec$, the q-NuMVC algorithm is able to find all the cliques in all the networks whatever the size. Thus, we kept for the application of q-NuMVC in our model the value $cutoff = 100$.

- γ and ρ: the NuMVC algorithm uses the forgetting mechanism to decrease edges weights periodically. In detail, if the averaged weight of all edges achieves a threshold γ, then all edge weights are multiplied by a constant factor ρ. In the q-NuMVC, we use the same values used by Cai et al. [25] in their experiments of the NuMVC algorithm : $\gamma = 0.5|V|$, where $|V|$ represents the number of nodes of the network on which we search for a vertex cover, and $\rho = 0.3$. We chose to use the same values because in [25], the authors performed a comparative test of NuMVC with various parameter combinations (γ, ρ) for the forgetting mechanism on 7 data instances having a size which can correspond to the size of our personal networks, and found that the parameter combination ($0.5|V|$, 0.3) yields relatively good performance among different other combinations for all instances, and exhibits a better robustness.

PERSON*EM* evolution model parameters. When possible, we used our data to estimate or to have an indication on the model parameters to select. Below, we will provide discuss the values of the three parameters used in the PERSON*EM* model: p_{node}, p_{sample} and β.

- p_{node}: in the first phase of the model, this parameter allows to determine whether we perform a node evolution step or an edge evolution step. In the clique-superposition model, the value used was $p_{node} = 0.1$ which indicates that the model performs in 10% of cases a node evolution step, and in 90% of cases an edge evolution step. This is an information that we cannot derive from our data; indeed, it is impossible to do in some sorts a reverse engineering on this process.

  Thus, in order to give more chance to the addition of nodes instead of the addition of edges, we fixed it at $p_{node} = 0.4$ instead of 0.1 as done in the original clique superposition model.

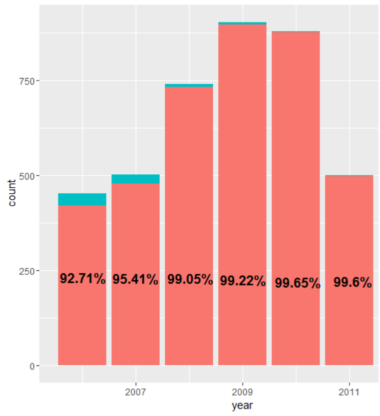
- p_{sample}: in the first phase of the mode, this parameter permits to choose whether we connect the $newMember$ node to the entire selected $q-clique$ or to a sub-clique of the $q-clique$. In the clique-superposition original model, the authors uses $p_{sample} = 0.6$, which indicates that in 60% of cases the model connects the selected node to a sub-clique, and in 40% of cases to the entire clique.

  We decided to use two values: $p_{sampleNew}$ and $p_{sampleOld}$, corresponding respectively to the connection probability of a new node arriving in the network at time $t+1$, and of an old existing node in the network at time t.

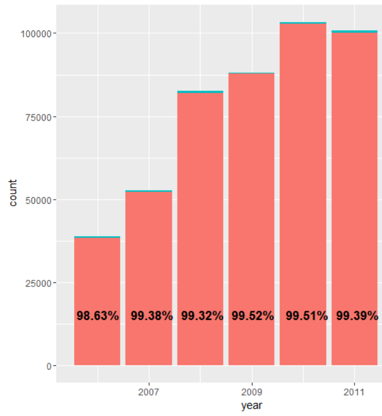
In order to estimate the two values, we computed on OPNs of $dataset_1$ for each k (1 to 5) and each *year* from 2006 to 2011, the percentages of connections between new nodes and cliques and between old nodes and cliques. We found that there is a difference of pattern connection whether the node to connect is a new or an old node. More precisely, we found that if the node that connects to the clique is a new node, it will connect in minimum in 92,71% of cases to a sub-clique (as it is the case on 2006 for $k = 1$ as given in Figure 5.6a), and up to 99,52% of cases to a sub-clique (as it is the case on 2009 for $k = 3$ as given in Figure 5.6b). While if the node is an old node, it has in average 75% of chances to connect to a sub-clique as we can observe it for $k = 3$ on Figure 5.6c (approximately, the same average was observed for the other k values).

While the value of $p_{sampleOld} = 0.75$ seems appropriate, the one of $p_{sampleNew}$ we observed on the data is very restrictive as it only gives 1% of chances to connect to a whole clique. However, we think that the situation where a new node connects to the entire cliques members is expected in our data (for example, in case a prior publication between 3 authors led to the creation of a 3-sized clique, with the arrival of new colleague, a next publication took place between this new colleague and the 3 other authors). So, in order to allow such situation we prefer to decrease $p_{sampleNew}$ value. Thus, we decided to use the following values: $p_{sampleNew} = 0.9$ and $p_{sampleOld} = 0.75$.

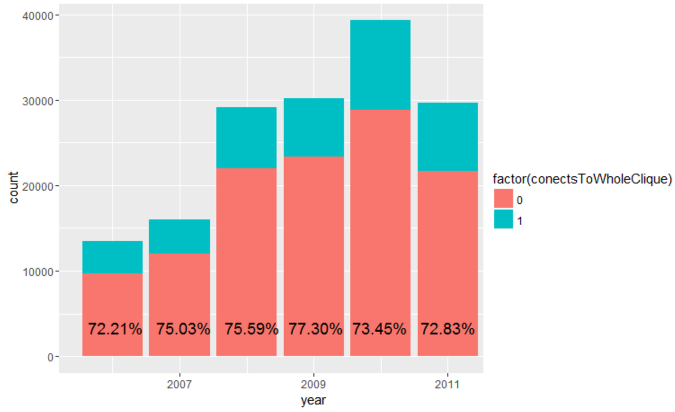
- β: the power law parameter β we use for the model was fixed starting from the power law distribution that the clique size of our data is following which is around $\beta = 3.4$ in average while most of β values are concentrated on 3.5 as we can observe it in Figure 5.8 for $k = 2, 3, 4$ on 2006, 2008, and 2010. However, the clique-superposition model uses a value of $\beta = 2.5$, and if we compare with the value found in our data, we can see that our β is too high and risks to limit the random selection of the clique size q from the power law distribution to very law values (2 or 3). We give in Figure 5.7 a graphics showing the effect of the power law parameter. We can see from this figure that the more we increase β, the less we have chances to select higher values from the tail on the x axis. Thus, we made the choice of setting its value to $\beta = 3.0$.



(a) Percentages of **new** nodes connections to whole/sub-clique, $k = 1$.



(b) Percentages of **new** nodes connections to whole/sub-clique, $k = 3$.



(c) Percentages of **old** nodes connections to whole/sub-clique, $k = 3$.

Figure 5.6: Percentages of new nodes connections (a, b), and old nodes connections (c) to whole/sub-clique, from 2006 to 2011, $dataset_1$.

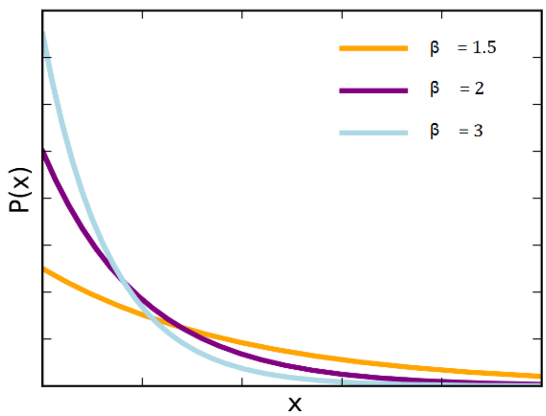


Figure 5.7: Power law β exponent effect.

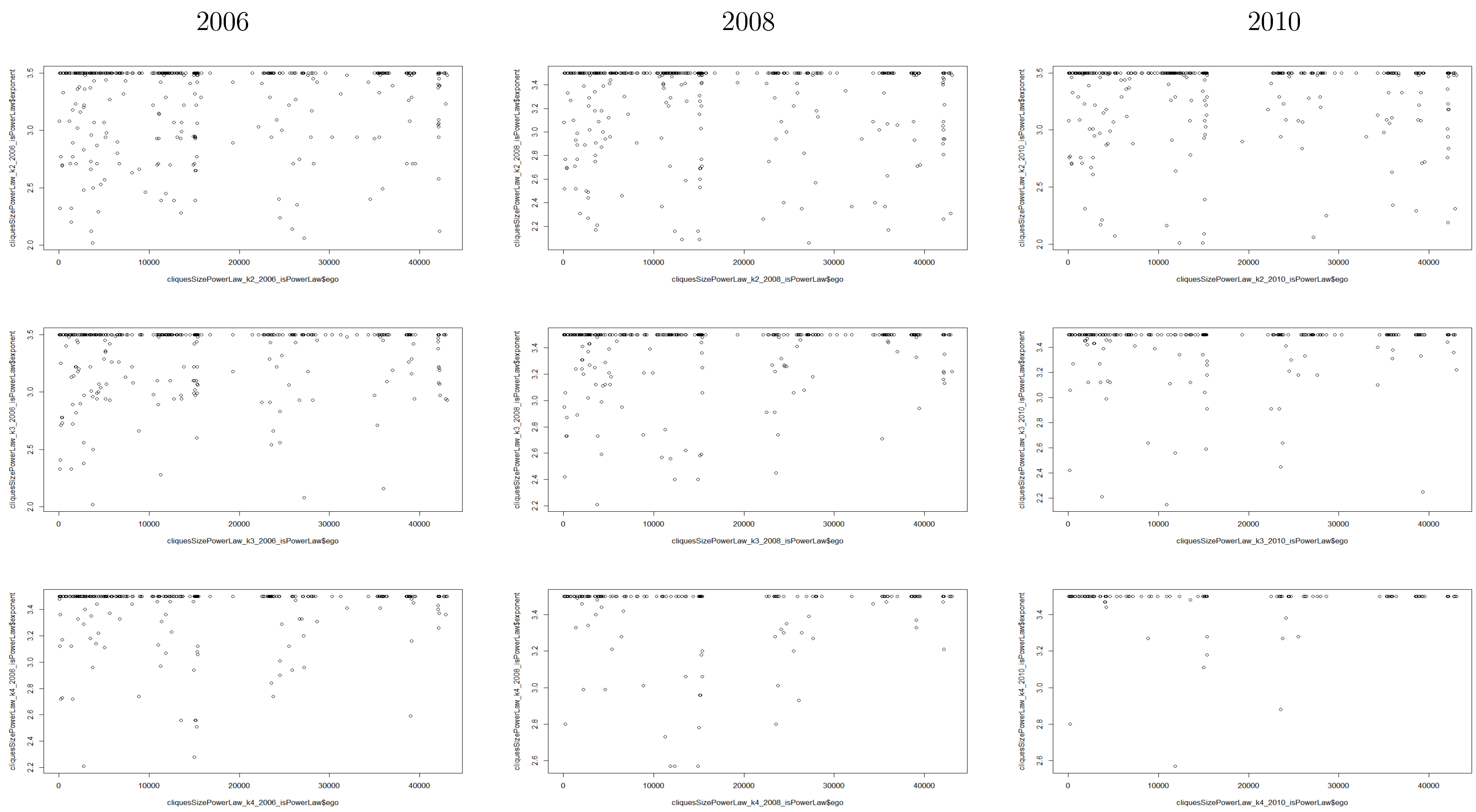


Figure 5.8: Clique size power law exponent for OPNs in $dataset_1$, for $k = 2, 3, 4$, and $year = 2006, 2008, 2010$.

5.3.5 Example of application of the PERSON*EM* model against the clique-superposition model

We consider the OPN of ego 1244 evolving from 2008 to 2009 for $k = 2$. The number of new nodes added from 2008 to 2009 is 4. Figure 5.9a shows the OPN of ego 1244 in 2008, and Figure 5.9b show the OPN of ego 1244 in 2009.

We applied on the OPN of ego 1244 in 2008 given in Figure 5.9a both the clique-superposition model and the PERSON*EM* model in order to predict the network in 2009.

For the clique-superposition model, we used the parameters $p_{node} = 0.1$, $p_{sample} = 0.6$, $\beta = 2.5$ used in the paper presenting the model [134], and the result is given in Figure 5.9c. From the figure, we can see two of the problems explained earlier: the isolated nodes, and the fact that depth of the network, $k = 2$, is over-passed in the predicted network. Indeed, from the four new nodes to connect, only two of them were connected (new_35310 and new_35312) while the two remaining new nodes (new_35313 and new_35311) are isolated. In addition, the connection of node new_35312 makes the network having a depth of $k = 3$.

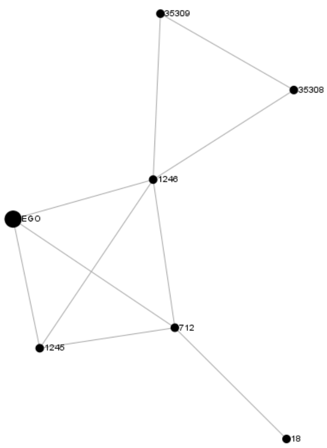
For the PERSON*EM* model, we used the parameters $p_{node} = 0.4$, $p_{sampleNew} = 0.9$, $p_{sampleOld} = 0.75$, $\beta = 3.0$; we can observe in Figure 5.9d that all the new nodes were connected and that the depth of the network, k, is still equal to 2.

5.4 PERSON*EM* model's experimental results

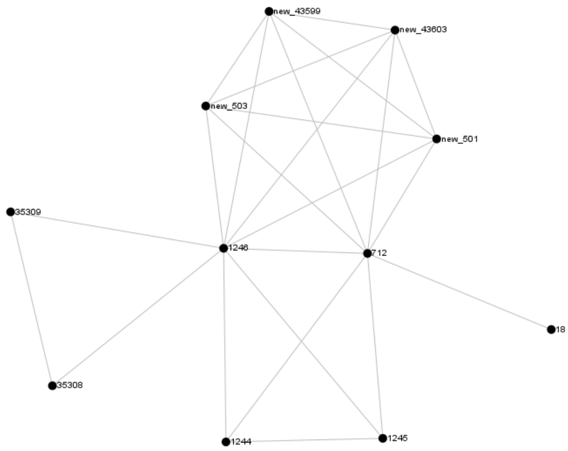
5.4.1 Methodology and objectives

In the following, we present the results of application of the PERSON*EM* evolution model on the personal networks of authors that started publishing on 2004 ($dataset_1$) for $k = 1, 2, 3$, and 4 on three different years 2006, 2008, and 2010 in order to predict the personal networks on the next year ($year + 1$), i.e. 2007, 2009, and 2011 respectively. We notice that we stopped at $k = 4$ since we found from the analysis presented in Chapter 4, that the personal networks of authors at $k = 4$ and $k = 5$ can be assimilated given their common structural and evolution properties.

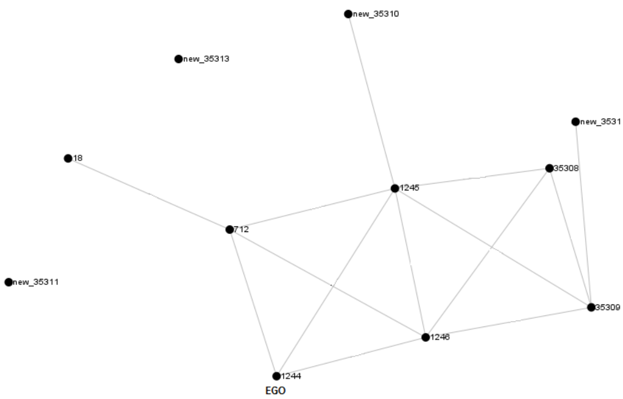
The graphics in Figure 5.10 and Figure 5.11 present a characterization of the networks analyzed via the distribution of the number of nodes and the distribution of the number of edges respectively. The analyzed personal networks have a size in terms of number of nodes and edges that grows with k value and with the years as



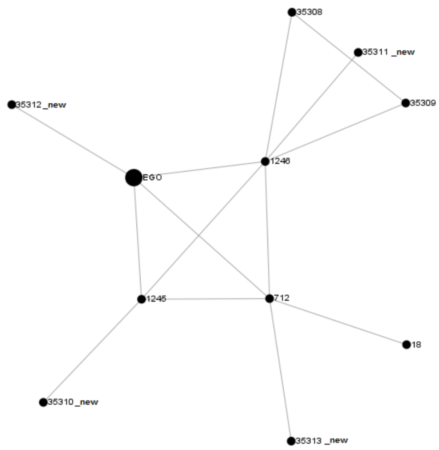
(a) OPN in 2008.



(b) OPN in 2009.



(c) Predicted OPN in 2009 with clique-superposition model.



(d) Predicted OPN in 2009 with PERSON*EM*.

Figure 5.9: Example of application of the clique-superposition model and PERSON*EM* model on the OPN of ego 1244, for $k = 2$ and $year = 2008$.

bigger networks of a size that can reach two thousands of nodes, and more than 4500 edges for $k = 4$ on 2010.

We first evaluate the results of prediction of the model via a set of metrics computed on each predicted OPN. The computed metrics are: the number of edges, the density, the global clustering coefficient, the average local clustering coefficient, the average degree, the ego degree and betweenness.

For each metric, we perform three different types of graphics:

- The value of the metric for each predicted OPN in order to check whether we detect the same evolution trends as the ones discovered on the real data.

- The difference in absolute value between the value of the metric in the predicted

OPNs and its value in the real OPNs ($delta = |predicted - real|$). This will allow us to see if the value of the metrics of the real and the predicted OPNs are close or not.

- The value of the metric in both the real and predicted OPNs on the same graphics with two different colors in order to observe more precisely for each OPN the value of the metric in both real and predicted cases.

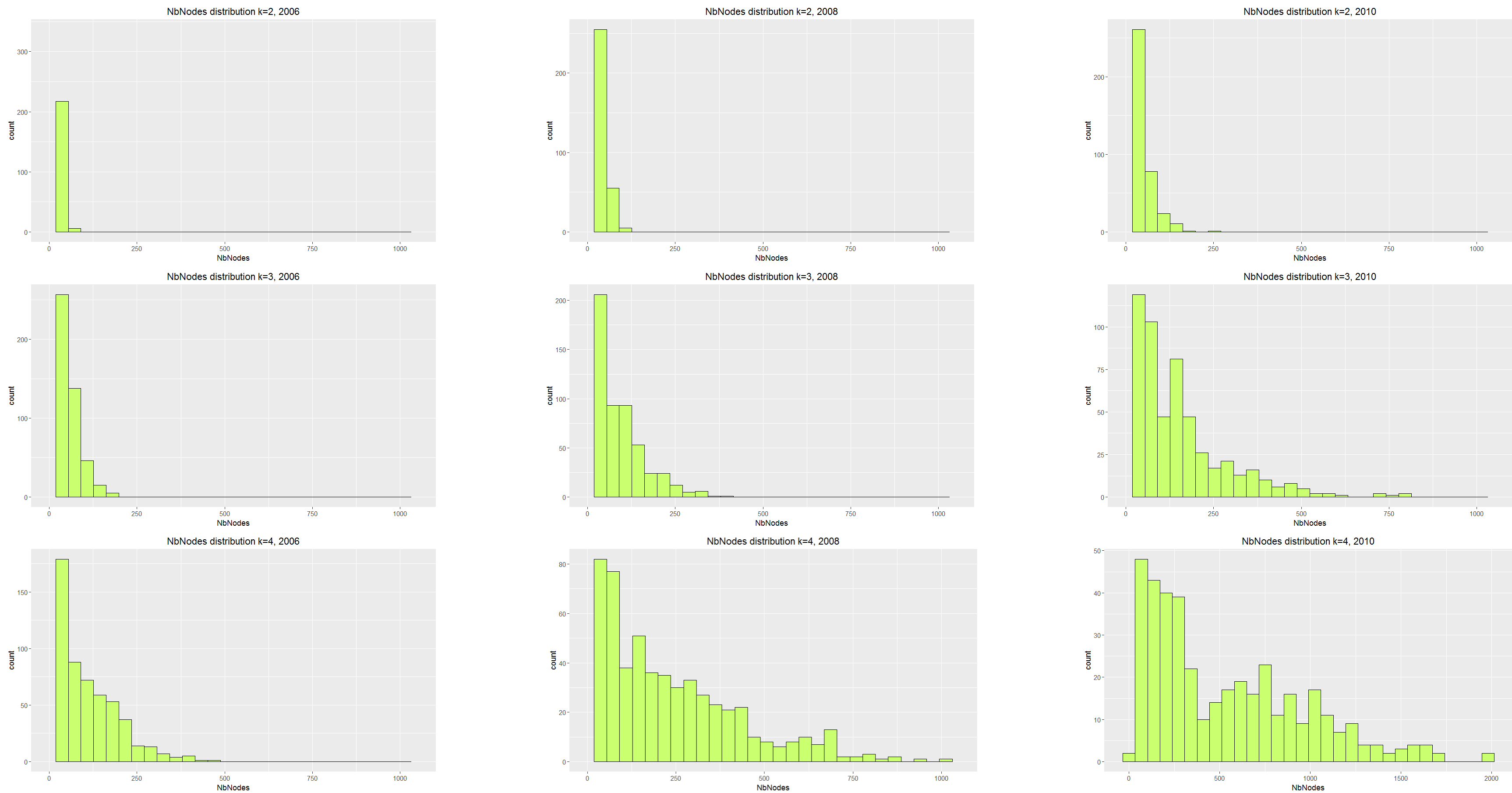


Figure 5.10: Distribution of the number of nodes of OPNs in $dataset_1$.

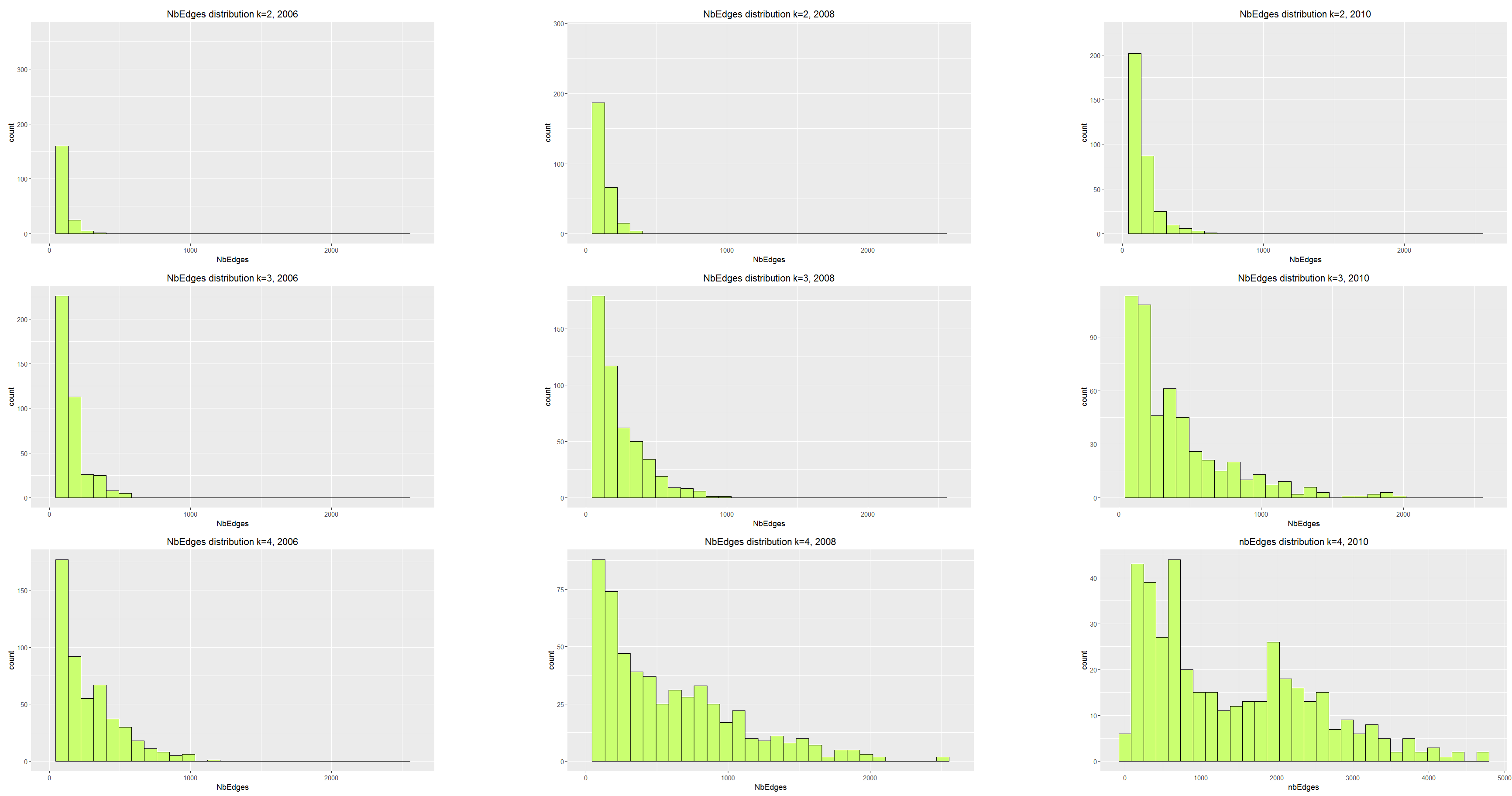


Figure 5.11: Distribution of the number of edges of OPNs in $dataset_1$.

5.4.2 Results for personal co-authoring networks with $k = 1$

1-level personal networks that we analyzed are a particular case. In fact, we found that the majority of the OPNs on the three years 2006, 2008, and 2010 constitute complete networks. The exact percentages are given in Table 5.3.

The problem with complete networks when applying our clique model is that in case we start with an edge evolution step, the model will run indefinitely trying to search for vertex cover with q-NuMVC inside a disconnected complementary graph where no edge exist among the nodes. Thus, the model is not applicable as it is on complete networks.

The remaining set of OPNs on which the model can work is rather small and risks to render the establishment of conclusions limited. We report in Figure 5.12 the results of computation of the difference of metrics values (not in absolute value) between the predicted and real OPNs on $year = 2008$, where we have the largest set of OPNs (about 60 OPNs) comparing to the other years. We notice that we obtained more or less the same results on 2006 and 2010.

Figure 5.12 gives the graphics with the difference between predicted and real OPNs (predicted - real) of each of the following metrics as presented in the Figure 5.12 from the top left to the top right and from the top to the bottom: number of edges, density, average degree, ego degree, ego betweenness, GCC, and ALCC. We discuss the results for each metric hereafter.

1. Number of edges: From the corresponding graphic, we can see that the majority of the values are around zero. We can also notice that when the model do not predict the exact number of edges, this one is lower than the real number of edges which means that the model assigns less edges than needed.

2. Density: The density of the predicted 1-level OPNs remains close to the one of the real 1-OPNs as the difference fluctuates around 0.

3. Average degree: As for the number of edges, the average degree is a bit lower than the average degree of real OPNs when it is not around 0. This observation is thus consistent with the fact that the model assigns less edges than needed.

4. Ego degree: There is a considerable proportion of predicted networks where the degree of the ego is exactly the same as in real networks. However, in many other networks the ego degree is lower. This again joins the fact that the model assigns in general less edges including edges with the ego.

Year	2006	2008	2010
Complete OPNs	90.86%	84.62%	87.84%
Analyzed OPNs	9.13%	15.36%	12.15%

Table 5.3: Percentage of complete OPNs and analyzed OPNs, $k = 1$.

5. Ego betweenness: The betweenness of the ego remains in most of the predicted 1-OPNs close to the betweenness in real OPNs with some cases where it can be either higher or lower.

6. GCC: The global clustering coefficient is also well maintained by the model as most of the values of GCC of the predicted OPNs are close to the values of real OPNs. The difference exceeds rarely 0.2 (in absolute value).

7. ALCC: From the last graphics in Figure 5.12, we can first observe that the average clustering coefficient of the predicted OPNs is lower. In addition, while for a set of OPNs the difference is situated between 0 and -0.2, a considerable set of OPNs have a higher difference that is between -0.2 and -0.4 .

As we have advanced, it is difficult to conclude on whether the model works well or bad on 1-level personal networks given the small size of the set on which we applied it. We are thus enable to confirm strongly a given tendency except the fact that the model assigns less edges. In addition, we might not be able to support the discussed observation when looking on the other levels ($k = 2, 3, ..$) of the OPNs since 1-personal networks are a particular case as we saw it in Chapter 4.

