Structural and Behavioral Analysis and Modeling of the Society

Teză destinată obţinerii
titlului ştiinţific de doctor inginer
la
Universitatea Politehnica Timişoara
în domeniul Calculatoare şi tehnologia informaţiei
de către

ing. **Alexandru TOPÎRCEANU**

Conducător ştiinţific: Prof. dr. ing. Mircea VLĂDUŢIU
Referenţi ştiinţifici: Acad. Mircea PETRESCU – UPB București
Prof. dr. ing. Liviu MICLEA – UTC Cluj-Napoca
Prof. dr. ing. Nicolae ROBU – UPT Timișoara

Ziua susţinerii tezei: 12.02.2016

This work was partially supported by the strategic grant POSDRU/159/1.5/S/ 137070 (2014) of the Ministry of National Education, Romania, co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007-2013.

Topîrceanu, Alexandru

Structural and Behavioral Analysis and Modeling of the Society

Teze de doctorat ale UPT, Seria 14, Nr. 30, Editura Politehnica, 2016, 229 pagini, 80 figuri, 38 tabele.
ISSN: 2069-8216
ISSN-L: 2069-8216
ISBN: 978-606-35-0049-7

Cuvinte cheie: social networks analysis, network topologies, graph theory, network similarity, diffusion modeling, simulation, software development.

Rezumat:

Cercetările din prezenta teză se axează pe două direcţii de mare interes din domeniul interdisciplinar al analizei de reţele sociale: pe de o parte propun analiza structurală a mai multor tipuri de reţele complexe, de colaborare, sintetizată prin modelarea topologică mai realistă a reţelelor sociale, şi pe de altă parte propun un nou model de interacţiune socială bazat pe un concept nou introdus denumit *toleranţă*. Prin aceste modele se reuşeşte îmbunătăţirea reproducerii dinamicii opiniei prin simulări. Elementele de originalitate constau în utilizarea algoritmilor genetici, a ataşamentului preferenţial folosind metrica de *betweenness*, şi a tiparelor de interacţiune cu limite dinamice a agenţilor sociali.

Lucrarea se evidenţiază prin aplicarea conceptelor noi din domeniul ştiinţei reţelelor pe date empirice, pentru a înţelege şi a modela mai bine procesele sociale din lumea reală.

Abstract

The recently introduced term, coined as *New Network Science,* facilitates the understanding of many emergent phenomena in nature and society, and is a major trend on the modern scientific scale. One branch of Network Science that has attracted much attention in the last decade is *Social Networks Analysis*. The benefit of understanding the complex processes behind how people adopt and form their own opinions about surrounding problems is an important concern for research fields like Psychology, Philosophy, Politics, Marketing, Finances, and even Warfare, and it can be alleviated using network analysis.

Social media is constantly modifying the way we create, share and consume information, and has become a powerful tool for understanding social trends, and society as a whole. The goal of this thesis is to help in the understanding and better prediction of diffusion phenomena, by using the computer as a tool for social networks analysis. Relying on computer science as a means for modeling and analysis of the underlying social topologies and individual interaction models, I focus on understanding systems of people and when they become stable, as well as the connections that cause events in social networks As such, I make use of the emerging interdisciplinary field of *Social Networks Analysis,* which sheds new light into the modeling of social opinion dynamics and personal opinion fluctuations, of how people influence each other and how they can be influenced.

The presented work begins with the analysis at the topological level of human relationship establishment, then explains and models network growth and interaction based on original and validated socio-psychological assumptions, and reaches the meta-level of human interaction models. These models are a current scientific (and also socio-political) barrier in predicting social emergence and being able to design more stable and safe social systems in the future.

I achieve the goals to model social interaction, network structure and network growth more accurately, and ultimately discuss how decision factors can be influenced at the macroscopic level of the society we live in. Like most of the sciences studying opinion and influence, this work models the decision process by combining elements from Psychology, Sociology, Anthropology