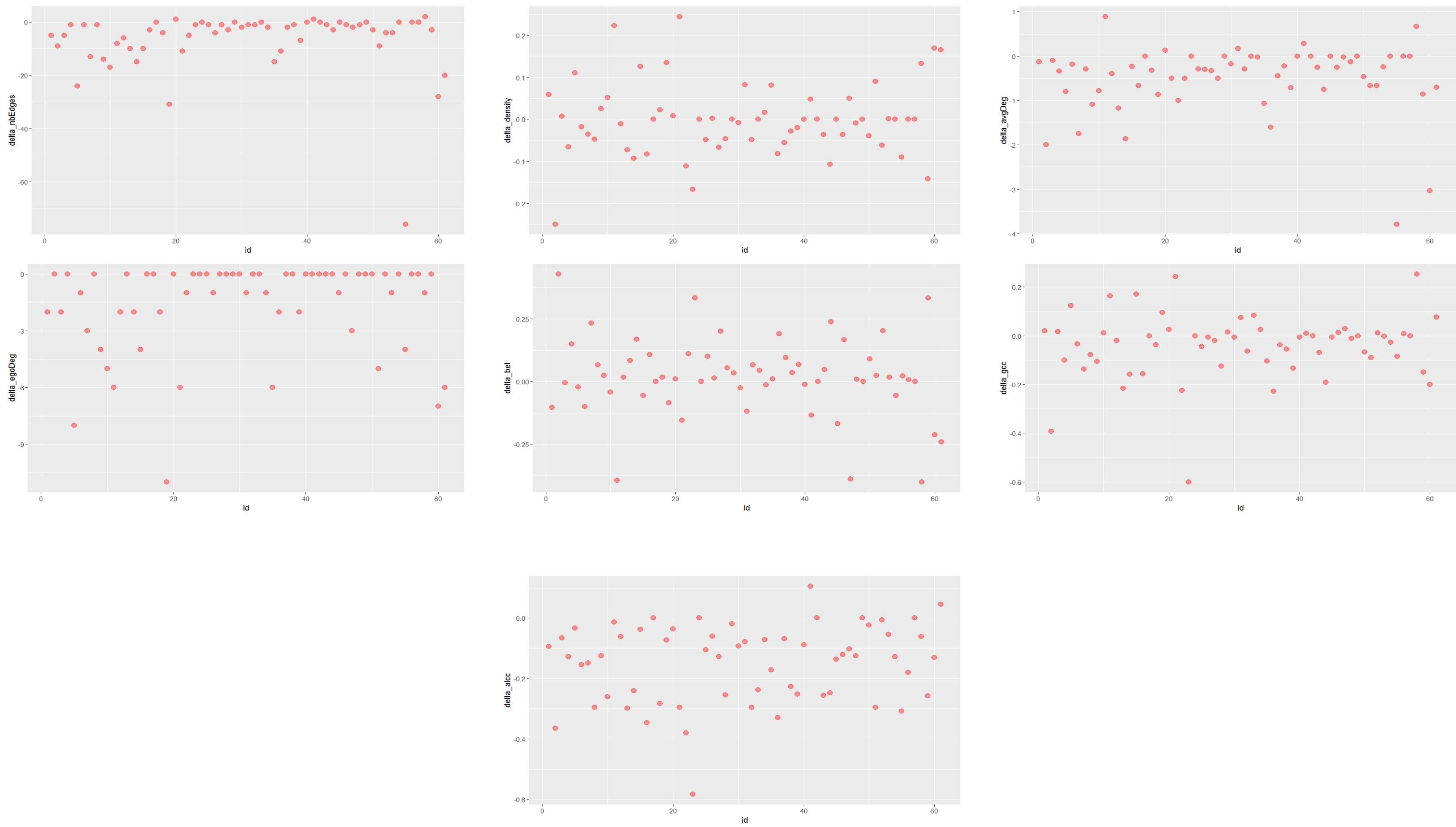


Figure 5.12: Metrics difference between predicted and real OPNs for $k = 1$, $year = 2008$.

5.4.3 Results for personal co-authoring networks with $k = 2, 3, 4$

In the following, we discuss the observations for each metric, for the different values of $k = 2, 3, 4$ and $year = 2006, 2008, 2010$.

Number of edges. For the number of edges, we did not plot the value for each predicted OPN since it is obvious that the evolution trend over the years would be the increase of the number of edges. Thus we concentrate on the difference between the number of edges in the predicted and real OPNs. From figure 5.13, we can notice that the prediction is better for k =2 and $k = 3$ than for $k = 4$. We can conclude that for OPNs with $k = 2, 3$, the difference between the two number of edges is close to 0 in many cases, and in specific cases the difference can be close to 200 edges. In comparison, for $k = 4$ the difference between the predicted and the real number of edges can reach important values such as 800. The difference between the predicted and real number of edges can seem important, but we are more interested in validating that the networks keep the same structure, than having exactly the same number of edges.

In addition, from Figure 5.14, we can observe that in OPNs with a high number of edges the model assigns in most of cases more edges (red dots) than in the real OPNs (green dots). For a better visibility of the later observation, we split the graphic for $k = 4$ and $year = 2006$ into three graphics: in the first one we focus on the 100 first OPNs (Figure 5.15a), then in the second on the 100 next OPNs (Figure 5.15b), and finally in the third graphic on the remaining set of OPNs (Figure 5.15c). In these three graphics, it is easier to observe that the red dots (the number of edges for the predicted OPNs) are above the green dots (the number of edges for the real OPNs).

However, we notice that the difference in the number of edges is not so large in general when focusing on each single OPN. To have a better idea on how important is the difference, we computed the percentage of OPNs with $k = 4$ on $year = 2006$ having a difference in absolute value of the number of edges between the real and predicted value that is between 0 and 50, 50 and 100, and between 100 and 200. The percentages are respectively of 70.34%, 22.46%, and 6.99%. Thus, the majority of predicted OPNs has a difference of number of edges with the real OPNs that is between 0 and 50, and only a small proportion is characterized by a large difference.

It is important to notice that even if the predicted OPNs do not have exactly the same number of edges as real OPNs, what we are interested in is rather predicting

OPNs that have the same structure as the real OPN, that will be verified by evaluating the rest of the metrics.

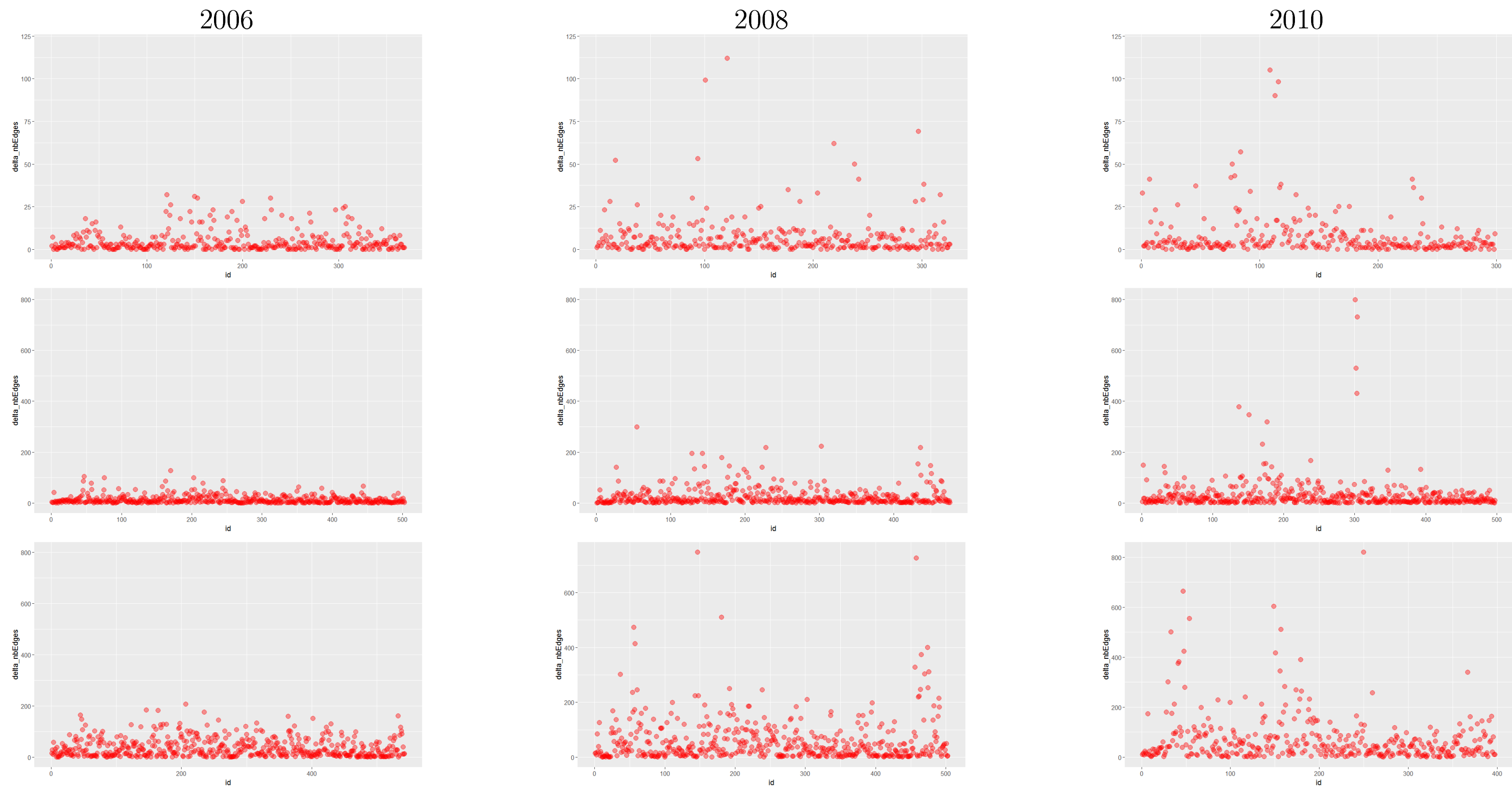


Figure 5.13: Number of edges: difference between predicted and real OPNs (absolute value), for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

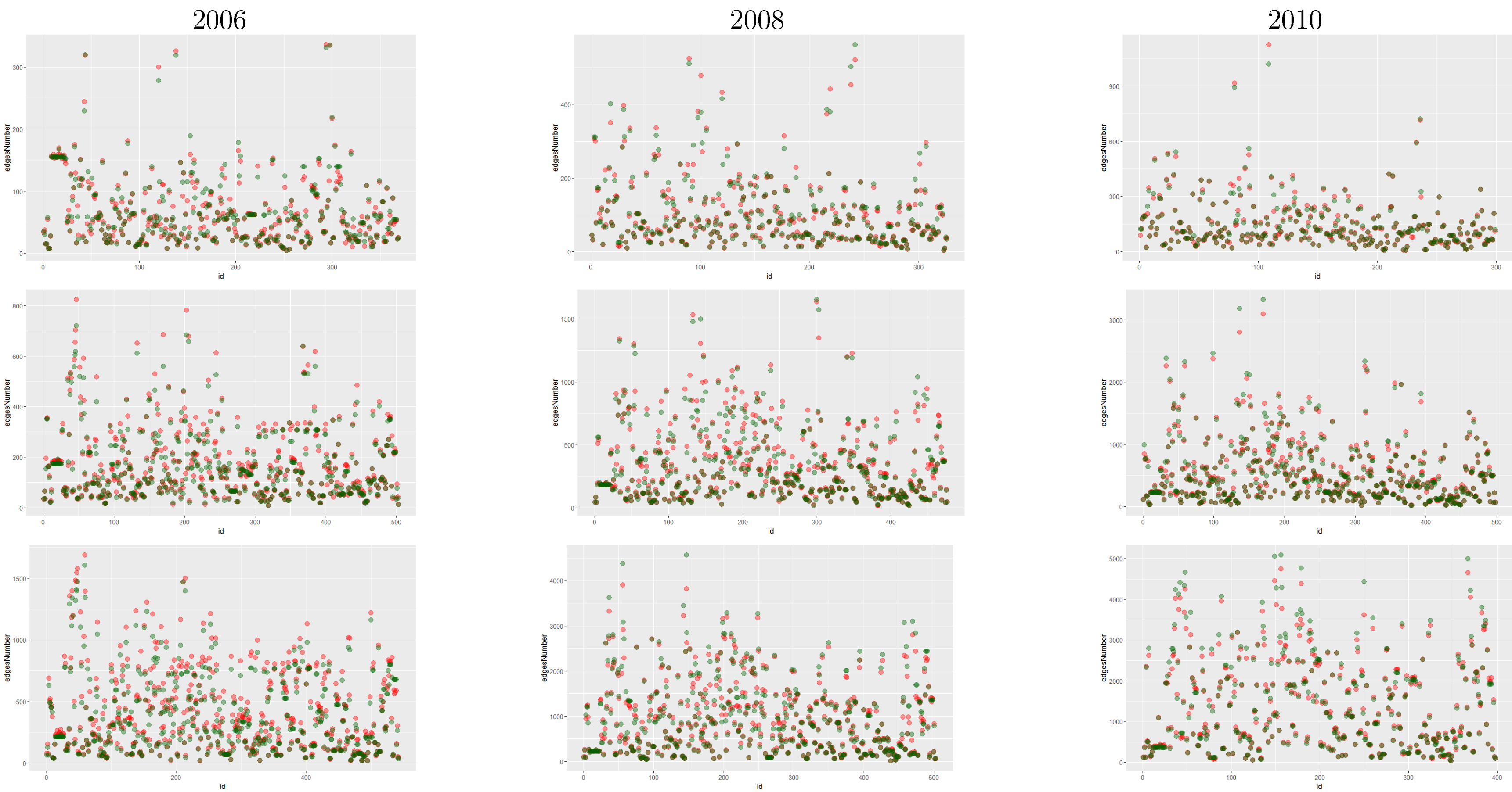
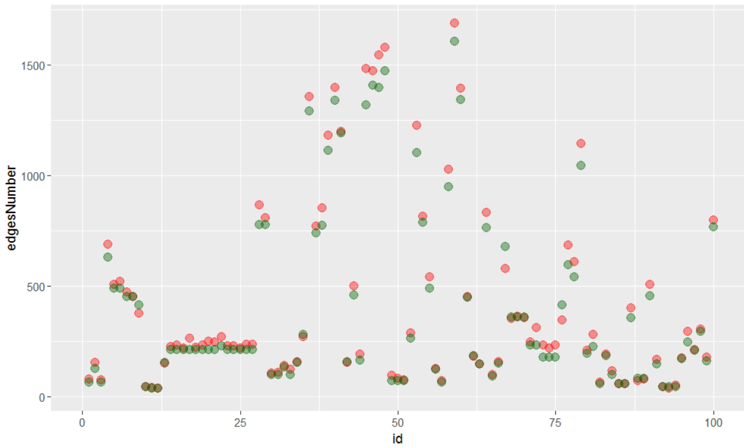
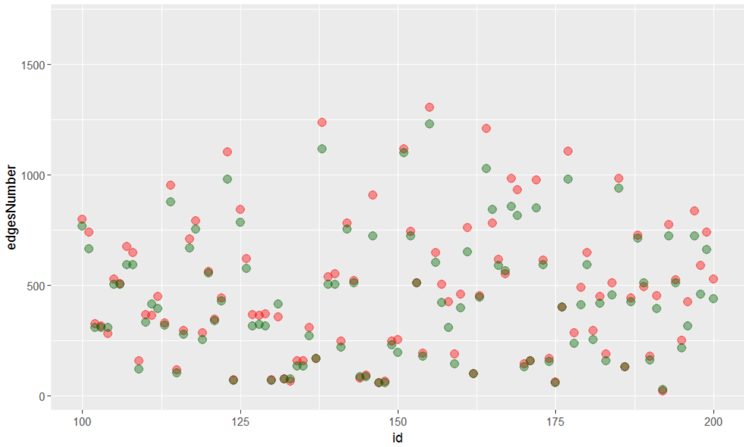


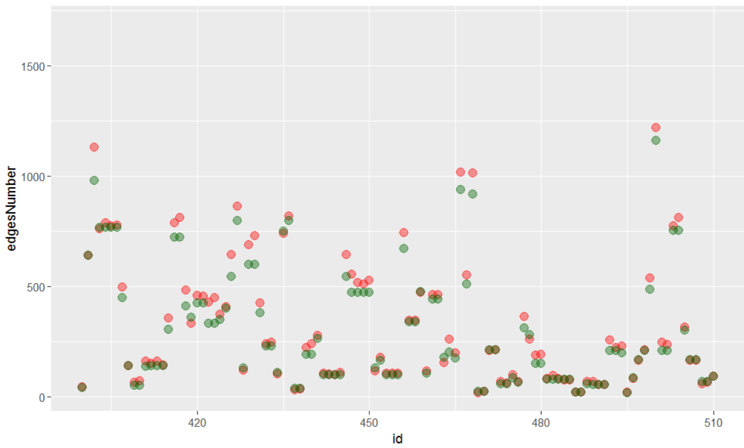
Figure 5.14: Number of edges: predicted (red dots) and real (green dots) OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).



(a) Zoom on OPNs 0-100.



(b) Zoom on OPNs 100-200.



(c) Zoom on OPNs 200-530.

Figure 5.15: Number of edges: zoom on predicted (red dots) and real (green dots) OPNs, for $k = 4$ and $year = 2006$.

Density. Following the same methodology as previously, we study the density of the of the predicted OPNs comparing to the real OPNs. We analyze the density via three phenomena: evolution tendency, difference in absolute value, and value for predicted vs. real OPNs, as describe hereafter:

1. Density evolution tendency: from Figure 5.17, the density of the predicted OPNs decreases over years as we observed it for real OPNs for k =2, 3 and 4. Its value is low in general.

   In order to confirm the decreasing tendency of the density observed from the graphics, we realized a Wilcoxon test as we did it on real OPNs in Chapter 4 to confirm the evolution trends of the different metrics. The results of the Wilcoxon test are given in Table 5.4. The table gives the results of the test applied on the density of the predicted personal networks evolving between 2006 and 2008, and between 2008 and 2010 for $k = 2$ to 4. We give the value of V the statistic value of the Wilcoxon test and the values $p - value_{decreasing}$ and $p - value_{increasing}$ representing the p-value of the Wilcoxon test whether we give as alternative hypothesis "greater" or "less" respectively. We can see that $p - value_{increasing}$ equals always 1 which indicates that the increasing trend cannot explain the evolution of the density, while $p - value_{decreasing}$ that is $<$ 2.2e-16 consolidate the observed decreasing trend as it is significantly lower than 5% (the threshold we use to reject the null hypothesis). Thus, as for real personal networks, the predicted personal networks by the model have also a decreasing density.

Year	2006-2008	2008-2010
$k = 2$	$V = 29155$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 18716$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$
$k = 3$	$V = 96538$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 88681$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$
$k = 4$	$V = 120430$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$	$V = 62963$, $p - value_{decreasing} < 2.2e - 16$, $p - value_{increasing} = 1$

Table 5.4: Wilcoxon test on the density of predicted OPNs, $dataset_1$.

2. Density difference in absolute value: from Figure 5.18, we can observe that there are only few cases where the difference of the density between the predicted and real OPNs is above 0.05, otherwise the density is well reproduced by the model.

The most intriguing case concerns the OPN with $k = 2$ in 2010 where we registered a difference of 0.3 which is the highest difference observed for the density. By looking more closely to this network, we found that it is about a very small OPN with only 5 nodes and 8 predicted edges instead of 5 in the real OPN which resulted in a high density (equal to 0.8). The real and predicted networks are given in Figure 5.16.

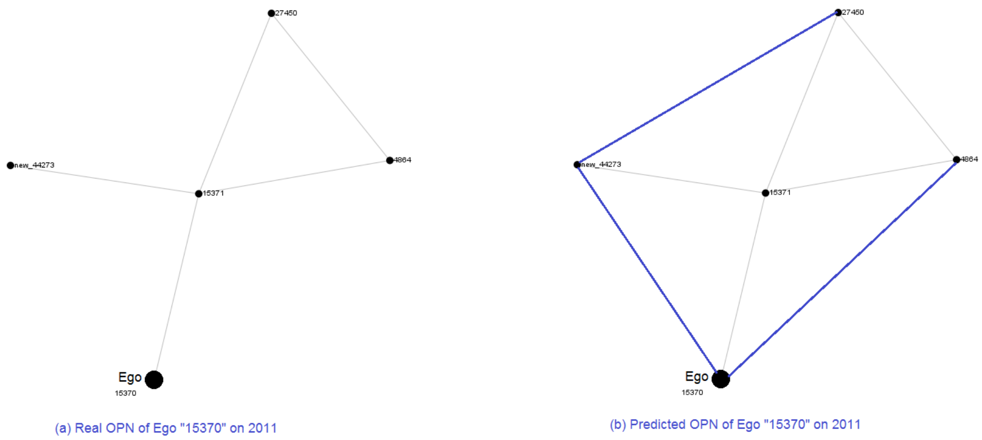


Figure 5.16: Real and predicted OPNs of Ego="15370" on 2011.

3. Density value in predicted and real OPNs: from Figure 5.19, we can observe that the values of the density in real and predicted OPNs are corresponding. As done previously for the number of edges, we make a zoom on the graphic of OPNs with $k = 4$ in 2006 in order to better observe the differences of the density values between predicted and real OPNs. The graphics are given in Figures 5.20a, 5.20b, and 5.20c. From these three graphics we can observe that the density values in real and predicted OPNs is very close and overlap in many cases. This demonstrates that even if we had in some cases a higher number of edges in the predicted OPNs, this does not affect the density of the networks.

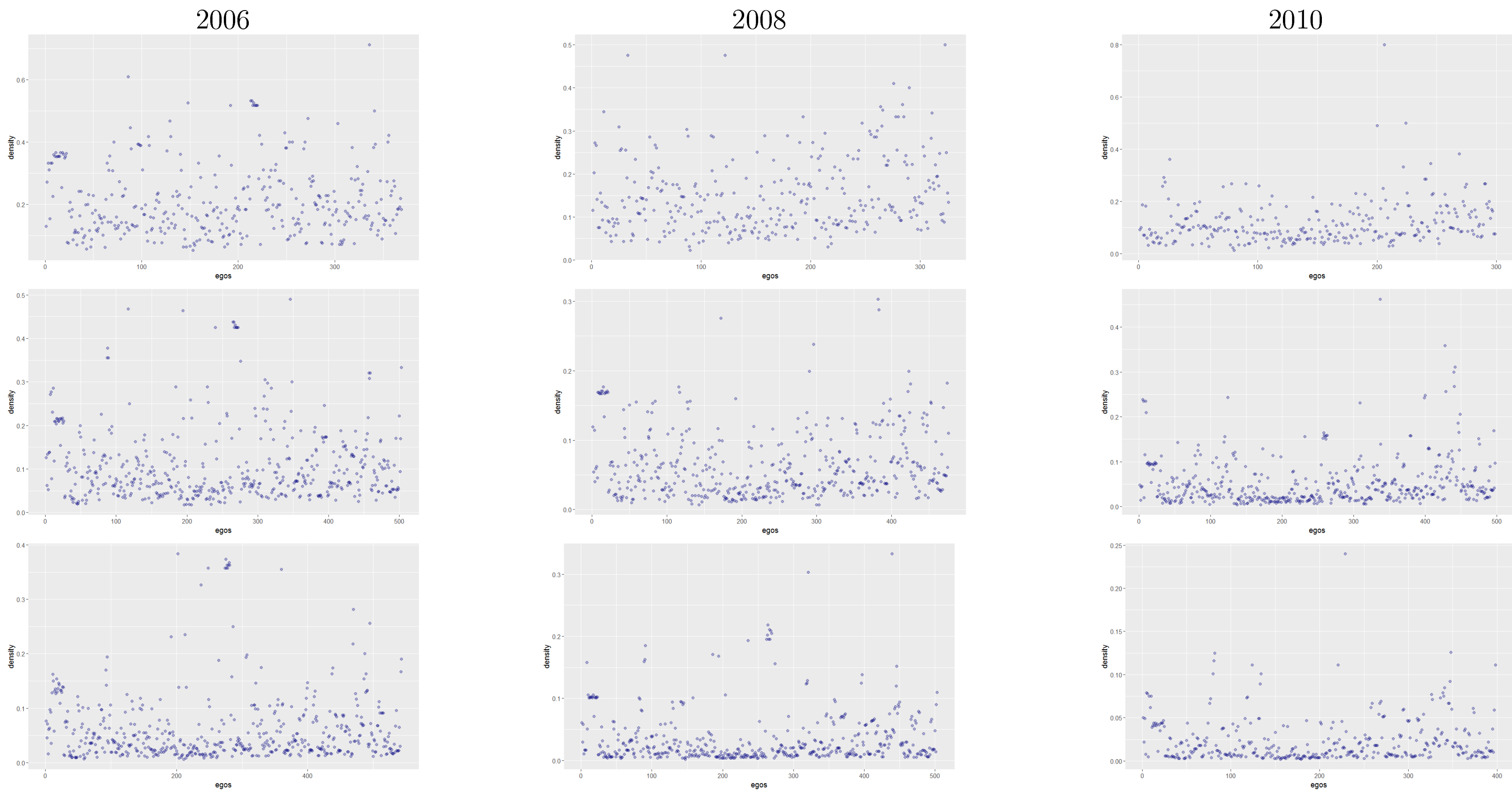


Figure 5.17: Density: predicted OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

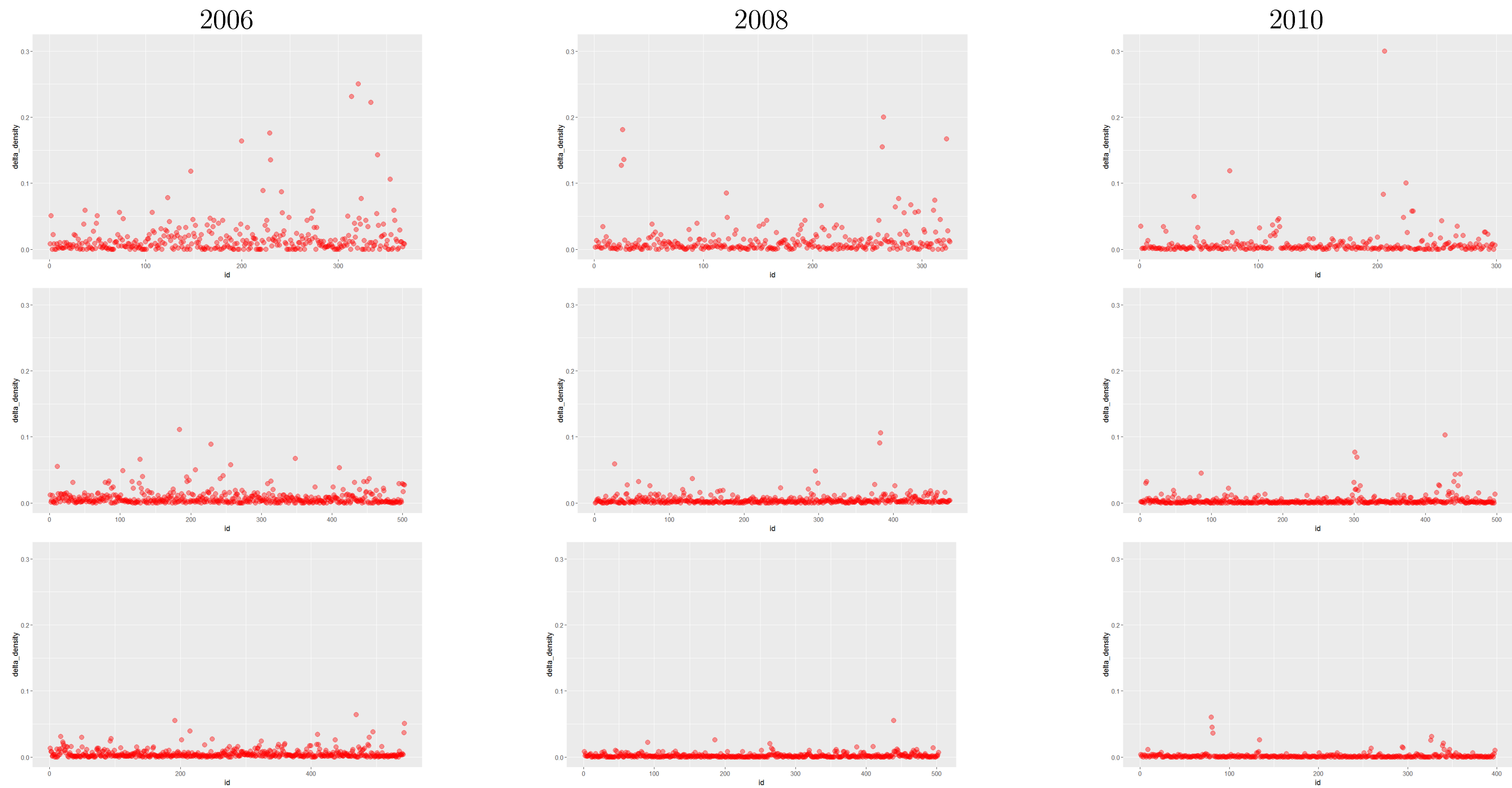


Figure 5.18: Density: difference in absolute value between predicted and real OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

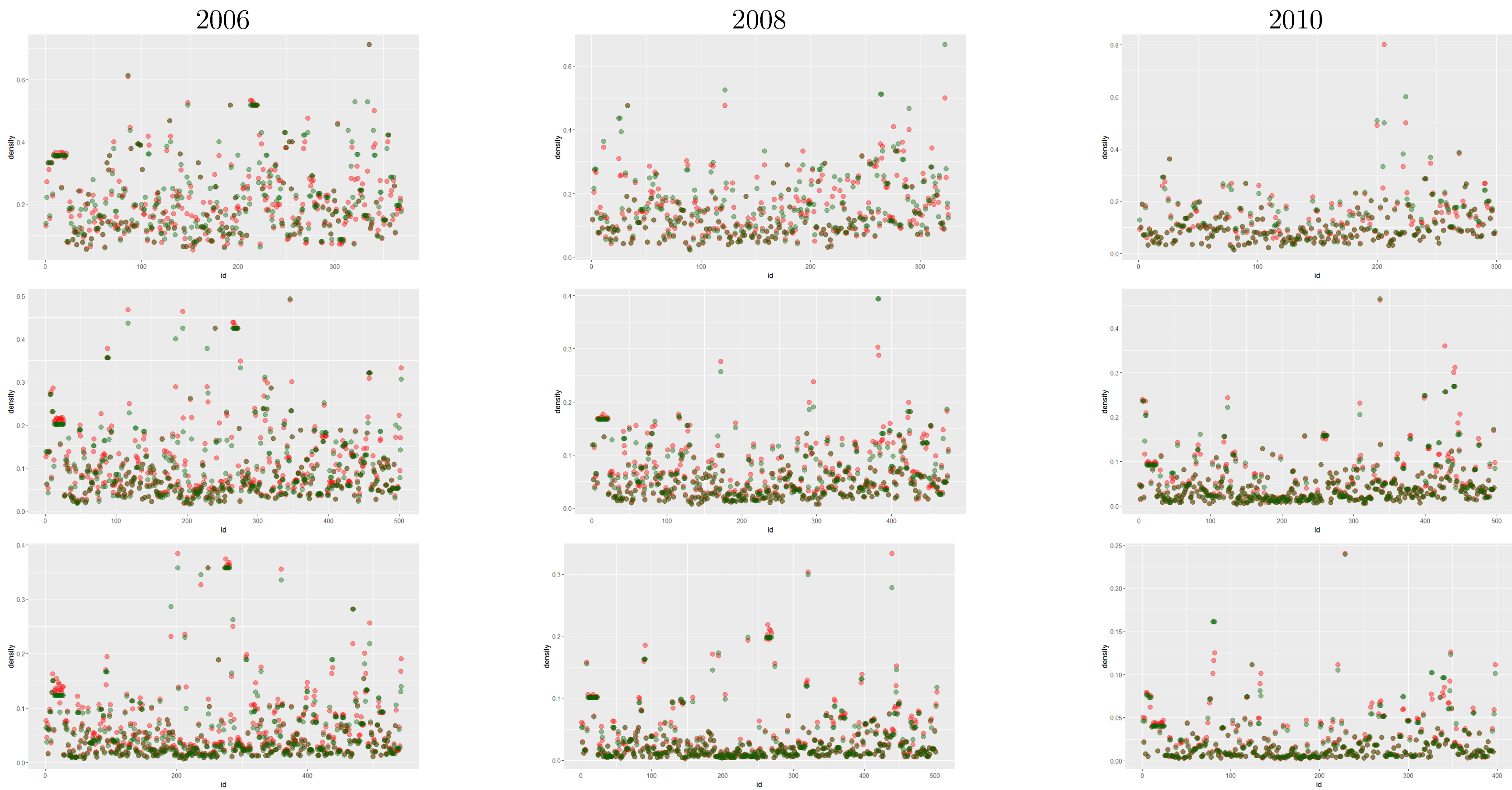
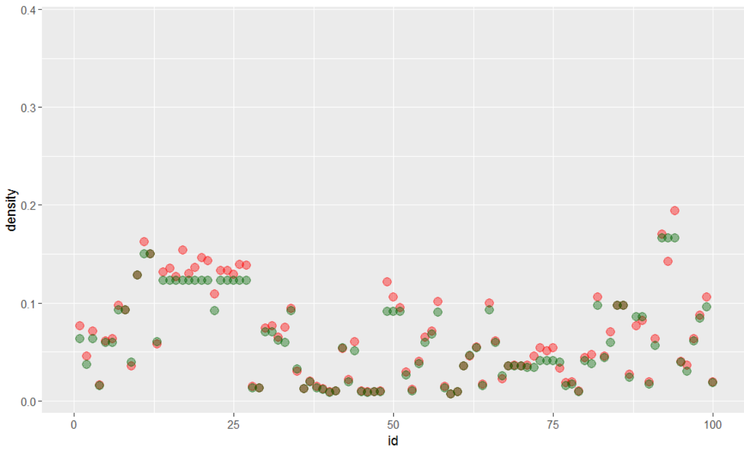
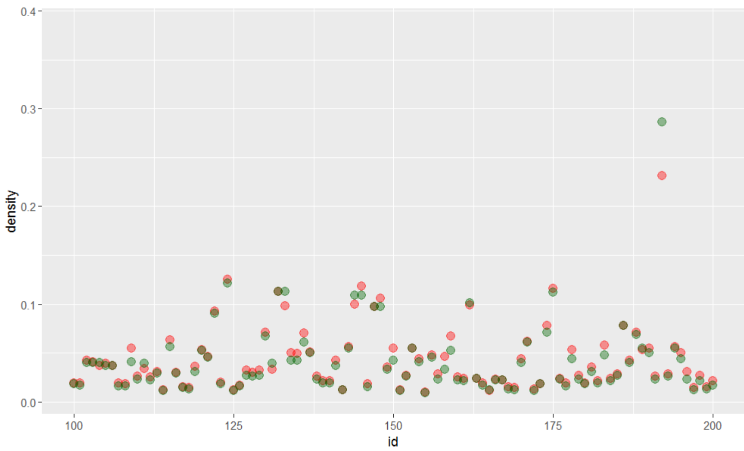


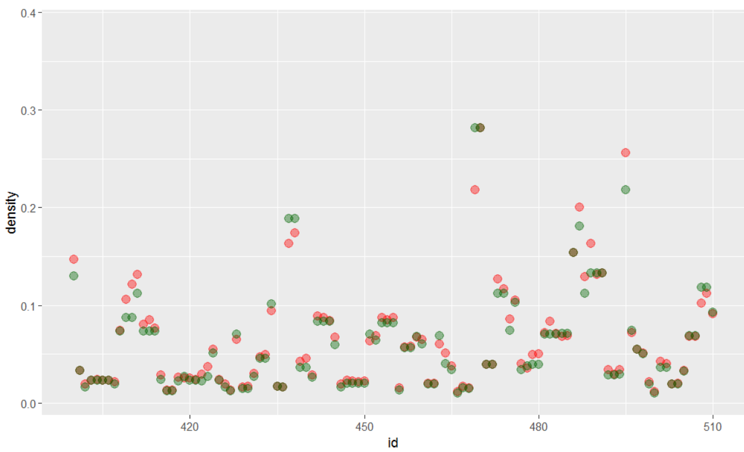
Figure 5.19: Density: predicted (red dots) and real (green dots) OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).



(a) Zoom on OPNs 0-100.



(b) Zoom on OPNs 100-200.



(c) Zoom on OPNs 200-530.

Figure 5.20: Density: zoom on predicted (red dots) and real (green dots) OPNs, for $k = 4$ and $year = 2006$.

Average degree. Following the same methodology as previously, we study the average degree of the of the predicted OPNs comparing to the real OPNs. We analyze the average degree via three phenomena: evolution tendency, difference in absolute value, and value for predicted vs. real OPNs, as describe hereafter:

1. Average degree evolution tendency: in Figure 5.21, we can seen that the average degree of the predicted OPNs increases then concentrates around the values 3, 4 and 5 as for real data.

   As done for the density, we performed a Wicoxon test to confirm the increase of the averge degree of the predicted personal networks over time. The results of the test are given in Table 5.5. We can observe that $p-value_{decreasing}$ equals always 1 (or approaches 1 in the case of $k=4$ between 2006 and 2008) which indicates that the increasing trend explains indeed the evolution of the average degree, while $p-value_{increasing}$ is less than the threshold 5%. This confirms the observed evolution behaviour from the graphics and joins the one explaining real personal networks given in chapter 4.

Year	2006-2008	2008-2010
$k=2$	$V=6940$, $p-value_{decreasing}=1$, $p-value_{increasing}=4.679e-14$	$V=6311$, $p-value_{decreasing}=1$, $p-value_{increasing}=8.798e-06$
$k=3$	$V=30168$, $p-value_{decreasing}=1$, $p-value_{increasing}=3.889e-13$	$V=32497$, $p-value_{decreasing}=1$, $p-value_{increasing}=1.171e-07$
$k=4$	$V=55506$, $p-value_{decreasing}=0.9675$, $p-value_{increasing}=0.03253$	$V=19700$, $p-value_{decreasing}=1$, $p-value_{increasing}=1.121e-10$

Table 5.5: Wilcoxon test on the average degree of predicted OPNs, $dataset_1$

2. Average degree difference in absolute value: in Figure 5.22, we can observe that there is no big difference among the predicted and real OPNs as it remains in the majority of OPNs between 0 and 0.5 and can reach from 1 to 4 only in very few cases.

3. Average degree value in predicted vs. real OPNs: Figure 5.23 shows that the value of the average degree per OPN is almost similar for the predicted and real OPNs. However we notice that the predicted OPNs have an average degree that is a little bit higher than the average degree of real OPNs.

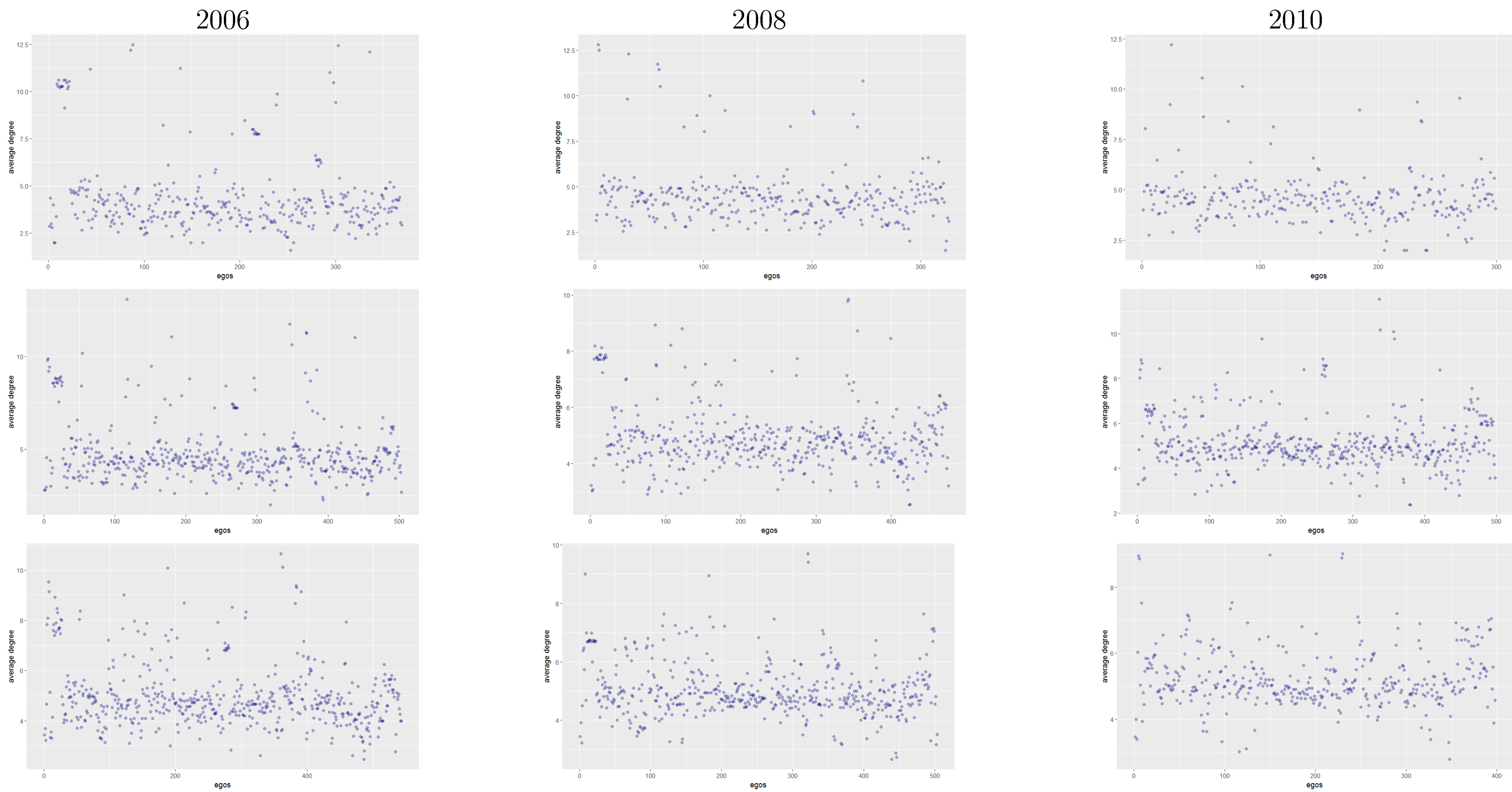


Figure 5.21: Average degree: predicted OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

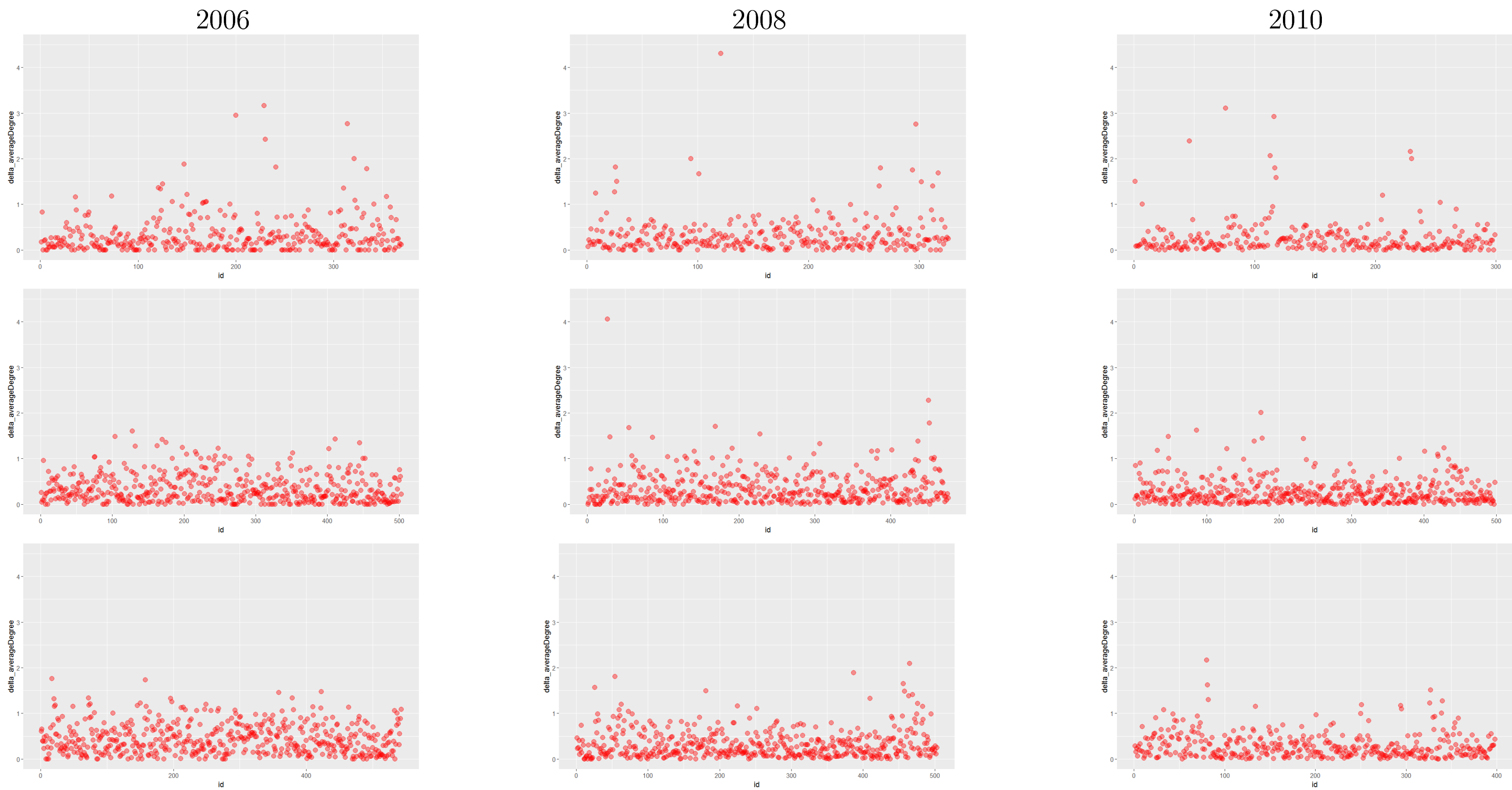


Figure 5.22: Average degree: difference in absolute value between predicted and real OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

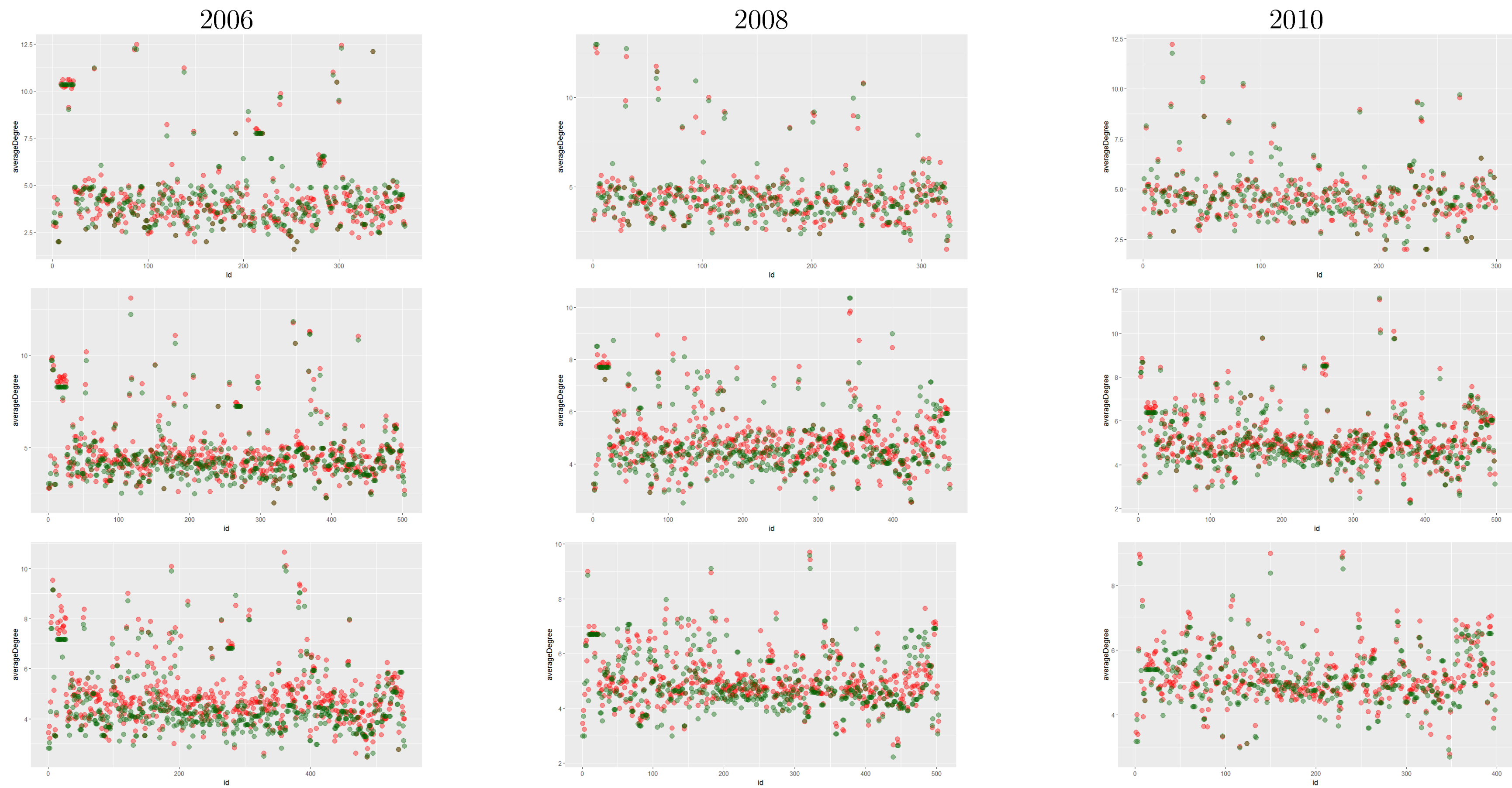


Figure 5.23: Average degree: predicted (red dots) and real (green dots) OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

Ego degree. Following the same methodology as previously, we study the ego degree of the of the predicted OPNs comparing to the real OPNs. We analyze the ego degree via three phenomena: evolution tendency, difference in absolute value, and value for predicted vs. real OPNs, as describe hereafter:

1. Ego degree value: in Figure 5.24, we can see that the ego degree is low in general, and only few egos have a high degree. The tendency is similar among the years.

2. Ego degree absolute difference: Figure 5.25 shows that there is a low difference between the ego degree in real and predicted OPNs concentrated around 0.

3. Ego degree value in predicted vs real OPNs: in Figure 5.26, we can observe that the value of ego degree of the predicted OPNs fits its value in real OPNs, but sometimes it is higher.

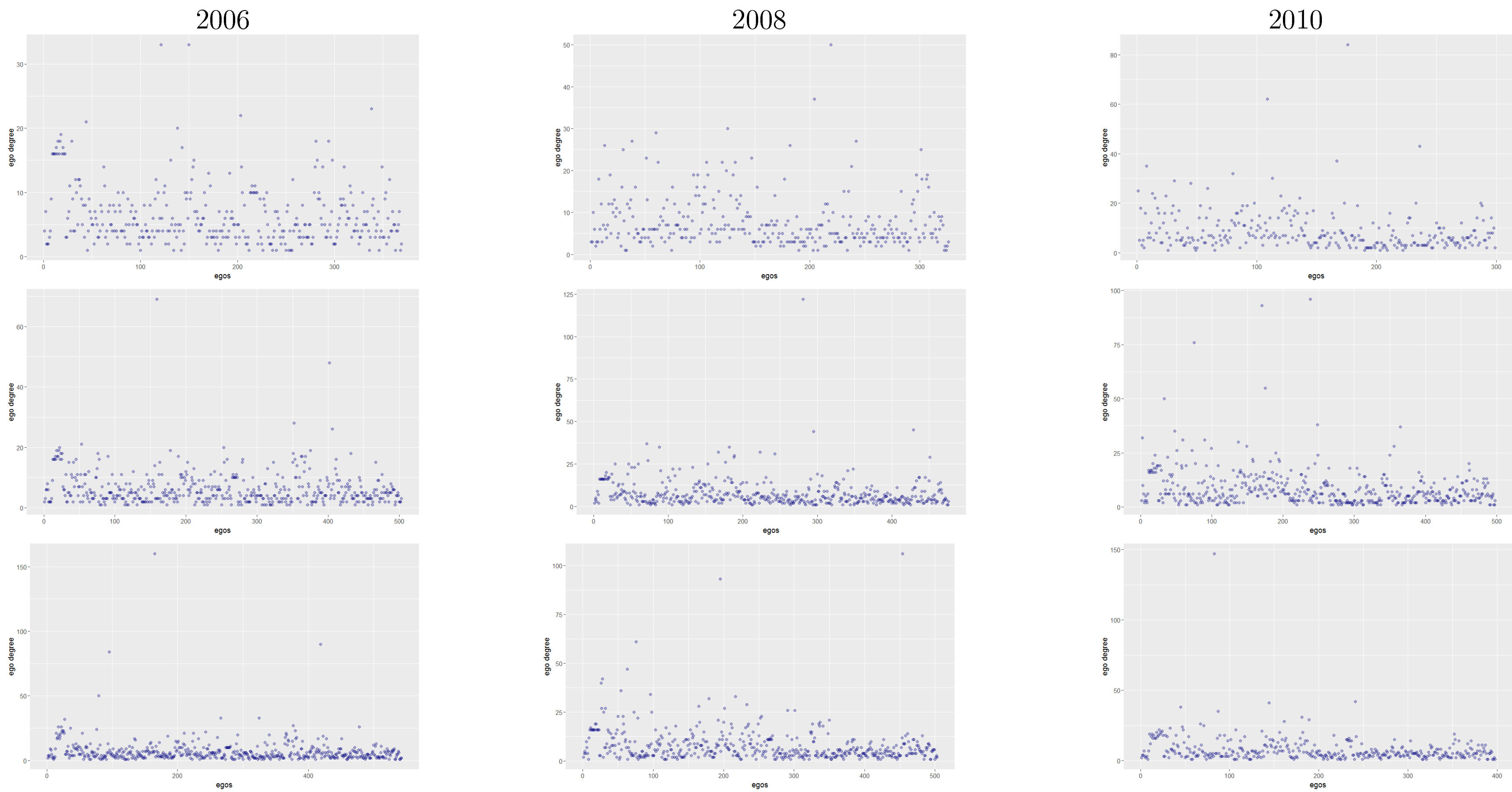


Figure 5.24: Ego degree: predicted OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

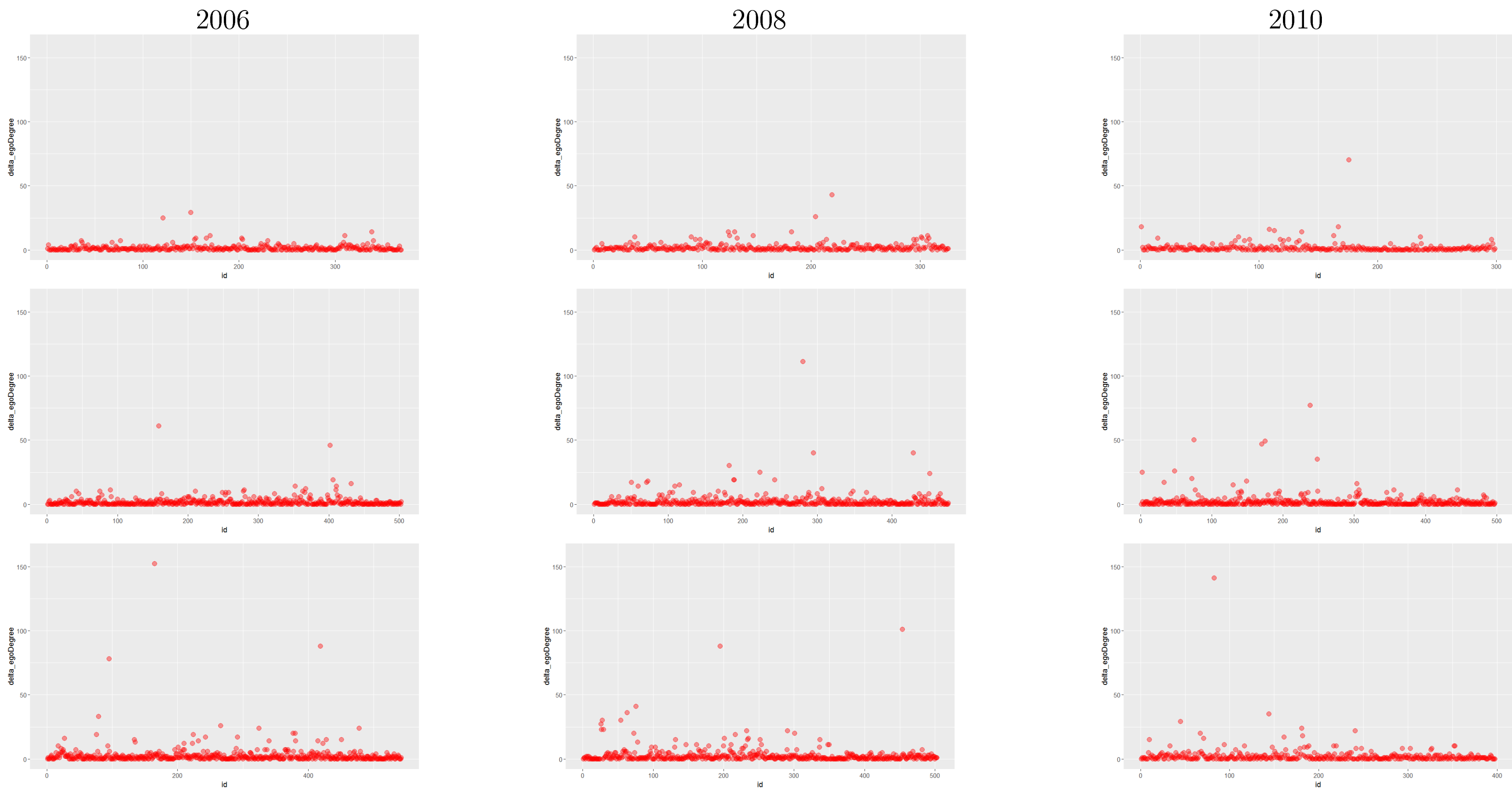


Figure 5.25: Ego degree: difference in absolute value between predicted and real OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

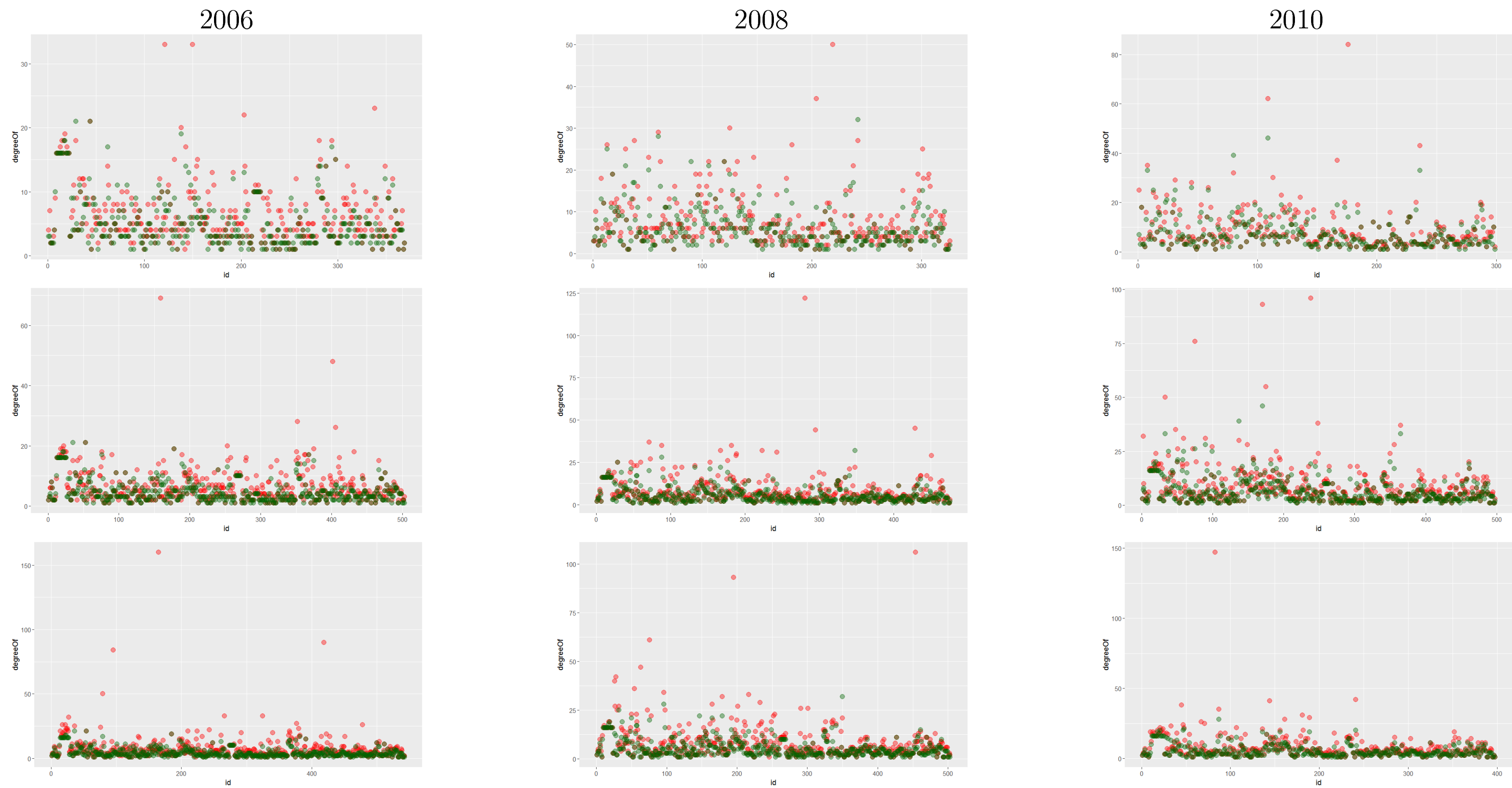


Figure 5.26: Ego degree: predicted (red dots) and real (green dots) OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row)

Ego betweenness. Following the same methodology as previously, we study the ego betweenness of the predicted OPNs comparing to the real OPNs. We analyze the ego betweenness via three phenomena: evolution tendency, difference in absolute value, and value for predicted vs. real OPNs, as describe hereafter:

1. Betweenness evolution tendency: in Figure 5.27, we can observe a higher ego betweenness for small k. The tendency of decreasing/increasing over the years is not clear from the graphic.

   We performed a Wilcoxon test to discover more accurately how the betweenness of the predicted personal networks evolve over time. The results of the test are given in Table 5.6. We comment the results in the following:

   - $k = 2$: $p-value_{decreasing}$ and $p-value_{increasing}$ are both > 0.05 from 2006 to 2008. Thus, we cannot reject the null hypothesis nor establish whether the betweennes increases or decreases. But, between 2008 and 2010 the betweenness is increasing as $p-value_{increasing} < 0.05$ and $p-value_{decreasing}$ approaches 1. Thus, the increasing betweenness of the predicted personal network joins what we discovered about real personal networks.

Year	2006-2008	2008-2010
$k = 2$	$V = 12732$, $p - value_{decreasing} = 0.4459$, $p - value_{increasing} = 0.5545$	$V = 6635.5$, $p - value_{decreasing} = 0.9845$, $p - value_{increasing} = 0.01556$
$k = 3$	$V = 44515$, $p - value_{decreasing} = 5.152e - 06$, $p - value_{increasing} = 1$	$V = 28284$, $p - value_{decreasing} = 0.9796$, $p - value_{increasing} = 0.0204$
$k = 4$	$V = 55768$, $p - value_{decreasing} = 1.255e - 07$, $p - value_{increasing} = 1$	$V = 21271$, $p - value_{decreasing} = 0.05199$, $p - value_{increasing} = 0.9481$

Table 5.6: Wilcoxon test on the betweenness of predicted OPNs, $dataset_1$

   - $k = 3$: the betweenness decreases between 2006 and 2008 as $p-value_{increasing}$ $=1$ and $p - value_{decreasing}$ <0.05. However, it is increasing from 2008 to 2010 given that $p-value_{increasing} < 0.05$ and $p-value_{decreasing}$ approaches 1. We recall that for real personal networks we could not conclude on the evolution trend of the betwenness for $k = 3$ between 2008 and 2010, while between 2006 and 2008 it was increasing.

- $k = 4$: the betweenness decreases as in real personal networks since $p - value_{increasing}$ equals or approaches 1 while $p - value_{decreasing}$ is almost equal or less than 0.05. This joins what we discovered when we considered real personal networks for $k = 4$ between 2008 and 2010 and for $k = 5$.

We can conclude that the Wilcoxon test allowed us to clarify the evolution trend of the betweenness of the predicted networks. These trends join the ones observed on the betweenness of real personal networks.

2. Betweenness absolute difference: we can observe in Figure 5.28, that in the majority of OPNs, the difference of the ego betweenness between real and predicted networks is very low and fluctuates around 0. But, there are cases where the difference is higher and ranges between 0.2 and 0.6, and can exceed 0.6 is some cases.

3. Betweenness value in predicted vs real OPNs: Figure 5.29 confirms the observation on the previous figure. However, we could not confirm whether the betweenness value of the predicted OPNs is higher or lower comparing to real OPNs. To this end, we give in Figures 5.30a, 5.30b, and 5.30c a focus on the OPNs with $k = 4$, in 2006 with both predicted and real value of the betweenness. Again, from these graphics, it is still not clear whether we predict OPNs having a higher or a lower ego betweenness since both situations take place.

   Thus, in order to conclude on this point, we computed the percentage on the differences of the betweenness between predicted and real OPNs (not the absolute value) to see what is the dominant percentage among positive and negative value for the OPNs with $k = 4$ on 2006. This way we can conclude on whether the model predicts OPNs with higher or lower ego betweenness. The obtained percentages are the following:

   - 30.38% of predicted OPNs have an ego betweenness exactly equal to the ego betweenness in real OPNs,
   - 44.19% of predicted OPNs have an ego betweenness that is higher than the ego betweenness in real OPNs, and
   - 25.41% of predicted OPNs have an ego betweenness that is lower than the ego betweenness in real OPNs.

   So, we can conclude that when the ego betweenness is not maintained by the model, it is higher in 44.19% of cases.

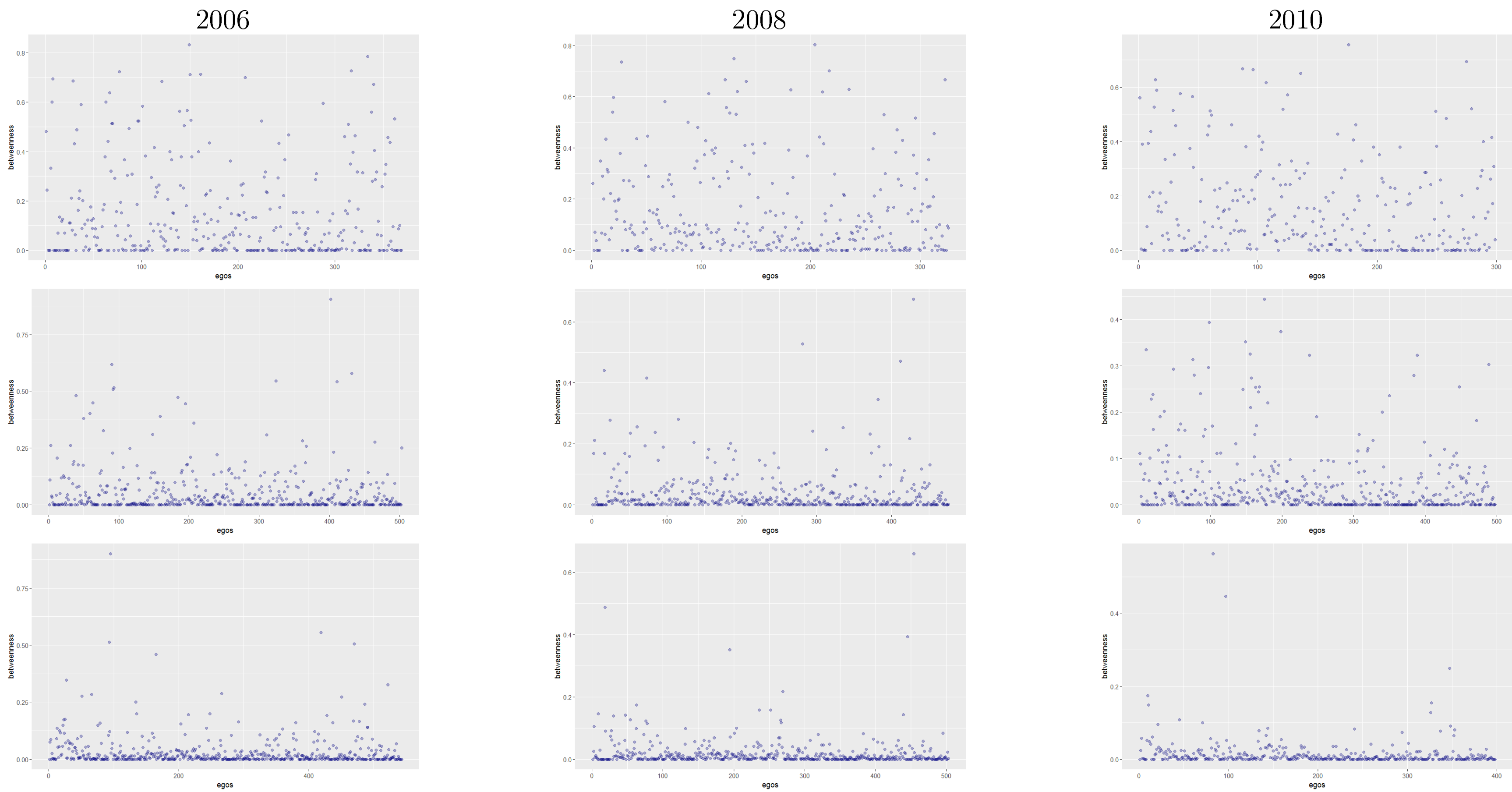


Figure 5.27: Ego betweenness: predicted OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

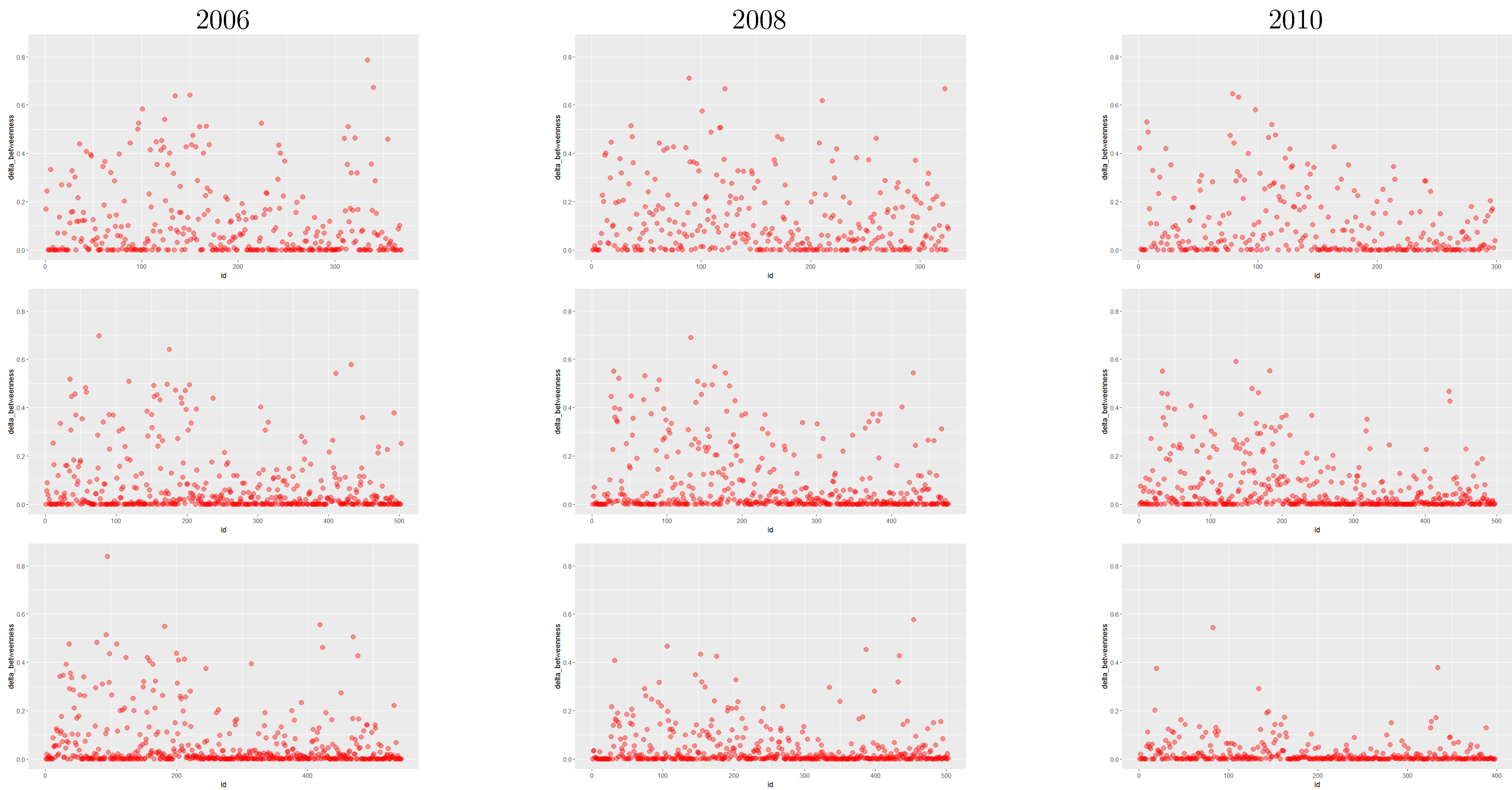


Figure 5.28: Ego betweenness: difference in absolute value between predicted and real OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

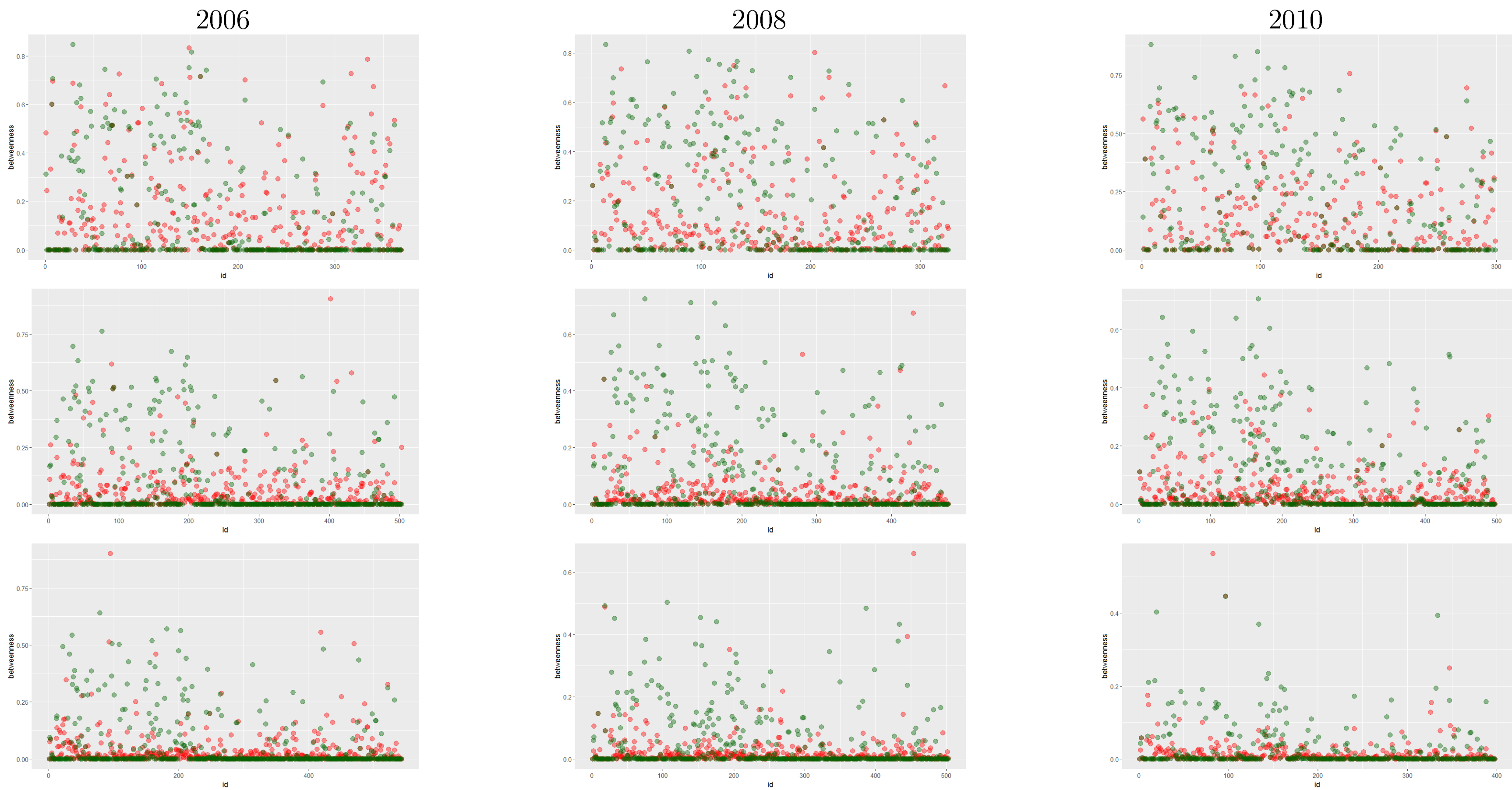
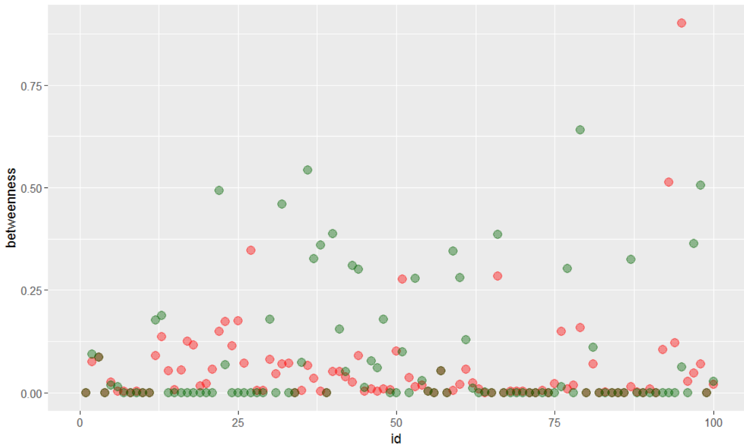
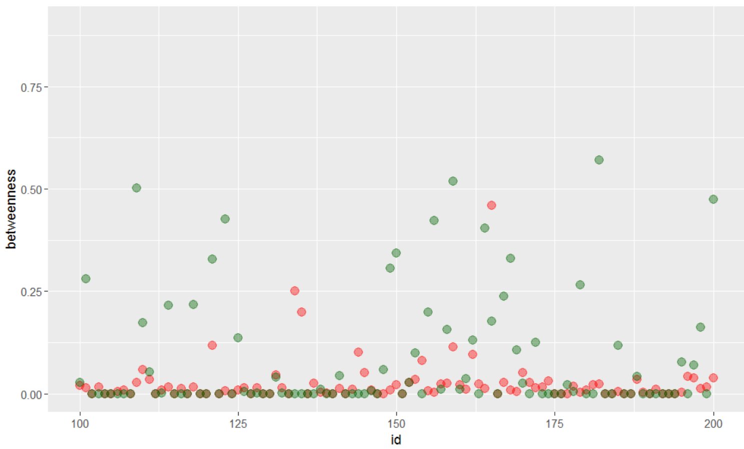


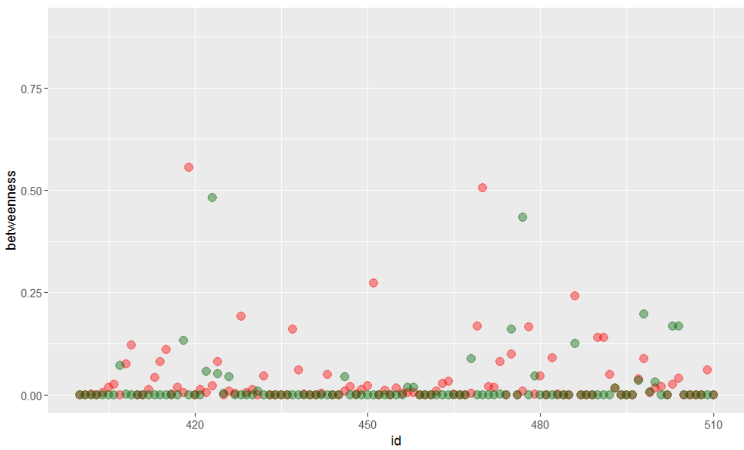
Figure 5.29: Ego betweenness: predicted (red dots) and real (green dots) OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).



(a) Zoom on OPNs 1-100.



(b) Zoom on OPNs 100-200.



(c) Zoom on OPNs 200-530.

Figure 5.30: Betweenness: predicted (red dots) and real (green dots) OPNs, for $k = 4$ and $year = 2006$.

Power law degree distribution. We tested if the degree distribution of the predicted networks follows a power law. The results of the test are given in Figure 5.31. From the figure we can easily confirm that the degree distribution of the predicted networks follows a power law for $k = 2, 3$, and 4. However, we can notice that the proportion of predicted personal networks that do not follow a power law degree distribution is higher for $k = 4$. These observations join what we observed for real OPNs where the personal networks at $k = 2$ and 3 were following a power law degree distribution, while at $k = 4$, the proportion of personal networks not following a power law degree distribution was higher.

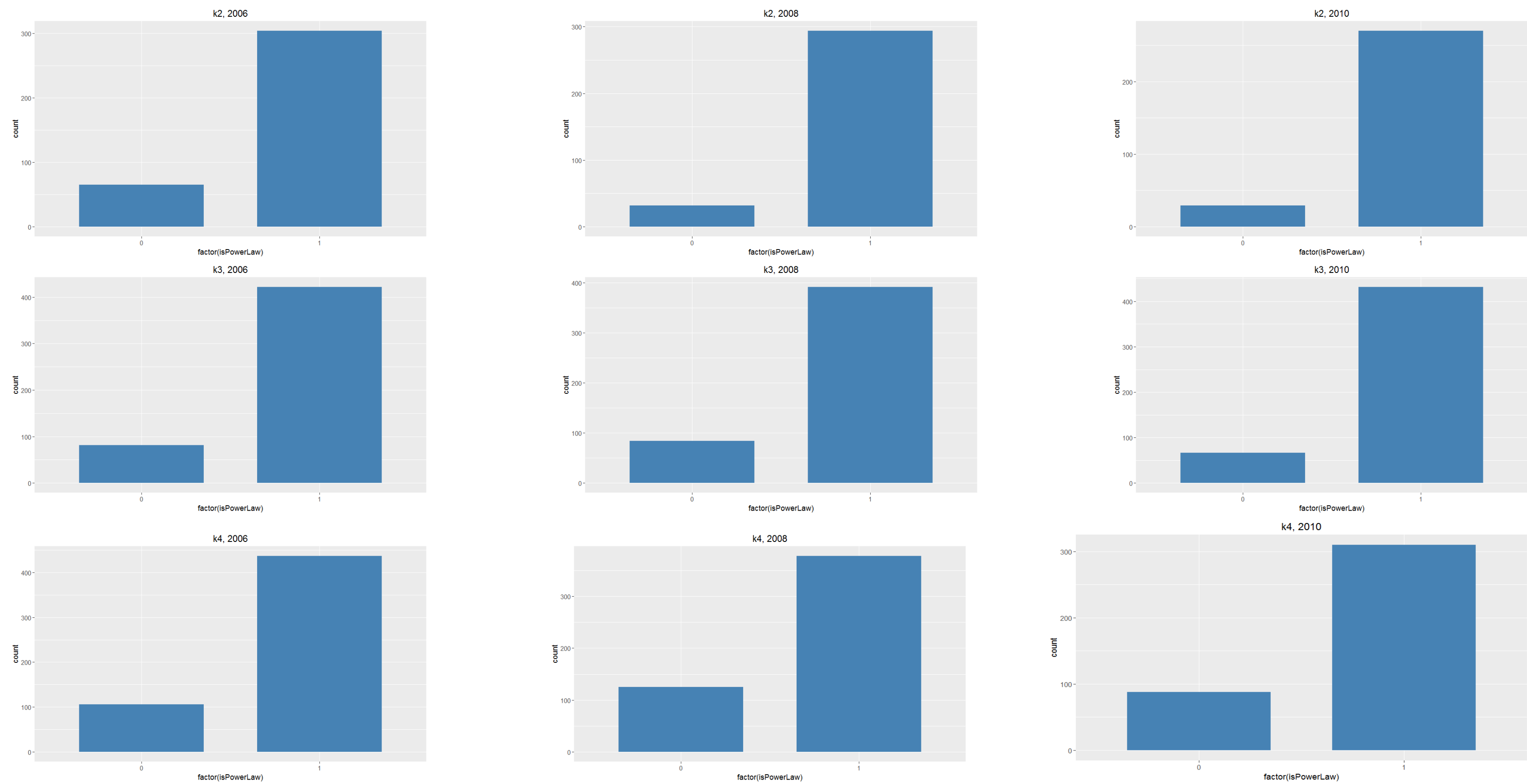


Figure 5.31: Power law degree distribution: test on the predicted OPNs.

Global clustering coefficient. Following the same methodology as previously, we study the global clustering coefficient of the of the predicted OPNs comparing to the real OPNs. We analyze the global clustering coefficient via three phenomena: evolution tendency, difference in absolute value, and value for predicted vs. real OPNs, as describe hereafter:

1. GCC evolution trend: From Figure 5.33, it is not obvious from this graphic if the GCC is decreasing (as for real data) or increasing over the years.

   However, using the Wilcoxon test, we could find that the global clustering coefficient of the predicted personal networks is decreasing over the years whatever k as it was the case for real personal networks. Indeed, the results of the Wilcoxon test given in Table 5.7 allows to conclude that the tendency of decreasing is verified since $p-value_{decreasing}$ is lower than 0.05 (so we reject the null hypothesis), and $p-value_{increasing}$ equals or approaches 1 indicating that the evolution tendency cannot be the increase of the global clustering value over years.

Year	2006-2008	2008-2010
k=2	V =21354, $p-value_{decreasing}$ = 1.73e-08, $p-value_{increasing}$ =1	V = 12406, $p-value_{decreasing}$ = 2.311e-05, $p-value_{increasing}$= 1
k=3	V = 65598 , $p-value_{decreasing}$ = 3.826e-11, $p-value_{increasing}$ =1	V = 48822 , $p-value_{decreasing}$ = 0.08006, $p-value_{increasing}$= 0.92
k=4	V = 78931 , $p-value_{decreasing}$= 1.031e-08, $p-value_{increasing}$ =1	V = 34817 , $p-value_{decreasing}$ =0.03886, $p-value_{increasing}$ = 0.9612

Table 5.7: Wilcoxon test on the global clustering coefficient of predicted OPNs, $dataset_1$

2. GCC difference in absolute value: Figure 5.34 shows that in most of time, the difference between predicted and real OPNs regarding the GCC is low and is rarely above 0.2.

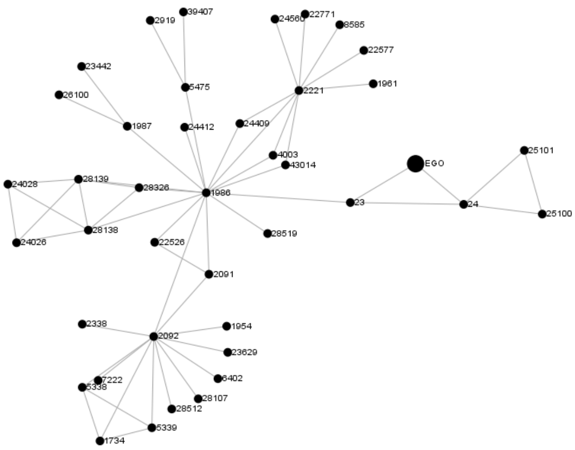
3. GCC value in predicted vs. real OPNs: Figure 5.35 shows that the GCC of the predicted networks is lower than real networks' GCC especially for $k = 4$.

   In order to validate the previous conclusion, we take the graphic for the OPNs with $k = 4$ in 2006 and zoom on the 100 first OPNs, the 100 next OPNs, and the rest of OPNs as done previously. The resulting graphics are given in Figures 5.36a, 5.36b, and 5.36c respectively. From these graphics, we can easily confirm that the GCC of predicted OPNs is lower than the GCC of real OPNs, However, by looking at the networks one by one, we can see that the difference is rarely significant. Indeed, we have:

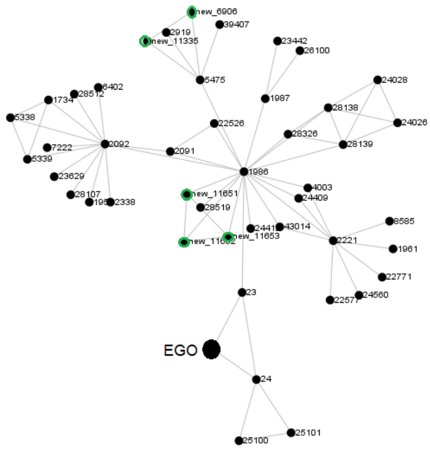
- 44.19% of cases where the difference is less that 0.1,
- 44.56% of cases where it is between 0.1 and 0.2, and
- 11.23% of cases only where it is above 0.2.

Next, we would like to explain the previous conclusion: the PERSON*EM* evolution model predicts networks with a lower GCC. In the first phase of the PERSON*EM* model, in the edge evolution step a node is selected in order to be connected to a (sub)-clique; when we select the size of the clique from the power law distribution to retrieve with q-NuMVC (q), we observed that it corresponds very often to the values 3 or 4. Moreover, the two p_{sample} parameters promote the connection of the node to a sub-clique, rather to an entire clique. Thus, very often, the model will connect the selected node to one single node of the sub-clique which means that we disfavor the creation of triangles and augment the number of connected triples; thus, the GCC gets lower.

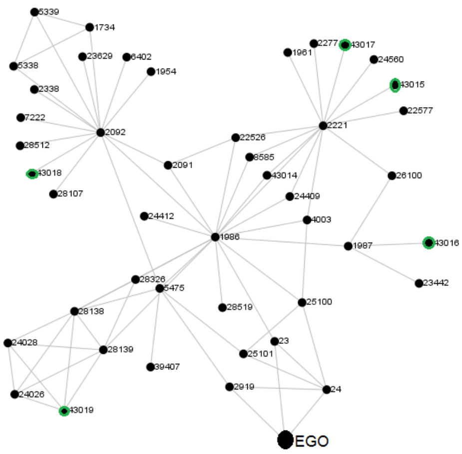
We give an example to illustrate this argumentation. Let us consider an OPN in 2006 with $k = 4$ to which 5 new nodes need to be added in 2007. We give in Figures 5.32a, 5.32b, and 5.32c the visual representation of the personal network of the author with the $id = 22$ (the ego) corresponding to the real OPN in 2006, the real OPN in 2007, and respectively the predicted OPN in 2007. The new added nodes in 2007 are surrounded by a green circle in the figures. From Figure 5.32c, we can see that 4 among the 5 new nodes where added to the OPN via a single connection forming a 2-sized clique with an old node of the OPN and only one new node (43019) created multiple connections (4 connections) and formed a 5-clique with the old nodes of the OPN. This node contributes in increasing the number of triangles and thus the GCC of the network, while the four other added nodes will increase the number of connected triple and thus decrease the GCC of the network.



(a) Real OPN in 2006.



(b) Real OPN in 2007.



(c) Predicted OPN in 2007.

Figure 5.32: Example of the OPN of ego 22, for $k = 4$ and $year = 2006$.

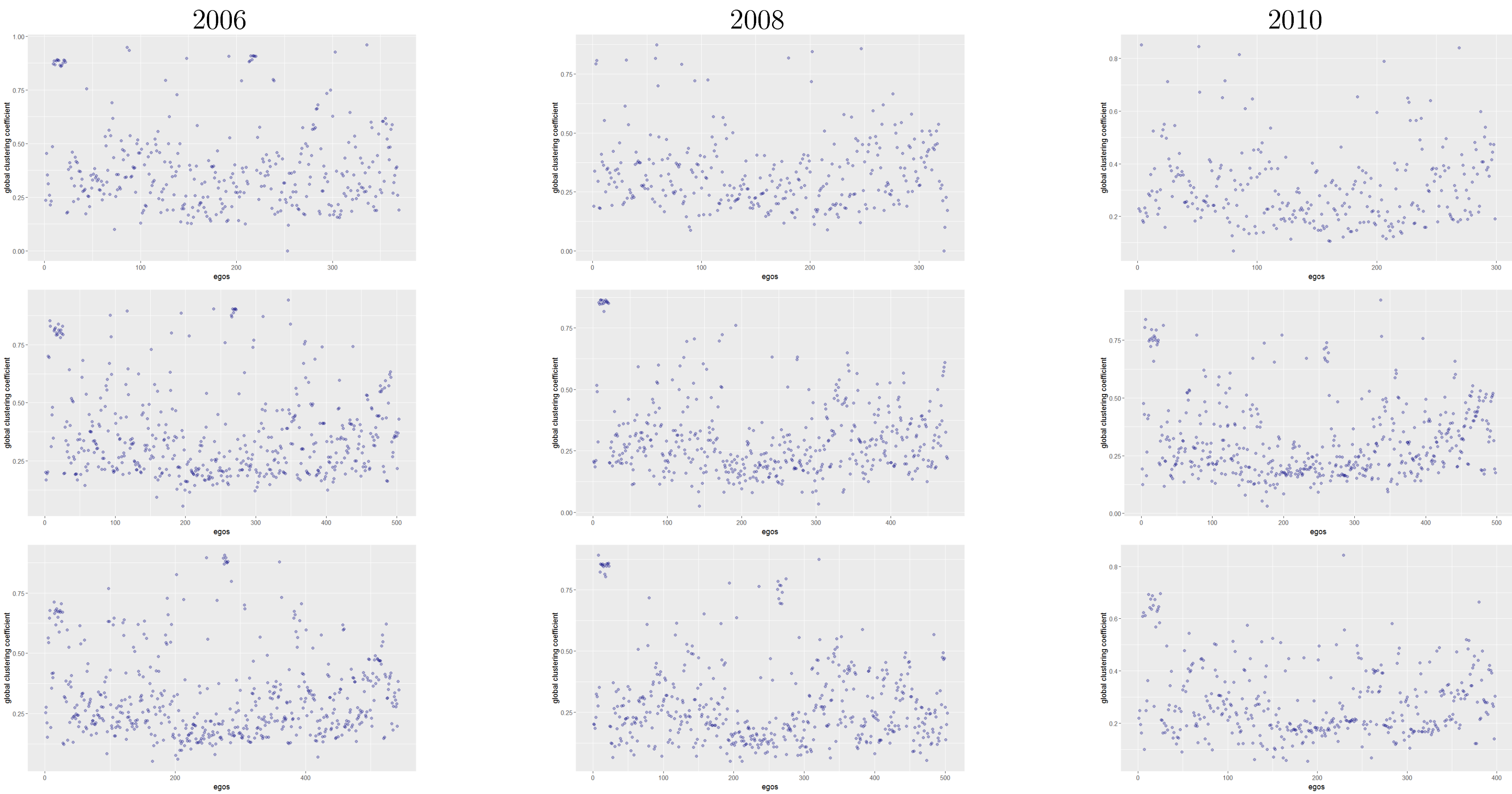


Figure 5.33: GCC: predicted OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

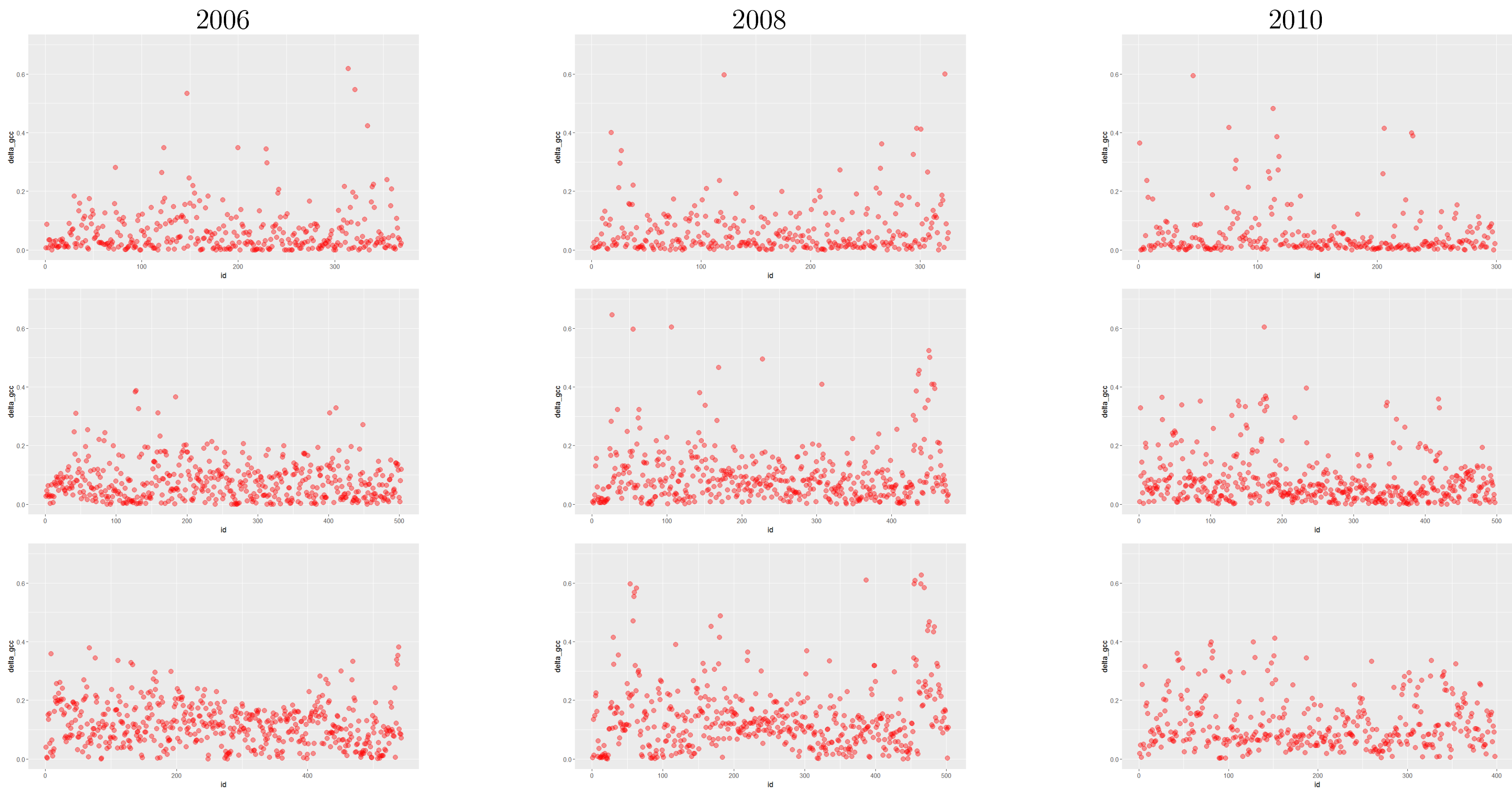


Figure 5.34: GCC: difference in absolute value between predicted and real OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

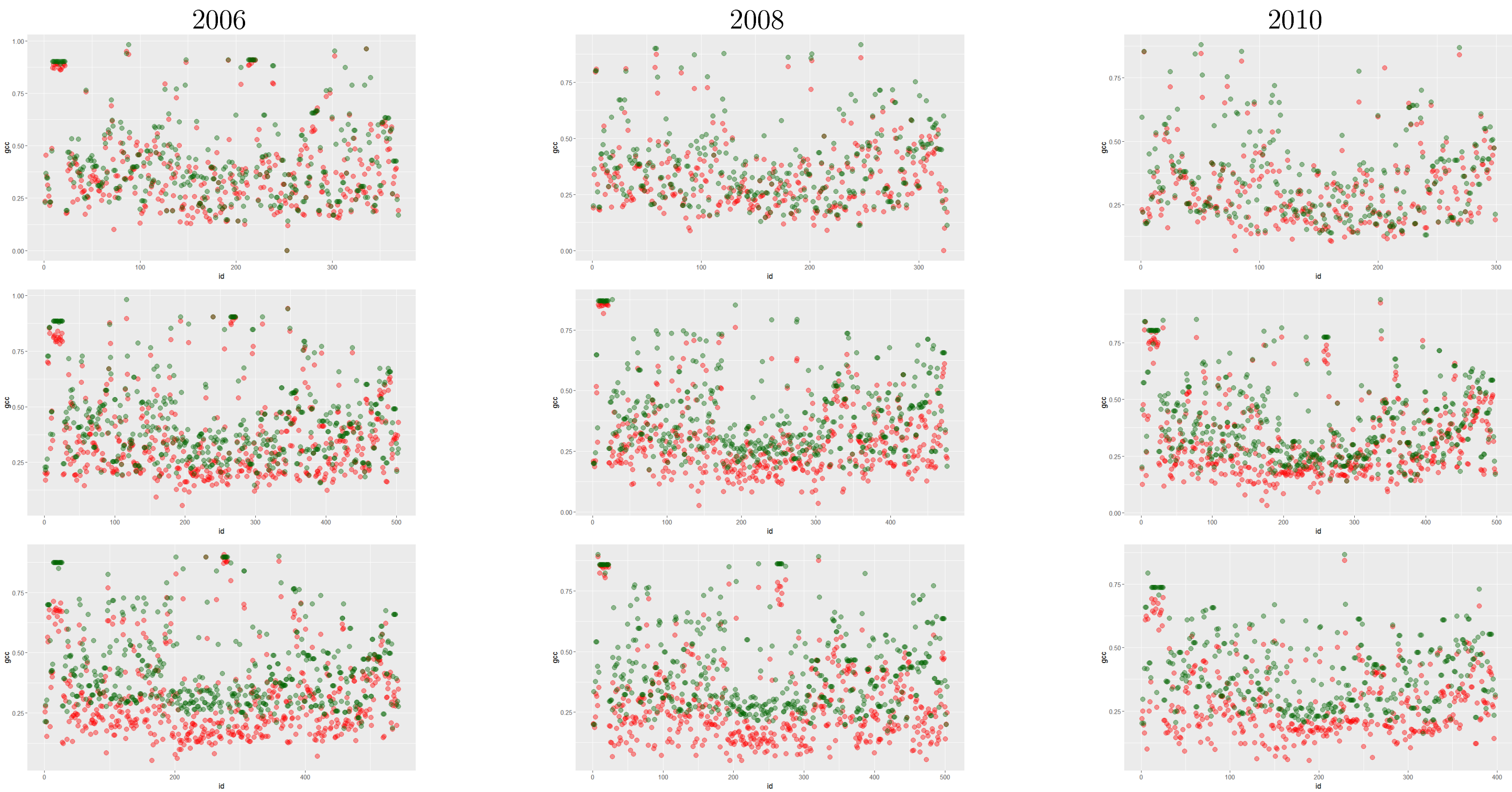
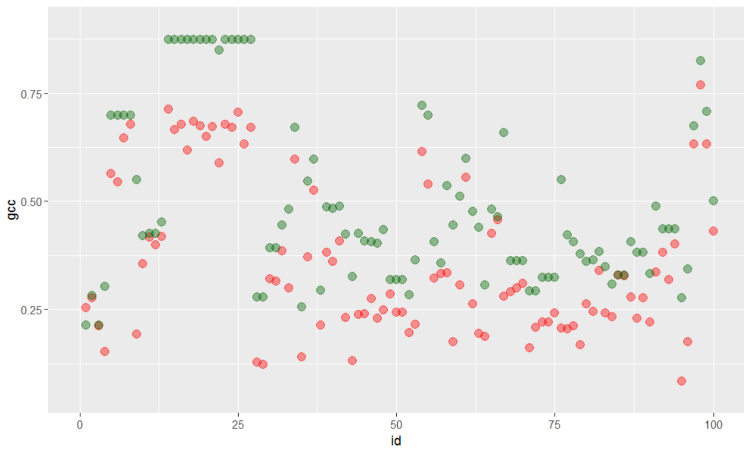
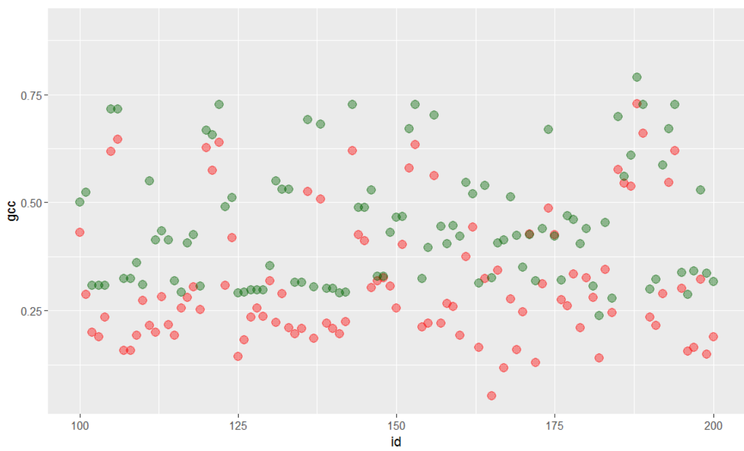


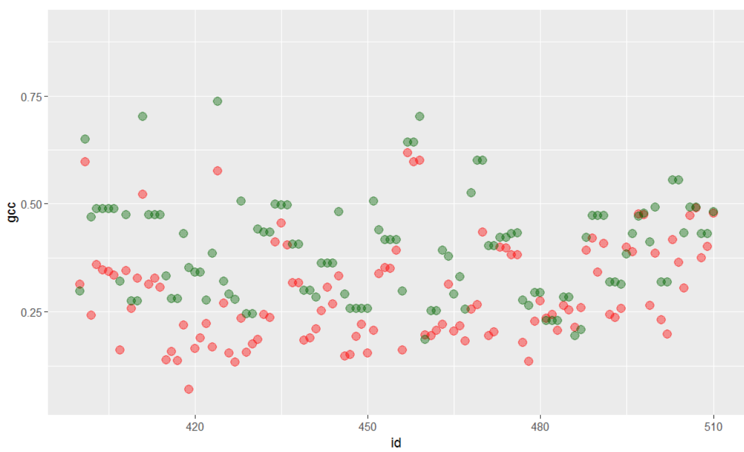
Figure 5.35: GCC: predicted (red dots) and real (green dots) OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).



(a) Zoom on OPNs 1-100.



(b) Zoom on OPNs 100-200.



(c) ZOom on OPNs 200-530.

Figure 5.36: GCC: zoom on predicted (red dots) and real (green dots) OPNs, for $k = 4$ and $year = 2006$.

Average local clustering coefficient. Following the same methodology as previously, we study the average local clustering coefficient of the of the predicted OPNs comparing to the real OPNs. We analyze the average local clustering coefficient via three phenomena: evolution tendency, difference in absolute value, and value for predicted vs. real OPNs, as describe hereafter:

1. ACC evolution tendency: From Figure 5.37, we can observe that the ACC is high as for real OPNs. However, it has not clear if it has a tendency of increasing over the years as for real OPNs.

   With the Wilcoxon test, we are able to clarify more the evolution tendency of the average clustering coefficient. Indeed, the results of the test given in Table 5.8 indicate that the average clustering coefficient increases over years for $k = 2$ since $p - value_{increasing}$ is less than 0.05, while $p - value_{decreasing}$ equals 1.

   For $k = 2, 3$, and 4, the average clustering coefficient increases over years from 2008 to 2010. But, we cannot keep this affirmation for predicted personal networks evolving from 2006 to 2008 since $p - value_{increasing}$ is higher than 0.05 (very sightly for $k = 3$ where $p-value_{increasing} = 0.05209$), even if the decreasing tendency cannot hold because $p - value_{decreasing}$ is close to 1.

Year	2006-2008	2008-2010
$k = 2$	$V = 10650$, $p - value_{decreasing} = 1$, $p - value_{increasing} = 3.484e - 05$	$V = 5696$, $p - value_{decreasing} = 1$, $p - value_{increasing} = 3.249e - 07$
$k = 3$	$V = 44998$, $p - value_{decreasing} = 0.948$, $p - value_{increasing} = 0.05209$	$V = 27598$, $p - value_{decreasing} = 1$, $p - value_{increasing} = 1.026e - 12$
$k = 4$	$V = 57628$, $p - value_{decreasing} = 0.8652$, $p - value_{increasing} = 0.1348$	$V = 25124$, $p - value_{decreasing} = 0.9998$, $p - value_{increasing} = 0.0002335$

Table 5.8: Wilcoxon test on the average clustering coefficient of predicted OPNs, $dataset_1$

2. ACC difference of absolute values: Figure 5.38 shows however a large difference between the predicted and real value of the ACC. If we take the example of $k = 4$ in 2006, we have 37.93% of cases where the difference is less than 0.2, 56.53% of cases where the difference is between 0.2 and 0.4, and 5.52% of cases where the difference is above 0.4. So, the larger proportion confirms a large difference of the ACC between the predicted and real OPNs.

3. ACC value in predicted vs real OPNs: from Figure 5.39, we can easily distinguish that the model predicts OPNs with a lower average ACC comparing to real OPNs. This observation is in line with the observation about the lower GCC of the predicted OPNs. Indeed, the fact that the model generates less triangle (as illustrated in Figure 5.32c) means also that the produced networks are locally less clustered in average.

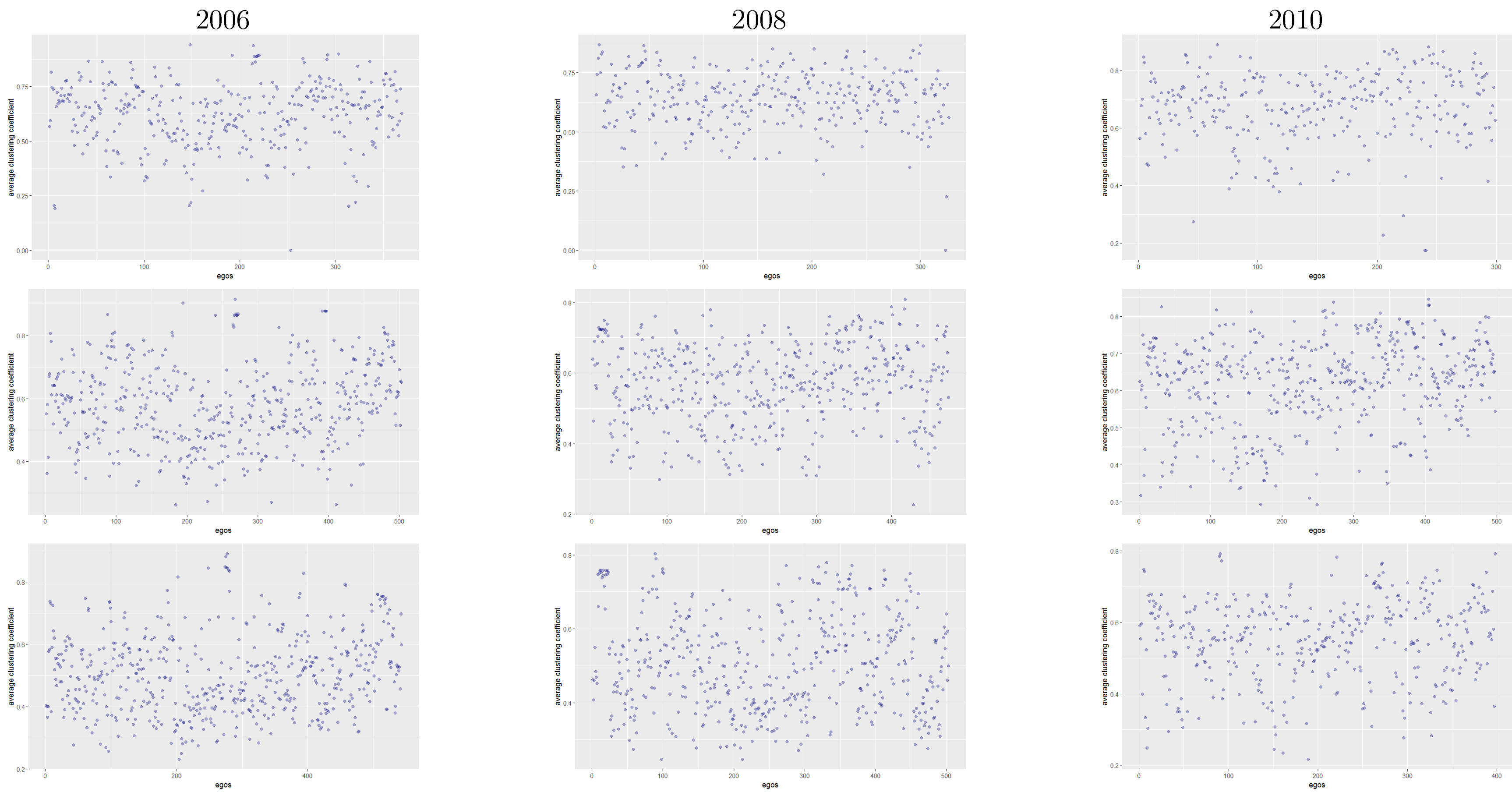


Figure 5.37: Average LCC: predicted OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

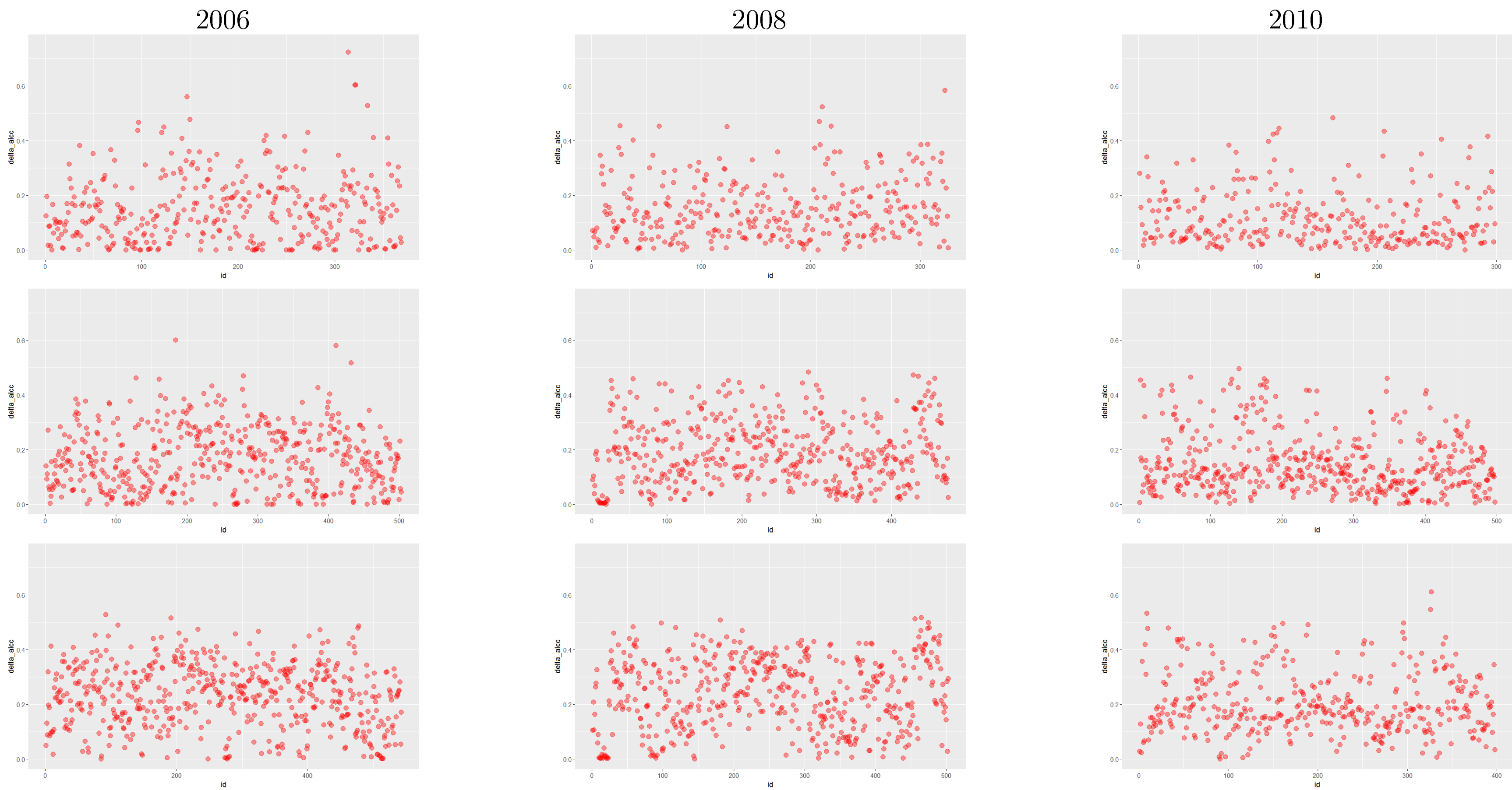


Figure 5.38: Average LCC: difference in absolute value between predicted and real OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

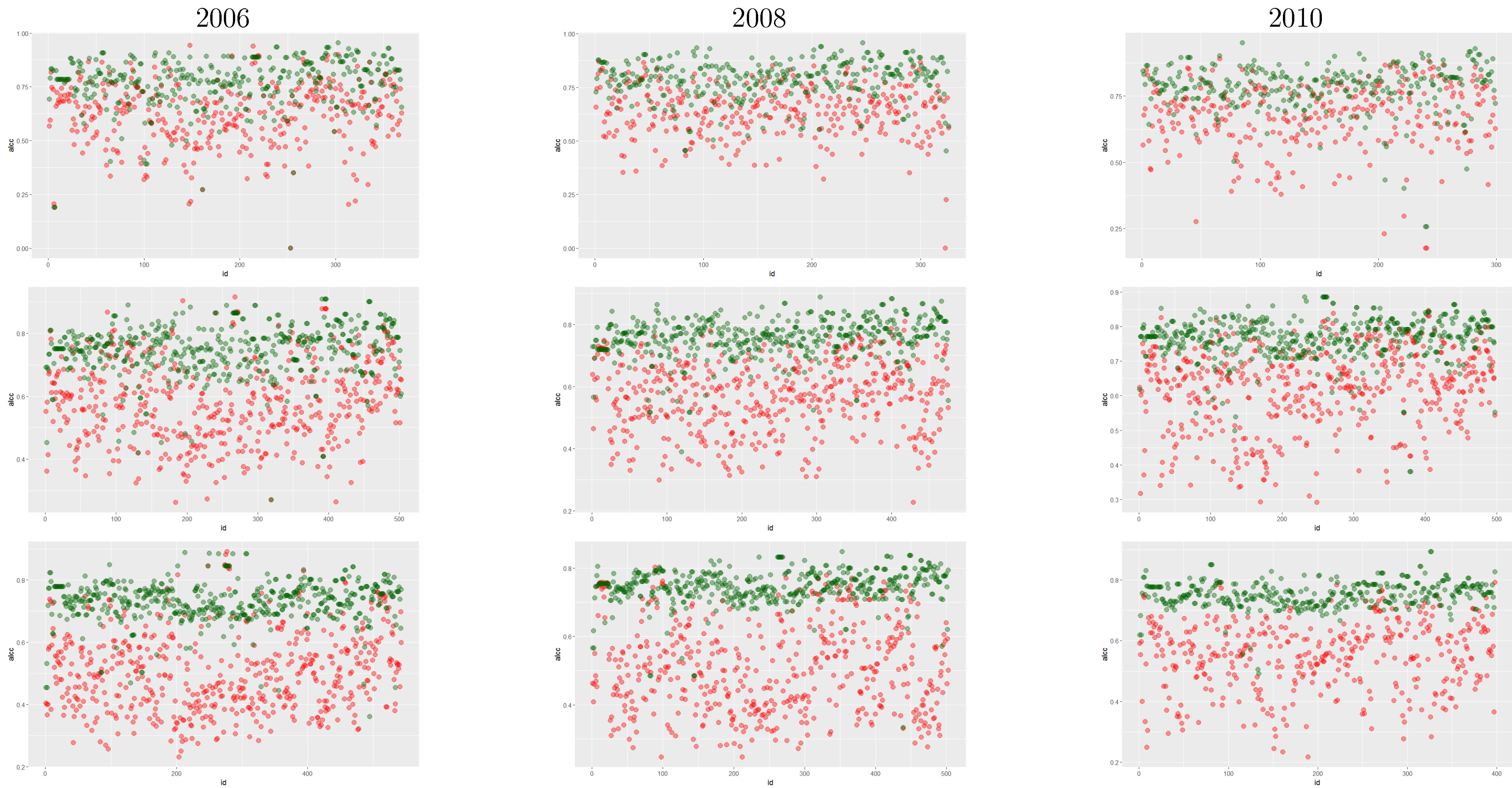


Figure 5.39: Average LCC: predicted (red dots) and real (green dots) OPNs, for $k = 2$ (1st row), $k = 3$ (2nd row), and $k = 4$ (3rd row).

5.4.4 Discussion

In the previous sections, we discussed the results of application of PERSON*EM* for the prediction of the evolution of real personal networks of authors of scientific publications. First, we discussed the results of application of PERSON*EM* on 1-personal networks ($k = 1$), and concluded that for this particular type of OPNs it was difficult to draw conclusions given that most of 1-OPNs of the dataset were complete (more than 80%), rendering not possible the application of the model, and we could test the model only on a small set. For this later set, we found that the model assigns less edges than needed but the density remains close to the one of real OPNs as well as the global clustering coefficient and the ego betweenness. In additon, we found that the predicted OPNs have a bit lower average degree and a lower average clustering coefficient.

Nevertheless, for personal networks with $k = 2, 3, 4$, we were able to prove that PERSON*EM* constitute a good model for predicting the evolution of co-authors personal networks given the results we obtained. Indeed, even if we have notice that the predicted OPNs by PERSON*EM* have a little less edges than expected, the structure of the predicted OPNs maintains the same characteristics as the one of the real OPNs. More precisely, density's values of the predicted OPNs are corresponding to its values in real OPNs and it is decreasing over years as well. The average degree per OPN for the predicted is almost similar to the average degree of the real OPNs and its evolution tendency is the same, i.e, increasing over years. The degree distribution of the predicted networks follows a power law. However, we can notice that the proportion of predicted personal networks that do not follow a power law degree distribution is higher for k = 4. This joins what we observed for real OPNs.

At the ego level, we found that ego degree of the predicted OPNs fits its value in real OPNs, and that there is a small difference for the ego betweenness between real and predicted networks. When the ego betweenness is not maintained by the model, it is a bit higher. The evolution trend of the betweenness for k =2 and 3 was not increasing as for real data between 2006 and 2008, but it was between 2008 and 2010. While for $k = 4$, the betweenness decreases as in real data.

Finally, concerning the transitivity inside the predicted OPNs, we found that PERSON*EM* predicts OPNs with lower GCC and ACC. While the difference is not very significant for the GCC, it is more important for the ACC (even if its value is high as in real OPNs). About the evolution of these two metrics, on the one hand, the Wilcoxon test allowed to conclude that the GCC of the predicted personal networks is decreasing over years whatever k as it was the case for real personal networks. On the other hand, with the Wilcoxon test, we found that the ACC increases for $k = 2$, but for $k = 3$, and 4, it increases only between 2008 and 2010. We explain this by the fact that in general, there are less nodes that are added to the OPNs between 2006 and 2008 than between 2008 and 2010.

However, the fact that the predicted personal networks have a lower ACC constitute a limitation of PERSON*EM*. As mentioned earlier, the model generates less triangle (as illustrated in Figure 1.32c) which means also that the produced networks are locally less clustered in average. This comes from the fact that very often, the model will connect the selected node to one single node of the sub-clique which means that we disfavor the creation of triangles and augment the number of connected triples; thus, the GCC gets lower but also the ACC. A lower p_{sample} parameter could overcome this situation. We could also modify the β parameter, corresponding to the power law exponent parameter in order to search for cliques of higher size.

5.5 PERSON*EM* model's performance

For the current discussion, we allowed our-self to provide again, in Algorithm 7, the PERSON*EM* evolution model that we proposed.

The part of the PERSON*EM* model that affects the most its performance consists in the $q-NuMVC$ function allowing to find cliques in a graph. Indeed, as mentioned in the introduction of this chapter, finding cliques in a graph is an NP-complete problem and constitute a real constraint especially when the graphs have an important size. Thus, given that PERSON*EM* is using $q - NuMVC$ for finding random cliques of size q, we could confirm that it becomes time consuming as the personal networks graphs given as input have an important size in terms of nodes and edges number, but also as the number of new nodes to add becomes high.

Moreover, it is important to note that in PERSON*EM* we call $q-NuMVC$ many times. In the following, we provide the different steps of the algorithm that require performing $q - NuMVC$:

- In each edges evolution step, the $q - NuMVC$ is called in order to provide a clique of size q (line 11 in Algorithm 7).
- Then, in the same edges evolution step, after the first call of $q - NuMVC$, the provided clique is tested for validity. If the clique is not valid, the $q - NuMVC$ is called until a valid clique is found (while loop from line 10 to 28 in Algorithm 7).

- Once the size of N nodes of the predicted personal network is reached, a set of additional edge evolution steps are performed for connecting the eventual disconnected nodes (line 31 in Algorithm 7). In this steps, the $q - NuMVC$ is called.

When performing $q - NuMVV$, we observed that the majority of running time is used in computing vertices *dscore* during the initialization part (line 2 of Algorithm 5), and updating vertices *dscores* once we modify uncovered edges weights (line 15 of Algorithm 5).

Besides the performance of the model related to the $q - NuMVC$ algorithm, we can also note that the model runs the step of nodes evolution and the step of edges evolution (with the probability p_{node}) till the predicted network has the required N nodes. Thus, we can not provide an exact number of executions of each step, but we know that the nodes evolution step is executed with a probability of 40%, while the edges evolution step is executed with a probability of 60%. Eventually, the nodes evolution step will be executed a maximum of N times, if the initial graph does not contain any nodes. Given the probability of 40%-60% between the steps, the edges evolution step should be executed a maximum of $1.5 * N$ times. In addition, the edges evolution step will be also executed in the $connectIsolatedNodes(G)$ function a number of time equal to the number of isolated nodes in the graph and that the maximum can be N. Thus, we can conclude that:

- the nodes evolution step is executed a maximum of N times;
- the edges evolution step is executed a maximum of $2.5 * N$.

In Figure 5.40, we provide the graphics describing the running time of PERSON*EM* as a function of the number of nodes (Figure 5.40b), the number of edges (Figure 5.40a), and the number of new nodes (Figure 5.40c) of the predicted personal networks with $k = 3$ on 2010.

From these graphics we can observe that the running time is related to the size of the personal networks in terms of number of nodes and edges, but also to the number of new nodes that need to be added. Indeed, the running time is increasing as personal networks have more nodes and edges and as the number of new nodes is growing. For a better visualization of such linear relationship, we focus on the graphic in Figure 5.40b on personal networks having a number of edges less than 1500 and a running time less than 9000 seconds (about 91% of the predicted networks). Then on personal networks having a number of edges less than 1000 and a running time less than 2500 seconds (about 80% of the predicted networks). The corresponding plots are given in Figure 5.41a and Figure 5.41b respectively. These figures show that the prediction of the majority of the predicted networks with PERSON*EM* is performed in a time less than 2000 seconds (about half an hour) for personal network whose edges number approaches 1000 edges.

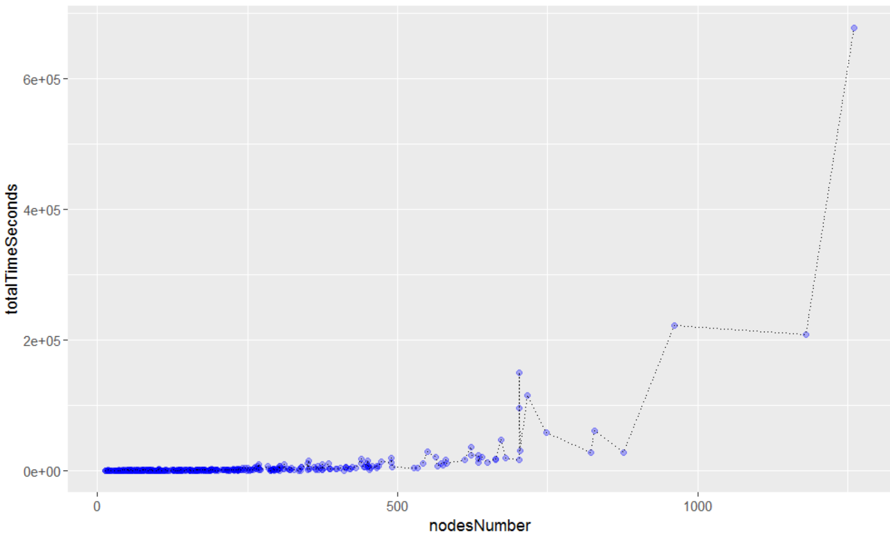
It is important to notice that $q - NuMVC$ performance is compatible with the original $NuMVC$ since we could observe a similar running time for networks of a same size.

Algorithm 7: PERSON*EM* (PERSonal cO-authoring social Networks Evolution Model) evolution model.

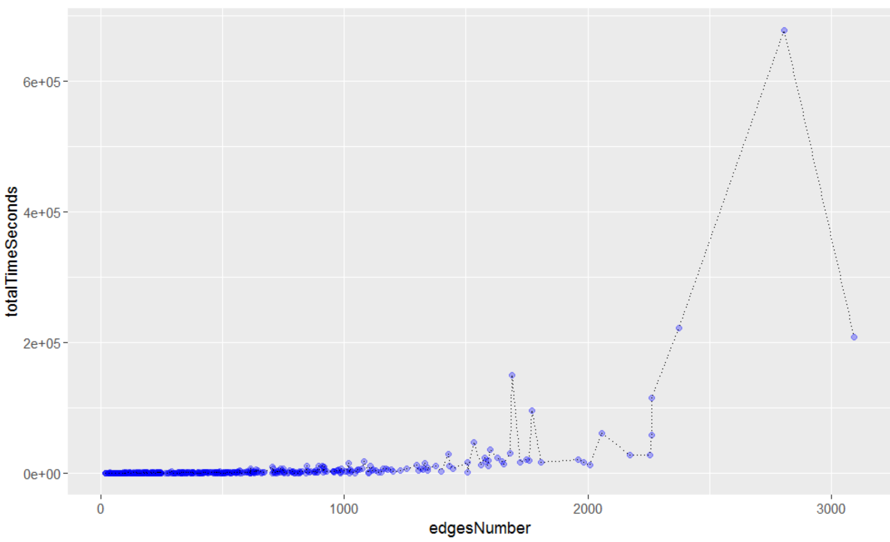
Input : G_{init}, k, p_{node}, $p_{sampleNew}$, $p_{sampleOld}$, N, β

Output: G

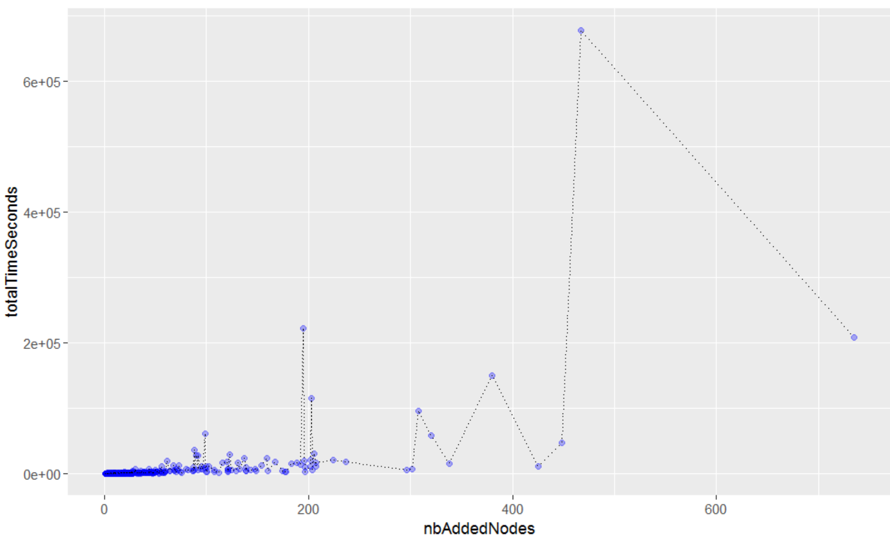
```
begin
Graph G ← newGraph();
while G.getNodes() < N do
    if Random() < p_node then
        Node current ← newNode();
        G.addNode(current);
    else
        Node newMember ← G.getRandomNode();
        q ← SamplePowerlaw(β);
        while ! valid_q_clique do
            Clique C ← V/q − NuMVC(Ḡ, |V| − q);
            if newNodes.contains(newMember) then
                p_sample = p_sampleNew;
            else
                p_sample = p_sampleOld;
            end
            if C ≠ null then
                if Random < p_sample then
                    Clique C′ ← SampleValidSubClique(C);
                else
                    Clique C′ ← C;
                end
                for i = 1, ..., Size(C′) do
                    G.addEdge(newMember, C′.node[i]);
                    // if the edge already existing, the edge weight plus 1
                end
            end
        end
    end
end
connectIsolatedNodes(G);
Return G;
end
```



(a) PERSONEM running time in second vs number of nodes.

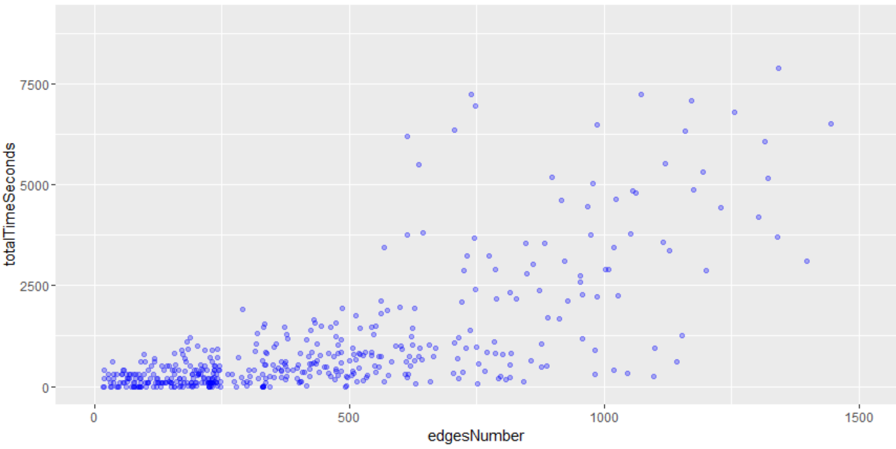


(b) PERSONEM running time in second vs number of edges.

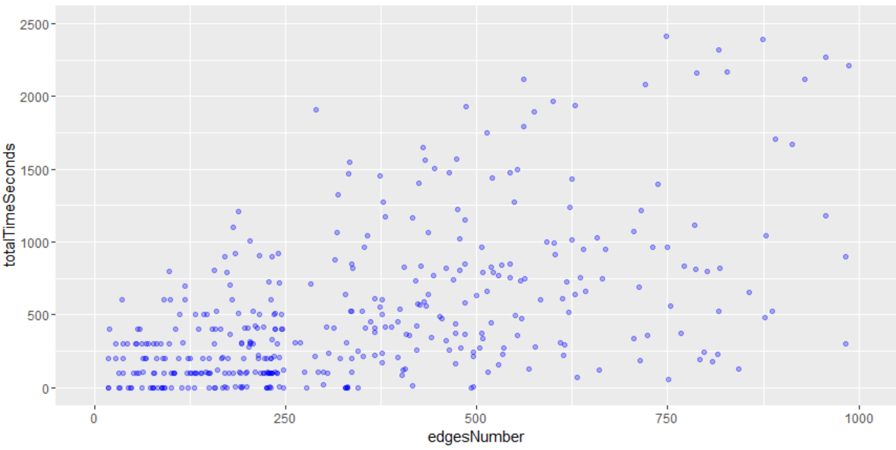


(c) PERSONEM running time in second vs number of new nodes.

Figure 5.40: PERSONEM running time vs number of nodes (a), number of edges (b), and number of new nodes (c) for predicted OPNs, for $k = 3$ and $year = 2010$.



(a) PERSONEM running time limited to OPNs with running time of 9000s and 1500 edges.



(b) PERSONEM running time limited to OPNs with running time of 2500s and 1000 edges.

Figure 5.41: PERSONEM running time vs number of edges for predicted OPNs, for $k = 3$ and $year = 2010$.

5.6 Conclusion

In this chapter, we presented PERSONEM, a new evolution model for personal co-authoring networks evolution based on the clique dimension that constitute a key substructure characterizing co-authorship personal networks' structure as concluded in Chapter 4. PERSONEM was presented in [41].

PERSONEM was designed starting from the clique superposition generative model proposed by Yan et al. in [134]. PERSONEM predicts a personal network at time $t + 1$ starting from the personal network at time t. It is based on the principle that a node (new arriving one or an old one) will connect to the network by the mean of connections to cliques. The new evolution model is based on two evolution steps: the nodes evolution step allowing to add a node to the network, and the edges evolution step allowing to connect a node (new or old) to a clique. PERSON*EM* uses $q-NuMVC$ algorithm (instead of the KEWLS algorithm used in the clique superposition model), a vertex cover finding algorithm that we adapted to the case of personal networks that allows to find a random clique of a given size.

We applied PERSON*EM* on a set of co-authoring personal networks in order to predict their evolution. These networks have a size that can reach two thousands of nodes, and more than 4500 edges.

The obtained results show that PERSON*EM* predicts personal networks with $k = 2, 3, 4$ having a structure which resembles the one of real personal networks. In addition, the predicted personal networks exhibit the same evolution trends we observed on real co-authoring personal networks. Thus, we believe that PERSON*EM* is a solution to problem of prediction of the evolution of personal networks such as co-authorship networks. However, some limitations were identified. The most important one is the fact that the predicted personal networks are locally less clustered in average comparing to real personal networks.

To overcome these limitations, several running of the model with a variation of the parameters are needed to evaluate the prediction and find mode appropriate parameters. In this work, the performance constraints prevented us to perform more runnings. Indeed, clique finding problem (or vertex cover finding problem) is time consuming when the networks are of bigger size. A more performing algorithm adapted to personal networks could improve the running time of PERSON*EM*, and so, would allow us to adjust the model's parameters for a better prediction efficiency.

Chapter 6

The Software Framework *PERSONA*

6.1 Introduction

In the previous Chapters, we built a large-scale experiment for analyzing the evolution of OPNs. To do so, we needed to compute a set of metrics and perform some algorithms on a large number of OPNs for several years. To this end, we first needed to retrieve a large set of personal networks from the whole social networks of authors that started publishing on a given year and we needed an automatic way for this task.

Existing softwares dealing with personal networks are not numerous and limited to extract personal networks fitting the traditional definition, i.e. the one including only the ego and its immediate alters at distance one or two plus the relationships among them.

Thus, we developed a software framework named PERSONA (PERSonal Online social Networks' Analytics) that allows us to extract OPNs from any networks and based on any of the definitions presented previously in Chapter 3. PERSONA allows to compute a set of metrics and perform algorithms on these OPNs. The development of the tool allowed us firstly to validate the definitions and secondly to perform the different analysis presented in Chapter 4. With PERSONA, we offer a versatile tool to anyone who wants to perform research around OPNs.

6.2 Existing tools for personal networks analysis

Most often, tools for analyzing networks as graphs offer only few functionalities for personal networks which are limited to highlighting the ego and its immediate connections (1-personal networks) with displaying basic statistics as the number of alters, the number of edges between alters, as it is the case of *Gephi* [16].

Some dedicated tools to personal networks exist. These tools are more evolution visualization oriented. For example, in [113] the authors propose techniques to visualize large scale personal networks evolution by considering the data as continuous streams. Furthermore, the visualization software *EgoLines*, presented in [136], proposes a dynamic analysis of personal networks. The system *EgoSlider* [132] allows to analyze a set of dynamic personal networks by summarizing their properties at three levels: a macroscopic level that concerns the entire personal network data, a mesoscopic level to capture specific individuals' personal network evolution, and a microscopic level providing detailed temporal information about the egos and their alters.

These tools allow to isolate a personal network and analyze its evolution with a visual support, but they are lacking the capabilities of performing massive scale analysis. Moreover, the integrated metrics (e.g. number of alters, edges between alters, etc.) are limited and provides only generic information about the OPN. Another important point is the fact that the considered personal networks are those of level 1 or two (1-personal networks or 2-personal networks) at most. But, in our case, we wanted to be able to extract personal networks following the definitions we proposed in Chapter 3. The only tool we found that allows to retrieve a personal network with specifying the radius (k) is *EgoNav* [71], a visual analytics system that allows to summarize a collection of personal networks in addition to the possibility of highlighting the network of a specific user in the whole network to allow the visualization and the analysis of its personal network properties. But, this tool focuses on the visualization and does not hold a lot of analyzing capabilities.

Given the limits of existing tools, we developed PERSONA software that allows:

- to connect to any database storing the whole network
- to extract personal networks according to the definitions proposed in Chapter 3
- to perform computations on OPNs via metrics and algorithms
- to visualize personal networks
- to export personal networks into a given graph format (GraphML, GML, DOT, CSV)

The software is presented in details in the rest of the chapter.

6.3 The choice of Technical Solutions

The framework PERSONA was implemented using the Java programming language under the development framework Eclipse. Choosing Java as programming language was decided by the numerous APIs available for graph representing and processing. We offer two versions, one as a standalone desktop tool and another as a set of web services that allows us to manipulate OPNs over the web and build the functionality into other applications and tools.

In the following, we present the main library *JGraphT.* that we used to deal with graph structure and the additional libraries we needed to perform more computations and algorithms missing in *JGraphT* .

6.3.1 JGraphT Library

JGraphT [1] is a Java library of graph theory data structures and algorithms which has the advantages of being flexible, powerful and efficient. Indeed, it is flexible because it allows to define nodes and edges with any object type. It enables having undirected edges as well as directed and weighted edges within a simple graph, multigraph, or a pseudograph. *JGraphT* is powerful as it implements specialized iterators for graph traversal (DFS, BFS, etc) and multiple algorithms for clique detection, path finding, coloring, centrality and the list goes on. It also offers exporters and importers for popular external representations such as GraphViz [2]. Finally, *JGraphT* is an efficient library because of its optimization of memory and its speed of performance.

Thus, these qualities motivated our choice for this library. In addition, a lot of libraries were built on top of *JGraphT* and thus brought the missing functionalities which render *JGraphT* a library that is quite complete when dealing with graphs and answers different needs. Another important advantage is that *JGraphT* is continuously updated and new versions are published.

In PERSONA software we use the version 1.1.0 of *JGraphT*. Due to all the advantages mentioned earlier about this library, we could define our personal networks sub-graphs with an appropriate data structure. The sub-graphs representing the personal networks were defined under the $SimpleWeightedGraph < V, E >$ class for the case of an undirected personal network The parameters V and E represent respectively the graph vertex type and the graph edge type which are user-defined types. In

[1] https://jgrapht.org/
[2] https://www.graphviz.org/

our case, the vertices are of integer type (*ids*) and the edges are of the type *Default-WeightedEdge*, a predefined type available with *JGraphT* for representing weighted edges such that the extremities of an edge are of integer type. When querying the database for the personal network of a given ego, the obtained result consisting in a set of nodes and edges is integrated to the data structure by using the methods *addVertex(v)* to add a vertex v and *addEdge(u,v)* to add an edge between nodes u and v. The edge weight w is set using *setEdgeWeight(e, w)* for a given edge e.

In addition to using the data structure offered with *JGraphT*, we could perform many computations thanks to the integrated metrics and algorithms.

6.3.2 Additional Libraries

In the following, we give a description of the libraries we required in addition to *JGraphT* to show the computational needs we had to integrate to PERSONA for implementing metrics and evolution models.

uncommons-maths. The Uncommons Maths library provides high-performance pseudorandom number generators (RNGs). Using the included probability distribution wrappers, these RNGs (and the standard JDK ones) can be used to generate values from Uniform, Normal, Binomial, Poisson and Exponential distributions.

jdistlib. A Java package that provides routines for various statistical distributions. It offers for instance the computation of the density (pdf), cumulative (cdf), quantile, and random variates of many popular statistical distributions (Chi square, Exponential, Normal, Poisson, etc), normality tests such as Kolmogorov-Smirnov and Anderson-Darling tests.

sorend-jgrapht-sna. This library implements several algorithms used in social network analysis (SNA) on top of the JGraphT library. The focus for this library was mainly on the centrality measures it implements as degree centrality, closeness centrality, betweenness centralty, eigenvector centrality, etc, and the improved support for GraphML import/export (weighted and directed) it includes.

jgrapht-metrics. Like the previous library, this one offers various metrics for JGraphT graphs like the average degree, the average weighted degree, the average neighbour degree, the average path length, the local clustering coefficient and the assortativity coefficient.

powerlaws. This is a java library for the analysis of power law distributed data. The methods implemented are all taken form the paper Power-Law Distributions in Empirical Data by Clauset, Shalizi and Newman (2007) [34]. It contains facilities for estimating parameters, uncertainty and significance and for generating power law distributed data. All methods are implemented for continuous data, discrete data and the approximation of discrete data with a continuous distribution.

neo4j-algorithms-master. Graph algorithms are used to compute metrics for graphs, nodes, or relationships. Neo4j Graph Algorithms is a library that provides efficiently implemented common graph algorithms. Many graph algorithms are iterative approaches that frequently traverse the graph for the computation using random walks, breadth-first or depth-first searches, or pattern matching. Due to the exponential growth of possible paths with increasing distance, many of the approaches also have high algorithmic complexity. Fortunately, optimized algorithms exist that utilize certain structures of the graph, memorize already explored parts, and parallelize operations. Neo4j integrates these optimizations. As an example of the algorithms it holds, we cite centrality, community detection, or path finding algorithms.

In the next section, we describe PERSONA system and provide details on its implementation and the different modules it involves.

6.4 PERSONA's design and implementation

PERSONA is a tool dedicated to automatically retrieve and analyze personal networks. It integrates the definitions we proposed presented in Chapter 3 and the different metrics and algorithms discussed in Chapter 4 and Chapter 5. It also integrates the evolution model presented in Chapter 5. With PERSONA, the user can analyze both the extracted OPN and the whole social network in order to compute various network metrics that might be useful in order to better understand the networks and further analyze them. Furthermore, PERSONA can perform a task on a large dataset of OPNs while existing tools focus on one single OPN [20] or only on a few set of OPNs [19].

Figure 6.1 gives an overview of PERSONA system. On a network stored as a graph in a database composed of a node table and an edge table, PERSONA performs the following tasks:

1. retrieve the personal network of a specific network member;
2. convert the personal network into a graph data structure;
3. perform a set of processing on the extracted personal network with a set of computations;
4. instantiate an evolution model and apply it on the personal network;

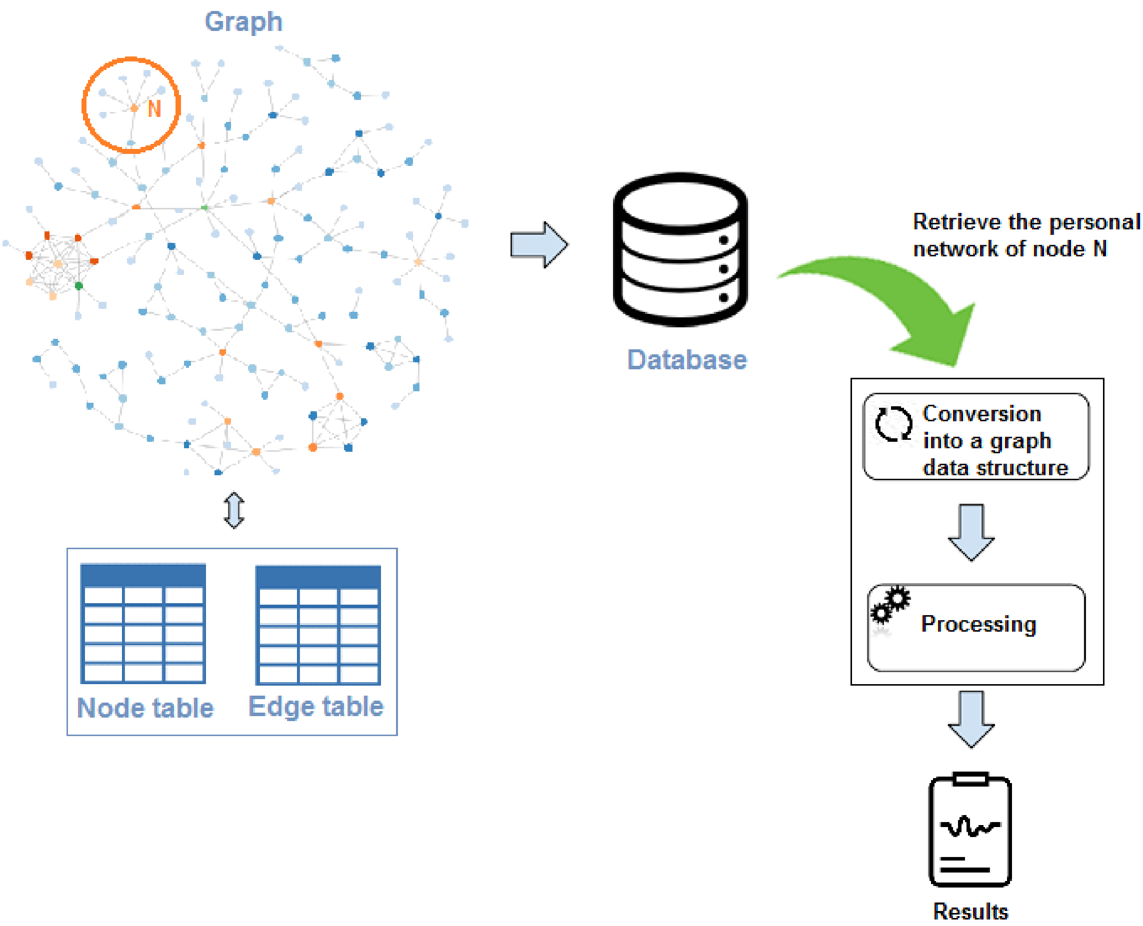


Figure 6.1: PERSONA System.

5. export of the personal network in a given format;

6. provide the visualization of the personal network graph.

6.5 Modules Development

The system presented in the previous section was implemented following the UML diagram given in Figure 6.2. The different modules of the framework will be presented in details in the following.

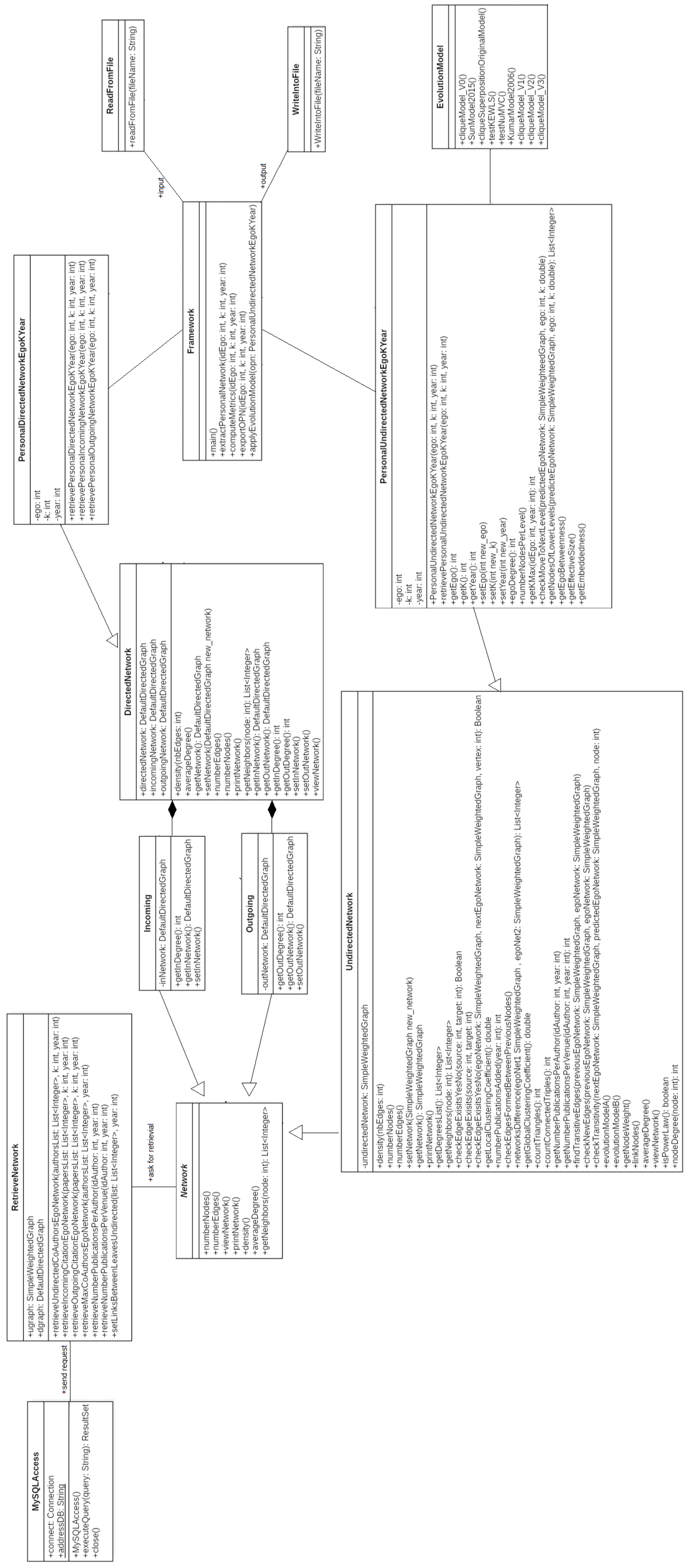


Figure 6.2: PERSONA's UML class diagram.

6.5.1 (Personal) Network Management/Storage Module

The data representing the whole networks is stored in a database. Then, PERSONA is able to access the database in order to extract either the overall network or the personal network of a given node in the network. Depending on the type of the database storing the network, nodes and edges between nodes have to be represented in the database in a form that makes easy the access to them.

In its current implementation, PERSONA uses a relational database to store the nodes and the edges in a set of tables, but different databases can be imagined with limited modification in the implementation.

6.5.2 Personal Network Retrieval Module

The retrieval module allows to set the exact configuration for the access to the database storing the network and to specify the parameters of the retrieval. It also has the task to convert the retrieved data into the corresponding data structure.

In Figure 6.3, we provide the structure of the retrieval module. *MySQLAccess* class ensures the connection with the MySQL database, via its address. Then, PERSONA queries the database via an SQL query and gets the answers from the database.

RetrieveNetwork class sends a specific request and takes in charge the answer from the database in order to convert it into the correct data structure. In this sens, the abstract class *Network* is used to contain the retrieved network and offers a set of generic operations, independently from the type of the network, as computing the number of nodes and edges, or the density of the network. This class allows also to visualize the retrieved network.

Depending on the query, the network retrieved can be of two types, as shown in Figure 6.4:

1. Undirected personal network, which are represented by the *UndirectedNetwork* class and implements all the operations applicable on undirected networks.

2. Directed personal network corresponds to the *incoming*, the *outgoing* or to the merge of the incoming and outgoing networks as defined in Chapter 3. *DirectedNetwork* class implements all the functionalities that are proper to directed networks.

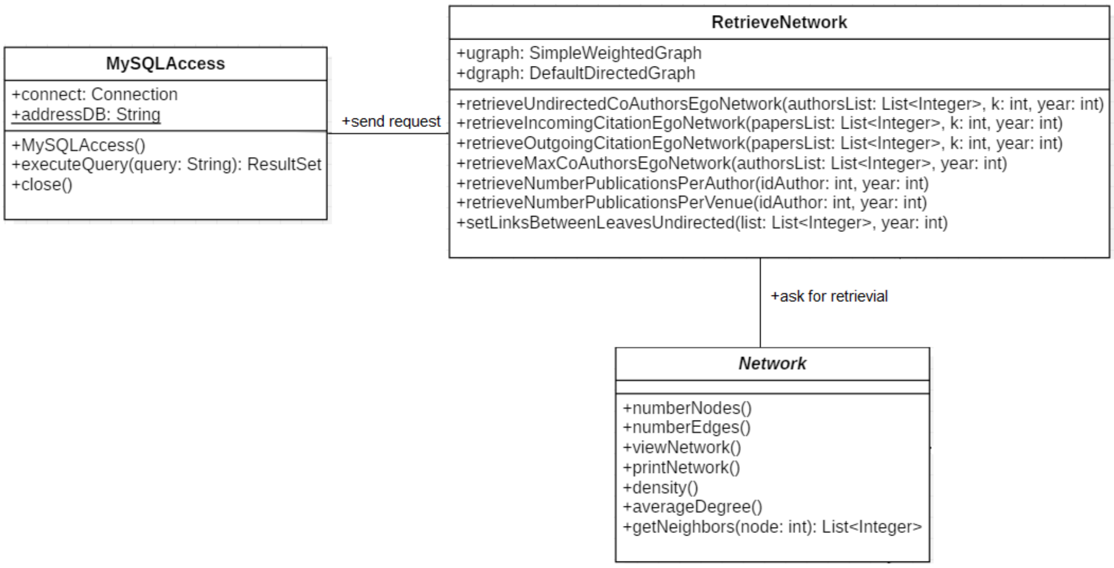


Figure 6.3: Retrieval Module.

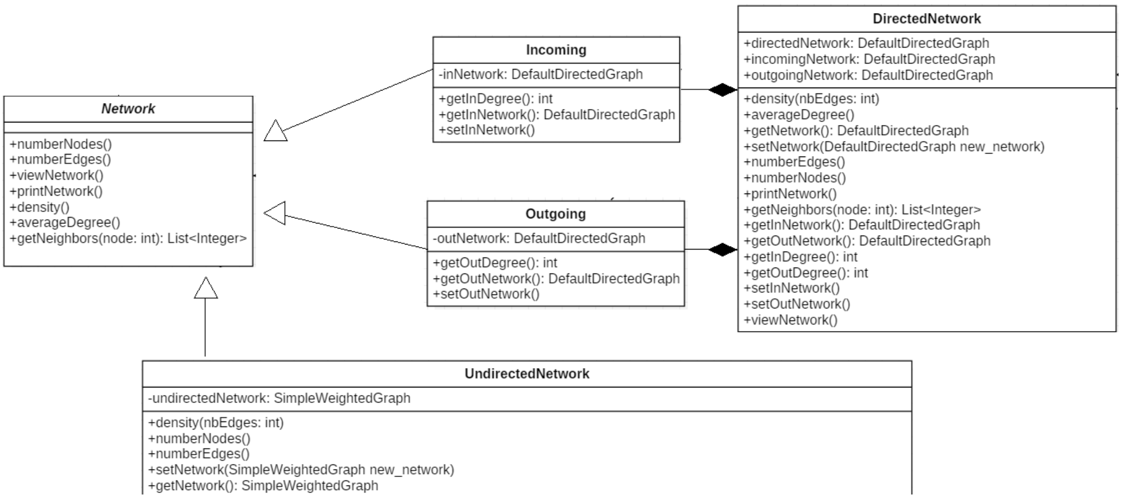


Figure 6.4: Network Type.

6.5.3 Metrics Computation Module

In order to analyze the extracted networks (OPNs or the overall network), we needed to compute a certain number of metrics. A set of them was already defined in the *JGraphT* library, and for the others, we needed either to implement them or to use other libraries. Moreover, some metrics needed to be adapted to ego-networks or to a given task. With PERSONA, we can expand the set of metrics in order to include additional ones and this makes the framework extensible. Figure 6.5 shows that some metrics are generic for all undirected networks, while, there are some that are specific for undirected personal networks. The metrics that PERSONA implements so far on a given network are:

- Number of nodes and edges
- Density
- Degree of nodes
- Average degree
- Local clustering coefficient of nodes
- Average local clustering coefficient
- Global clustering coefficient
- Number of triangles
- Number of connected triples
- Nodes weight
- Check the power law distribution
- Check Poisson distribution
- Betweenness centrality
- Closeness centrality
- Geodesic distance
- kmax
- Effective size

These metrics were already defined in Chapter 2. This list does not include metrics that are proper to the data we used.

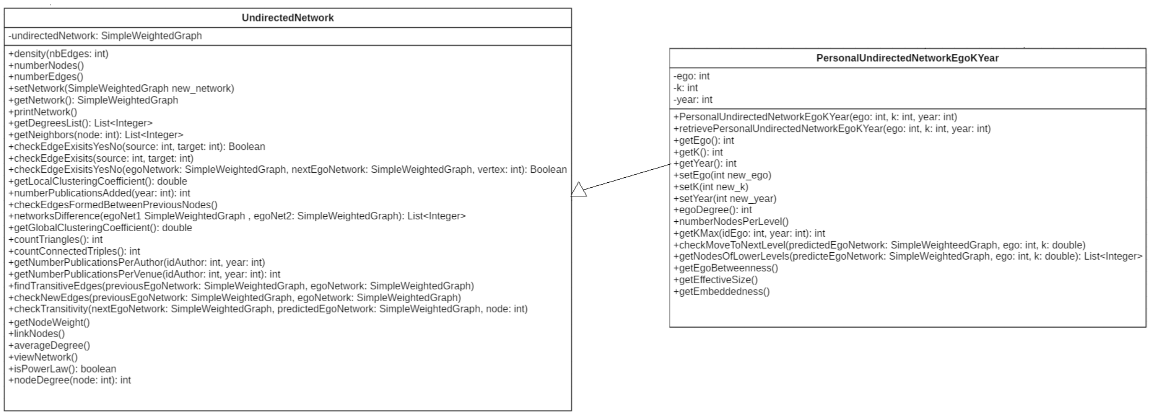


Figure 6.5: Metrics Computation Module.

6.5.4 Evolution Models Management Module

PERSONA integrates a module allowing to apply the evolution models over the personal networks. The class *EvolutionModel* allows to instantiate a given evolution model and to apply it on a given personal network as shown in Figure 6.6. More precisely, the class takes an original network (composed of nodes and edges) and returns a new network that was changed by following the evolution model that was applied on the original network.

To this end, a set of evolution models existing in the literature were implemented as the model proposed by [118], and the one proposed by [78], but also to the clique model we proposed in Chapter 5 and all the different versions of this later that we tested. In addition to the algorithms used by the model (KEWLS, NuMVC) and to the original clique superposition model proposed by [134].

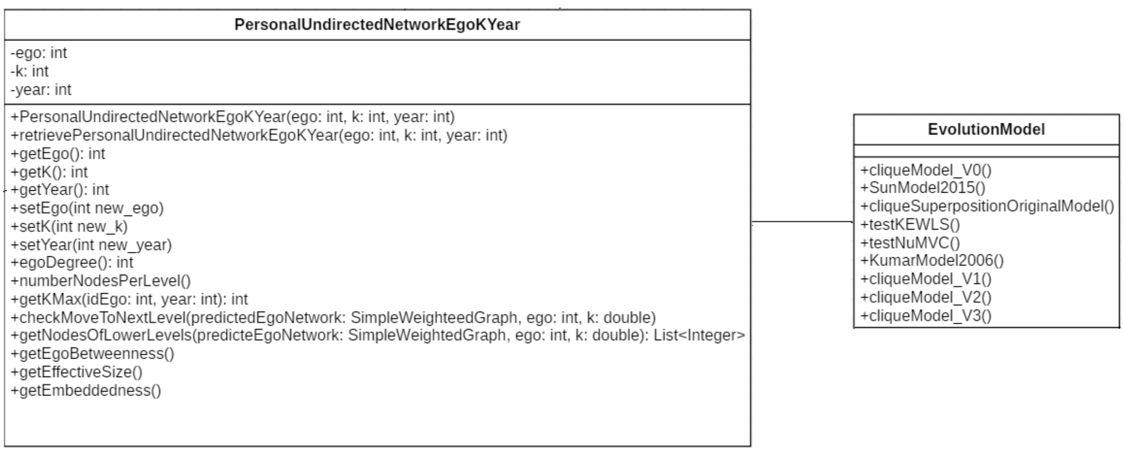


Figure 6.6: Evolution Models Module.

6.5.5 Network visualisation Module

When analyzing personal networks, we need to visualize them in order to explain some special cases regarding a given metric for instance. To this end, a visualization tool is required in order to observe how nodes and edges composing the networks are organized.

In PERSONA, we made the choice of using *GraphStream*[3] library to manage networks' visualization. *GraphStream* is a java library for graph allowing modeling and processing in addition to visualizing graphs. Moreover, it allows the modeling of the dynamics of graphs as stream of graph events. *GraphStream* provides a set of properties to display graphs and the elements of the displayed graphs can be customized. Although *GraphStream* was not built on the top of *JGraphT* since it defines its proper way for representing graphs, it allows to store any kind of data attribute on the graph elements: numbers, strings, or any object.

We precise that in the PERSONA framework we only exploit the visualization functionality of *GraphStream* in order to visualize the OPNs and we use a standard dispaly comparing to the multiple possibilities of customizing graphs with *Graph-Stream*.

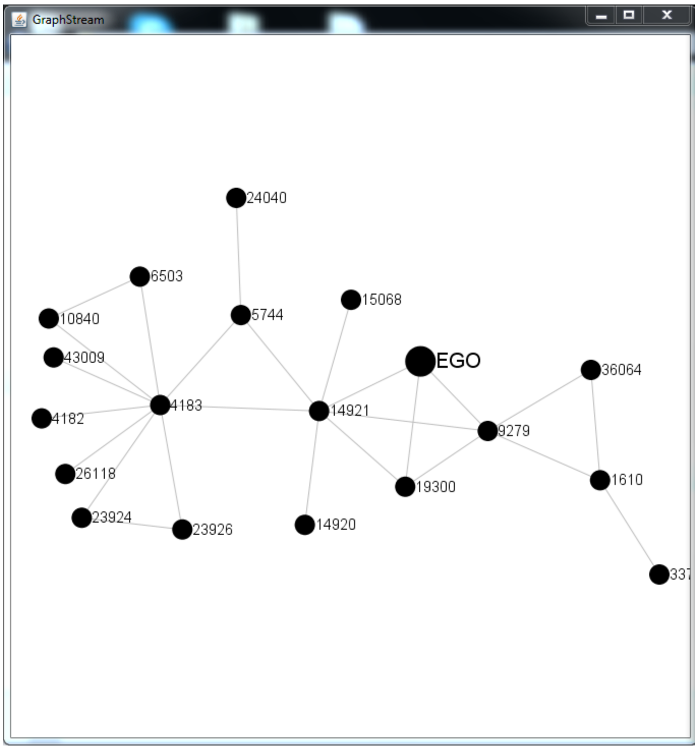


Figure 6.7: Example of an OPN visualization with *GraphStream* library.

[3]http://graphstream-project.org/

The method we use is *Graph.display()* which is a shortcut that creates a viewer. This method by default will try to place the nodes automatically in space to make the graph as readable as possible. In Figure 6.7, we give an example of an OPN visualization using *GraphStream* library.

6.5.6 Personal Network Exportation Module

After the extraction of an OPN, in addition to the analyzing and visualization functionality provided by PERSONA, we offer the ability to export it in a set of different formats. Exporting an OPN allows to save that personal network graph to avoid retrieving it from the database a second time for example. It also allows to use it with another software following the extraction format. We thus render the extracted OPNs usable for another purpose and even built a new graph-formatted database. Among the possible exporting graph formats we cite GraphML, GML, DOT, CSV, etc. The exportation module is given in Figure 6.8.

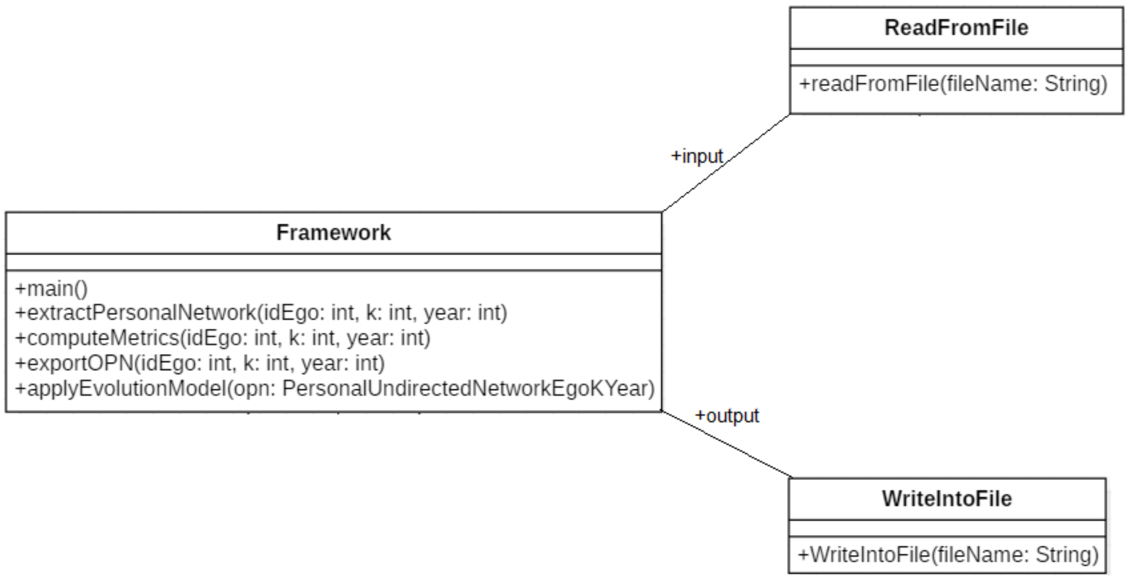


Figure 6.8: Exportation Module.

6.6 PERSONA framework's usage with DBLP data

6.6.1 Global network storage

The framework can connect to any kind of relational database, it can retrieve the necessary data, independently of the kind of the network or the rest of its characteristics. In our case, the overall network of collaboration in the field of *computer networks* was stored within a MYSQL relational database that we named *dblp*. In order to keep all the information about this network, we needed to create the three main tables allowing us to represent the graph, and a set of five additional tables helping us with the computations. These tables described in the following.

The main tables are:

- **authors (nameAuthor, idAuthor).** In this table we associate to each author name a unique identifier. The identifier (*id*) is used to retrieve, from the database, the personal network of a given author.
- **papers (idPaper, title, yearPaper, venue).** The *papers* This table contains each publication or paper with the corresponding information. In our case we have the title of each paper, the year of its publication, and the venue on which it has been published. A screen-shot of this table is given Figure 6.9.
- **author_paper (author_id, paper_id).** This table makes the link between each author and the paper(s) he/she wrote.

idPaper	title	yearPaper	venue
6	Active Database Systems.	1995	Modern Database Systems
44	Indexing Techniques for Object-Oriented Databases...	1989	Object-Oriented Concepts
725	Algebraic Specification.	1990	Handbook of Theoretical Computer Science
1563	Audience-driven Web Design.	2001	Information Modeling in the New Millennium
2079	Attribute-Oriented Induction in data Mining.	1996	Advances in Knowledge Discovery and Data Mining
2196	ACTA: The SAGA Continues	1992	Database Transaction Models for Advanced Applicat...
2204	Facility for Non Standard Database Systems	1992	Database Transaction Models for Advanced Applicat...
2524	Introduction to Information Storage and Retrieval...	1992	Information Retrieval: Data Structures & Algorith...
2564	A Two-Phase Approach to Predictably Scheduling Re...	1996	Performance of Concurrency Control Mechanisms in ...
2952	Run-Time Systems for Coordination.	2001	Coordination of Internet Agents
3143	The MOSIX Distributed Operating System - Load Bal...	1993	Lecture Notes in Computer Science
3213	Network Calculus: A Theory of Deterministic Queui...	2001	Lecture Notes in Computer Science
3498	Petri Nets: An Introduction	1985	Monographs in Theoretical Computer Science. An EA...
3631	Resource Allocation in Wireless Networks: Theory ...	2006	Lecture Notes in Computer Science
3794	The Datacenter as a Computer: An Introduction to ...	2009	Synthesis Lectures on Computer Architecture

Figure 6.9: Papers table for dblp database.

The additional tables are:

- **authors_apparition (idAuthor, year_apparition).** The table allows us to know the year of apparition of authors, i.e. the year on which each author of the underlined network has published for the first time. We built this table because when studying the evolution, it is important to know when a given author joined the network.

- **authors_component (idAuthor, idComponent).** In the network of collaboration in the *Computer Networks* domain, we detected more than one component (a component is a sub-graph such that it exists a path between every nodes pair belonging to that component) including a giant component. Thus, in order to deal only with those authors that belong to the giant component, we created the table *authors_component* which tells us to which component belong each author of the network.

- **citations (citedPaper, idPaper).** The table citations contains for each paper, the paper(s) that have cited it.

- **coauthors (egoAuthor, coAuthor, weight, yearPaper).** This table is the one defining collaboration relationship among network's authors since it records for each author, its coauthor(s). In addition, it records the year of the collaboration (year when both have written a paper together) and also the number of time they collaborate together (number of paper with each other) that can be used for weighting the links between authors.

6.6.2 Connection to the global network

In order to connect to the *dblp* database, we used the JDBC driver for MYSQL [4], version 5.1.38. Once the connection established, the software query the database for retrieving a given information. We mainly query the database via an SQL query for retrieving a personal network of a given author but it is also possible to ask for the entire co-authorship network, or for a specific information as the number of publications of a given author or the maximum number of authors per publication.

In the case of retrieving the personal network of a given author, we need not only to specify its *id* but we need also to specify *k* to get its *k-personal network* and the *year* in order to get its personal network at a given point in time. We give below an example of the query allowing to get a personal network for the DBLP case.

[4]https://dev.mysql.com/

Example 27 *SQL query: select egoAuthor, coAuthor from coauthors where egoAuthor=59 && yearPaper $\leq$ 2010 group by coAuthor*

Output:

egoAuthor	coAuthor
59	56
59	57
59	58
59	7879
59	8240
59	32027
59	33427

In this example, we retrieve from dblp database the set of coauthors of the author having id=59 by querying the coAuthors table which constitute the table storing the edges where each edge is given by the pair (egoAuthor, coAuthor). egoAuthor and coAuthor are the nodes represented in the node table that is authors. The second condition, in addition to ego's id, is to specify the year, for example 2010, which means that we are asking for the personal network of the ego 59 until 2010. The condition regarding k is not part of the query as we can see. Indeed, after retrieving the set of co-authors of the ego author (k=1) we will recursively do the same for each one of its co-authors to get the second level (k=2) and so one until reaching the desired k level.

By using another query, we can retrieve additional properties existing within the database for the nodes (as the names of authors) and, for the edges (as the weights according to Definition 3.5.1).

6.7 Conclusion

In this chapter, we presented the software framework PERSONA for OPNs' extraction, visualization and analysis. With PERSONA, we could validate the definition presented in Chapter 3. This framework was designed in a way to support further extensions as connecting to different types of databases, or supporting additional metrics, and allowing for building on top of it evolutionary predictive models to express the dynamics of OPNs. Thus, PERSONA can be used by anyone wishing to perform analysis and computation on any personal network data.

Chapter 7

Conclusion and Perspectives

Online social networks area is of great interest for the researchers in recent years given the availability of large amounts of users interaction data. Works in this area have mainly analyzed the whole graph representing the users' interaction. More precisely, for many years, scientists have been interested in the dynamics of OSNs and a large set of evolution models was proposed. However, discovering what is happening in an individual's personal online social network through time has not been enough investigated and there are no models devoted to OPNs' evolution.

In this thesis we focused on understanding the dynamics of OSNs at the individual user level, i.e. at the user' OPN level in the aim of proposing new evolution models.

To do so, we first needed to define OPNs since existing definitions have limitations. Thus, we proposed a set of formal definitions for OPNs. The definitions are independent of any application or OSN and capture all diverse existing cases. To the best of our knowledge there is no other systematic and formal recording of these concepts in a way that provides a universal framework for studying OPNs in general.

Then, we performed an analytical study of a large set of personal networks of authors of scientific publications, with the goal to use the results of this study to understand the evolution of the corresponding personal networks. First, we analyzed the way edges are organized among new nodes entering the personal networks at the first level, and nodes that already exist. We found that most of the new edges are formed among new coming nodes which reveals that nodes join the personal network under the form of cliques. Thus, a new collaboration involves most of the time new authors that were not in the network at time t with inevitably the ego author (because we consider 1-level personal networks) and eventually his alters (old nodes) which explains the second important proportion of edges made between new and old nodes. However, we found that it is unlikely for two old authors not connected at time t to connect at time $t+1$ given the very small proportion of links made between old alters. These findings gave us a first a hint about the way edges evolve over time at the first level of the personal networks but was not sufficient to describe the change of the overall structure of these networks.

Thus, in a second stage, we selected a set of metrics that characterizes personal networks' structure and that allowed us to capture the change over time of this structure. We were interested to understand not only the specific values of the metrics but mainly how these values change over time when considering k-personal network with k varying from 1 to 5. These analysis were performed using PERSONA, an extensible software framework for OPNs' extraction, visualization and analysis that we developed. From the metrics analysis, we could discover important properties:

- We first distinguished a clear difference in the topology of 1-level personal networks comparing to upper levels (2 to 5) since 1-level personal networks can have a very specific structure as complete networks or star networks. The evolution of 1-personal networks affects mostly the connections among alters who were characterized at the beginning with an important number of connections and then become less connected over time as new alters join the 1-level personal network (the density, the local clustering and the efficiency decrease), while egos' degrees and egos' betweenness are not affected. Moreover, we found that the addition of nodes (alters) keeps the average degree of the 1-level personal network low.

- k-personal networks (k=2 to 5) are characterized by a decreasing density and transitivity at the personal network level (GCC). While, at the local level, the average clustering coefficient was observed to be high and gets higher when the personal networks are growing. This divergence in the evolution trend between the global clustering coefficient and the average clustering coefficient is a very important finding as it characterizes a particular structure of networks which is the *windmill* graph structure as demonstrated by Estrada in [48]. This demonstrates that our k-personal networks have a structure reassembling the one of the *windmill* graphs.

- Another interesting finding concerns the different evolution behaviour between $k = 2, 3$, and $k = 4, 5$ regarding two metrics, the betweenness centrality and the power law degree distribution. Ego's betweenness centrality was increasing for $k = 2, 3$, but decreases for $k = 4, 5$. On the one side, the increase for $k = 2, 3$ is due to the addition of nodes around the ego that form disconnected groups of alters (as the windmill graph example) and the ego plays a key role among them. On the other side, when we consider a $k = 4$ or 5. This behaviour is justified by the fact that OPNs with $k = 4, 5$ are often characterized by the presence of large cliques as discovered with cliques analysis that we discuss in the following.

The degree power law distribution test results are different depending on k as it is better satisfied when $k = 2, 3$ than for $k = 4, 5$. In addition, for $k = 5$, the networks are not anymore following a power law degree distribution starting from 2010. Again, we explained this changing behaviour when we analyzed the composition of the OPNs when considering cliques, as we found that OPNs with $k = 4, 5$ are often characterized by the presence of large cliques. We discuss cliques analysis in the following.

The last bloc of analysis is based around the composition of personal networks in terms of cliques. Via these analysis, we could effectively justify the behaviour of the metrics evolution trends (betweenness, power law degree distribution) discussed above. It also allowed us to discover interesting characteristics about personal network in terms of their composition of cliques. Indeed, we saw that for a high number of OPNs the minimum size of its cliques is 3, while a high number of OPNs contains cliques with a maximum size of 18 or 20 which constitute the same large cliques appearing in different OPNS as they belong to the giant component. Furthermore, we discovered a linear relationship between the number of cliques and the number of nodes as well as with the number of edges. We also found that, in general, co-authorship OPNs have more small cliques than big ones, and that the number of cliques of a given size is increasing with k and with the years. Finally, by analyzing the connections between new nodes and cliques, we could identify that new nodes tend to connect to sub-cliques rather than to whole cliques and they connect more often to around 30%-50% of the clique nodes, and rarely above 50% of them. At last, new nodes are more tended to connect to few cliques (up to 3) than to a large number of cliques.

The findings from all these analysis outlined above contributed to bringing us a better understanding around the structure and evolution of personal networks. Such a deep investigation has not been carried out before. In addition, the discovered patterns constituted a sufficient baggage to approach modeling the evolution of the studied networks.

To do so, given the dominance of the clique structure in the analyzed personal networks and its role during their evolution, we proposed a new evolution model for personal networks based on cliques. Evolution models in general including the clique dimension are very rare in the existing literature and such a model constitute a different approach for modeling the evolution. The evolution model we proposed, PERSON*EM*, is a predictive model that is based on the clique superposition model proposed by Yan in [134]. PERSON*EM* predicts starting from a personal network at time t the resulting personal network at time $t + 1$. It allows to connect the N new nodes entering the personal network at time $t+1$ with the cliques composing the network. It integrates $k - NuMVC$ algorithm for random k-clique finding that we propose as an extension of $NuMVC$ algorithm of Cai in [25]. PERSON*EM* parameters were fixed from our data and integrates findings discovered via the experimental analysis we performed.

The results of applying PERSON*EM* on our data show good results; indeed, the metrics of the predicted personal networks have, in general, similar values and comparable evolution tendency than the metrics of the real personal networks (with some differences stated in the manuscript); we based our results on the following metrics: number of edges, density, average degree, power law distribution, ego degree and betweenness. However, some lacks were detected concerning the limited expression of the transitivity at the global level and at the local level since we obtain a lower GCC and ACC than in real personal networks. We explain this result by the fact that, during the creation of the edges, PERSON*EM* is less tended to close triangles which induces less transitivity in the resulting network both at the global and local levels. To overcome this lack, multiple runnings of the model with varying the parameters β and p_{sample} could be performed. Another limitation of PERSON*EM* is its low performance in term of execution time; when personal networks grow in size, $k - NuMVC$ might be needed to be run several times. A more efficient algorithm for clique finding problem adapted to personal networks will make PERSON*EM* performing faster.

At last, this thesis is a first effort to understand the characteristics and the evolution of Online Personal Social Networks and to provide algorithms and models for it. While we have focused on online personal collaboration networks with specific characteristics (co-authorships), the results produced could be usable by the greater community.

Bibliography

[1] Alireza Abbasi, Kon Shing Kenneth Chung, and Liaquat Hossain. Egocentric analysis of co-authorship network structure, position and performance. *Information Processing & Management*, 48(4):671–679, 2012.

[2] Charu Aggarwal and Karthik Subbian. Evolutionary network analysis: A survey. *ACM Computing Surveys (CSUR)*, 47(1):10, 2014.

[3] Richard D Alba. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3(1):113–126, 1973.

[4] Reka Albert. Scale-free networks in cell biology. *Journal of cell science*, 118(21):4947–4957, 2005.

[5] Miltiadis Allamanis, Salvatore Scellato, and Cecilia Mascolo. Evolution of a location-based online social network: analysis and models. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 145–158. ACM, 2012.

[6] Frederic Amblard, Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. On the temporal analysis of scientific network evolution. In *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, pages 169–174. IEEE, 2011.

[7] Sinan Aral and Dylan Walker. Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science*, 60(6):1352–1370, 2014.

[8] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin Dunbar. Dynamics of personal social relationships in online social networks: a study on twitter. In *Proceedings of the first ACM conference on Online social networks*, pages 15–26. ACM, 2013.

[9] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Analysis of ego network structure in online social networks. In *SocialCom/PASSAT*, pages 31–40. IEEE, 2012.

[10] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Ego networks in twitter: An experimental analysis. In *INFOCOM*, pages 3459–3464. IEEE, 2013.

[11] Valerio Arnaboldi, Robin IM Dunbar, Andrea Passarella, and Marco Conti. Analysis of co-authorship ego networks. In *International Conference and School on Network Science*, pages 82–96. Springer, 2016.

[12] Valerio Arnaboldi, Andrea Passarella, Maurizio Tesconi, and Davide Gazzè. Towards a characterization of egocentric networks in online social networks. In *OTM Workshops*, volume 7046 of *Lecture Notes in Computer Science*, pages 524–533. Springer, 2011.

[13] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada A. Adamic. The role of social networks in information diffusion. *CoRR*, abs/1201.4145, 2012.

[14] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[15] Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3):590–614, 2002.

[16] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[17] Alex Bavelas. A mathematical model for group structures. *Applied anthropology*, 7(3):16–30, 1948.

[18] Aartjan TF Beekman, Dorly JH Deeg, Theo van Tilburg, Jan H Smit, Chris Hooijer, and Willem van Tilburg. Major and minor depression in later life: a study of prevalence and risk factors. *Journal of affective disorders*, 36(1-2):65–75, 1995.

[19] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[20] Lisa F Berkman. Social support, social networks, social cohesion and health. *Social work in health care*, 31(2):3–14, 2000.

[21] Claire Bidart and Daniel Lavenu. Evolutions of personal networks and life events. *Social networks*, 27(4):359–376, 2005.

[22] Katy Börner, Luca Dall'Asta, Weimao Ke, and Alessandro Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, 10(4):57–67, 2005.

[23] Ronald S Burt. Network items and the general social survey. *Social networks*, 6(4):293–339, 1984.

[24] Shaowei Cai, Kaile Su, and Qingliang Chen. Ewls: A new local search for minimum vertex cover. In *AAAI*, 2010.

[25] Shaowei Cai, Kaile Su, Chuan Luo, and Abdul Sattar. Numvc: An efficient local search algorithm for minimum vertex cover. *Journal of Artificial Intelligence Research*, 46:687–716, 2013.

[26] Maria Chiara Caschera, Arianna D'Ulizia, Fernando Ferri, and Patrizia Grifoni. Studying network dynamics in digital ecosystems. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, page 4. ACM, 2009.

[27] Tanmoy Chakraborty, Sikhar Patranabis, Pawan Goyal, and Animesh Mukherjee. On the formation of circles in co-authorship networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–118. ACM, 2015.

[28] Pei-Shan Chang, I-Hsien Ting, and Shyue-Liang Wang. Towards social recommendation system based on the data from microblogs. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 672–677. IEEE, 2011.

[29] Michael Chau and Jennifer Xu. Using web mining and social network analysis to study the emergence of cyber communities in blogs. In *Terrorism Informatics*, pages 473–494. Springer, 2008.

[30] France Cheong and Brian J Corbitt. A social network analysis of the co-authorship network of the pacific asia conference on information systems from 1993 to 2008. *PACIS 2009 Proceedings*, page 23, 2009.

[31] Chaochang Chiu, Yungchang Ku, Ting Lie, and Yuchi Chen. Internet auction fraud detection using social network analysis and classification tree approaches. *International Journal of Electronic Commerce*, 15(3):123–147, 2011.

[32] Kenneth KS Chung, Liaquat Hossain, and Joseph Davis. Exploring sociocentric and egocentric approaches for social network analysis. In *Proceedings of the 2nd international conference on knowledge management in Asia Pacific*, pages 1–8, 2005.

[33] Kon Shing Kenneth Chung and Walter Christian Paredes. Towards a social networks model for online learning & performance. *Journal of Educational Technology & Society*, 18(3):240, 2015.

[34] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[35] Marco Conti, Andrea Passarella, and Fabio Pezzoni. A model for the generation of social network graphs. In *WOWMOM*, pages 1–6. IEEE, 2011.

[36] Marco Conti, Andrea Passarella, and Fabio Pezzoni. A model to represent human social relationships in social network graphs. In *Proceedings of the 4th International Conference on Social Informatics*, SocInfo'12, pages 174–187, Berlin, Heidelberg, 2012. Springer-Verlag.

[37] Stephanie H Cook, José A Bauermeister, Deborah Gordon-Messer, and Marc A Zimmerman. Online network influences on emerging adults' alcohol and drug use. *Journal of youth and adolescence*, 42(11):1674–1686, 2013.

[38] Edsger Wybe Dijkstra. *A short introduction to the art of programming*, volume 4. Technische Hogeschool Eindhoven Eindhoven, 1971.

[39] Sarra Djemili, Claudia Marinica, Maria Malek, and Dimitris Kotzinos. A definitions' framework for personal/egocentric online social networks. In *7 ème conférence sur les modèles et l'analyse des réseaux: Approches mathématiques et informatiques (MARAMI'16)*, 2016.

[40] Sarra Djemili, Claudia Marinica, Maria Malek, and Dimitris Kotzinos. Personal networks of scientific collaborators: A large scale experimental analysis of their evolution. In *International Workshop on Information Search, Integration, and Personalization*, pages 116–139. Springer, 2016.

[41] Sarra Djemili, Claudia Marinica, Maria Malek, and Dimitris Kotzinos. Un nouveau modéle pour l'évolution de réseaux sociaux personnels. In *11 ème conférence sur les modèles et l'analyse des réseaux: Approches mathématiques et informatiques (MARAMI'20)*, 2020.

[42] Sergei N Dorogovtsev and José FF Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford, 2013.

[43] R. Dunbar and M. Spoors. Social networks, support cliques, and kinship. *Human Nature*, 6(3):273–290, 1995.

[44] Catherine Dwyer, Starr Hiltz, and Katia Passerini. Trust and privacy concern within social networking sites: A comparison of facebook and myspace. *AMCIS 2007 proceedings*, page 339, 2007.

[45] John P Eaton, James C Ward, Ajith Kumar, and Peter H Reingen. Social-structural foundations of publication productivity in the journal of consumer research. In *Essays by Distinguished Marketing Scholars of the Society for Marketing Advances*, pages 199–220. Emerald Group Publishing Limited, 2002.

[46] Ergin Elmacioglu and Dongwon Lee. On six degrees of separation in dblp-db and more. *ACM SIGMOD Record*, 34(2):33–40, 2005.

[47] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[48] Ernesto Estrada. When local and global clustering of networks diverge. *Linear Algebra and Its Applications*, 488:249–263, 2016.

[49] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM, 1999.

[50] Katherine Faust and Stanley Wasserman. *Social network analysis: Methods and applications*, volume 249. Cambridge: Cambridge University Press, 1994.

[51] Scott L Feld. The focused organization of social ties. *American journal of sociology*, 86(5):1015–1035, 1981.

[52] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Alessandro Provetti. The role of strong and weak ties in facebook: a community structure perspective. *CoRR*, abs/1203.0535, 2012.

[53] Simon M Firestone, Michael P Ward, Robert M Christley, and Navneet K Dhand. The importance of location in contact networks: Describing early epidemic spread using spatial social network analysis. *Preventive veterinary medicine*, 102(3):185–195, 2011.

[54] Kai Fischbach, Johannes Putzke, and Detlef Schoder. Co-authorship networks in electronic markets research. *Electronic Markets*, 21(1):19–40, 2011.

[55] Claude S Fischer. *To dwell among friends: Personal networks in town and city.* University of chicago Press, 1982.

[56] Ove Frank and David Strauss. Markov graphs. *Journal of the american Statistical association*, 81(395):832–842, 1986.

[57] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[58] Linton C Freeman. Centered graphs and the structure of ego networks. *Mathematical Social Sciences*, 3(3):291–304, 1982.

[59] Joel Friedman. Computing betti numbers via combinatorial laplacians. *Algorithmica*, 21(4):331–346, 1998.

[60] JA Gallian. Dynamic survey ds6: Graph labeling, electronic j. *Combinatories, DS6*, pages 1–58.

[61] Bo Gao and Bettina Berendt. Circles, posts and privacy in egocentric social networks: An exploratory visualization approach. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 792–796. IEEE, 2013.

[62] Maira A. de C. Gatti, Ana Paula Appel, Cicero Nogueira dos Santos, Claudio Santos Pinhanez, Paulo Rodrigo Cavalin, and Samuel Barbosa Neto. A

simulation-based approach to analyze the information diffusion in microblogging online social network. In *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, WSC '13, pages 1685–1696, Piscataway, NJ, USA, 2013. IEEE Press.

[63] Sharad Goel, Roby Muhamad, and Duncan Watts. Social search in small-world experiments. In *Proceedings of the 18th international conference on World wide web*, pages 701–710. ACM, 2009.

[64] Bruno Goncalves, Nicola Perra, and Alessandro Vespignani. Validation of dunbar's number in twitter conversations. *arXiv preprint arXiv:1105.5170*, 2011.

[65] Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 131–144. ACM, 2012.

[66] Roger V Gould and Roberto M Fernandez. Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological methodology*, pages 89–126, 1989.

[67] Lars Groeger and Francis Buttle. Word-of-mouth marketing: Towards an improved understanding of multi-generational campaign reach. *European Journal of Marketing*, 48(7/8):1186–1208, 2014.

[68] Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):795–813, 2006.

[69] Daniel S Halgin and Stephen P Borgatti. An introduction to personal network analysis and tie churn statistics using e-net. *Connections*, 32(1):37–48, 2012.

[70] Frank Harary and Robert Z Norman. *Graph theory as a mathematical model in social science*. University of Michigan, Institute for Social Research Ann Arbor, 1953.

[71] Martin Harrigan, Daniel Archambault, Pádraig Cunningham, and Neil Hurley. Egonav: Exploring networks through egocentric spatializations. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 563–570. ACM, 2012.

[72] Penelope Hawe and Laura Ghali. Use of social network analysis to map the social relationships of staff and teachers at school. *Health education research*, 23(1):62–69, 2007.

[73] R.A. Hill and R.I.M. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, 2003.

[74] Jian Huang, Ziming Zhuang, Jia Li, and C Lee Giles. Collaboration over time: characterizing and modeling network evolution. In *Proceedings of the 2008 international conference on web search and data mining*, pages 107–116. ACM, 2008.

[75] Mohsen Jamali, Gholamreza Haffari, and Martin Ester. Modeling the temporal dynamics of social rating networks using bidirectional effects of social relations and rating patterns. In *Proceedings of the 20th international conference on World wide web*, pages 527–536. ACM, 2011.

[76] Alexander P Kartun-Giles and Ginestra Bianconi. Beyond the clustering coefficient: A topological analysis of node neighbourhoods in complex networks. *arXiv preprint arXiv:1901.10978*, 2019.

[77] Peter D Killworth and H Russell Bernard. The reversal small-world experiment. *Social networks*, 1(2):159–192, 1978.

[78] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer, 2010.

[79] Jérôme Kunegis, Damien Fay, and Christian Bauckhage. Network growth and the spectral evolution model. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 739–748. ACM, 2010.

[80] Massimiliano La Gala, Valerio Arnaboldi, Marco Conti, and Andrea Passarella. Ego-net digger: a new way to study ego networks in online social networks. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, pages 9–16. ACM, 2012.

[81] Silvio Lattanzi, Alessandro Panconesi, and D Sivakumar. Milgram-routing in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 725–734. ACM, 2011.

[82] Edward O Laumann, John H Gagnon, Stuard Michaels, Robert T Michael, and James S Coleman. Monitoring the aids epidemic in the united states: A network approach. *Science*, 244(4909):1186–1189, 1989.

[83] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470. ACM, 2008.

[84] Menglin Li and Colm O'Riordan. The effect of clustering coefficient and node degree on the robustness of cooperation. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 2833–2839. IEEE, 2013.

[85] Miranda J Lubbers, José Luis Molina, Jürgen Lerner, Ulrik Brandes, Javier Ávila, and Christopher McCarty. Dynamic analysis of personal networks. the case of argentinean migrants in spain.

[86] R Duncan Luce and Albert D Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

[87] Peter V Marsden. Social networks. *Encyclopedia of sociology*, 4:2727–2735, 2000.

[88] C McCarty and JL Molina. Personal networks: research and applications. *Florida, USA: University of Florida*, pages 1–165, 2013.

[89] Christopher McCarty, José Luis Molina, Claudia Aguilar, and Laura Rota. A comparison of social network mapping and personal network visualization. *Field methods*, 19(2):145–162, 2007.

[90] Jean Marie McGloin and David S Kirk. An overview of social network analysis. In *Advancing Quantitative Methods in Criminology and Criminal Justice*, pages 75–87. Routledge, 2014.

[91] Donnay McNally and Jonathan Alston. The use of social network analysis (sna) in the examination of an outlaw motorcycle gang. *Journal of Gang Research*, 13(3):1, 2005.

[92] Stanley Milgram. I967. the small world problem. *Psychology Today, 1 (1), May*, page 60, 1967.

[93] James Clyde Mitchell. *Social networks in urban situations: analyses of personal relationships in Central African towns*. Manchester University Press, 1969.

[94] JL Moreno. Psychological and social organization of groups in the community. In *Proceedings & Addresses. American Association on Mental Deficiency*, 1933.

[95] Gail P Myers, Gene A McGrady, Clementine Marrow, and Charles W Mueller. Weapon carrying among black adolescents: a social network perspective. *American Journal of Public Health*, 87(6):1038–1040, 1997.

[96] Alan Neaigus, Samuel R Friedman, Richard Curtis, Don C Des Jarlais, R Terry Furst, Benny Jose, Patrice Mota, Bruce Stepherson, Meryl Sufian, Thomas Ward, et al. The relevance of drug injectors' social and risk networks for understanding and preventing hiv infection. *Social science & medicine*, 38(1):67–78, 1994.

[97] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

[98] Mark EJ Newman. Ego-centered networks and the ripple effect. *Social Networks*, 25(1):83–95, 2003.

[99] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[100] Mark EJ Newman. Who is the best connected scientist? a study of scientific coauthorship networks. In *Complex networks*, pages 337–370. Springer, 2004.

[101] Paul Nieuwbeerta and Henk Flap. Crosscutting social circles and political choice: Effects of personal network composition on voting behavior in the netherlands. *Social Networks*, 22(4):313–335, 2000.

[102] A James O'malley, Samuel Arbesman, Darby Miller Steiger, James H Fowler, and Nicholas A Christakis. Egocentric social network structure, health, and prosocial behaviors in a national panel study of americans. *PLoS One*, 7(5):e36250, 2012.

[103] José Luis Ortega. Influence of co-authorship networks in the research impact: Ego network analyses from microsoft academic search. *Journal of Informetrics*, 8(3):728–737, 2014.

[104] Emmanuel Pannier. L'analyse des réseaux sociaux: théories, concepts et méthodologies.

[105] Andrea Passarella, Robin I. M. Dunbar, Marco Conti, and Fabio Pezzoni. Ego network models for future internet social networking environments. *Comput. Commun.*, 35(18):2201–2217, November 2012.

[106] Thomas V Pollet, Sam GB Roberts, and Robin IM Dunbar. Use of social network sites and instant messaging does not lead to increased offline social network size, or to emotionally closer relationships with offline network members. *Cyberpsychology, Behavior, and Social Networking*, 14(4):253–258, 2011.

[107] Derek J De Solla Price. Networks of scientific papers. *Science*, pages 510–515, 1965.

[108] Liudmila Ostroumova Prokhorenkova and Egor Samosvat. Global clustering coefficient in scale-free networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 47–58. Springer, 2014.

[109] Daniele Quercia, Licia Capra, and Jon Crowcroft. The social world of twitter: Topics, geography, and emotions. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[110] Daniel Richter, Kai Riemer, and Jan vom Brocke. Internet social networking. *Wirtschaftsinformatik*, 53(2):89–103, 2011.

[111] Sam G. B. Roberts, Robin I. M. Dunbar, Thomas V. Pollet, and Toon Kuppens. Exploring variation in active network size: Constraints and ego characteristics. *Social Networks*, 31(2):138–146, 2009.

[112] Nicola Santoro, Walter Quattrociocchi, Paola Flocchini, Arnaud Casteigts, and Frederic Amblard. Time-varying graphs and social network analysis: Temporal indicators and metrics. *arXiv preprint arXiv:1102.0629*, 2011.

[113] Rui Sarmento, Mário Cordeiro, and João Gama. Visualization of evolving large scale ego-networks. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 960–962. ACM, 2015.

[114] Fabian Schneider, Anja Feldmann, Balachander Krishnamurthy, and Walter Willinger. Understanding online social network usage from a network perspective. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pages 35–48. ACM, 2009.

[115] Konstantinos Semertzidis and Evaggelia Pitoura. Time traveling in graphs using a graph database. In *EDBT/ICDT Workshops*, 2016.

[116] Tom AB Snijders. Stochastic actor-oriented models for network change. *Journal of mathematical sociology*, 21(1-2):149–172, 1996.

[117] Tom AB Snijders. The statistical evaluation of social network dynamics. *Sociological methodology*, 31(1):361–395, 2001.

[118] Xin Sun, Junyu Dong, Ruichun Tang, Mantao Xu, Lin Qi, and Yang Cai. Topological evolution of virtual social networks by modeling social activities. *Physica A: Statistical Mechanics and its Applications*, 433:259–267, 2015.

[119] Alistair Sutcliffe, Robin Dunbar, Jens Binder, and Holly Arrow. Relationships and the social brain: integrating psychological and evolutionary perspectives. *British journal of psychology*, 103(2):149–168, 2012.

[120] John Tang, Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, and Vito Latora. Small-world behavior in time-varying graphs. *Physical Review E*, 81(5):055101, 2010.

[121] Dieudonné Tchuente. *Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux : approche à partir de communautés de réseaux k-égocentriques*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, janvier 2013.

[122] Jeffrey Travers and Stanley Milgram. The small world problem. *Phychology Today*, 1:61–67, 1967.

[123] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

[124] Thomas W Valente, Susan C Watkins, Miriam N Jato, Ariane Van Der Straten, and Louis-Philippe M Tsitsol. Social network associations with contraceptive use among cameroonian women in voluntary associations. *Social science & medicine*, 45(5):677–687, 1997.

[125] Marijtje AJ Van Duijn, Jooske T Van Busschbach, and Tom AB Snijders. Multilevel analysis of personal networks as dependent variables. *Social networks*, 21(2):187–210, 1999.

[126] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.

[127] Caroline S Wagner and Loet Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research policy*, 34(10):1608–1618, 2005.

[128] Stanley Wasserman. *Advances in social network analysis: Research in the social and behavioral sciences*. Sage, 1994.

[129] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[130] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world'networks. *nature*, 393(6684):440–442, 1998.

[131] Qinghua Wu and Jin-Kao Hao. A review on algorithms for maximum clique problems. *European Journal of Operational Research*, 242(3):693–709, 2015.

[132] Yanhong Wu, Naveen Pitipornvivat, Jian Zhao, Sixiao Yang, Guowei Huang, and Huamin Qu. egoslider: Visual analysis of egocentric network evolution. *IEEE transactions on visualization and computer graphics*, 22(1):260–269, 2016.

[133] Erjia Yan and Ying Ding. Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10):2107–2118, 2009.

[134] Fei Yan, ShaoWei Cai, Ming Zhang, GuoJun Liu, and ZhiHong Deng. A clique-superposition model for social networks. *Science China Information Sciences*, 56(5):1–19, 2013.

[135] Tso-Jung Yen. Analysis of egocentric network data, August 2014. Lecture available at http://www3.stat.sinica.edu.tw/summerschool2014/document/0801-2.pdf Retrieved on 23/02/2015.

[136] Jian Zhao, Michael Glueck, Fanny Chevalier, Yanhong Wu, and Azam Khan. Egocentric analysis of dynamic networks with egolines. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5003–5014. ACM, 2016.

[137] Thomas Zimmermann and Nachiappan Nagappan. Predicting defects using network analysis on dependency graphs. In *Proceedings of the 30th international conference on Software engineering*, pages 531–540. ACM, 2008.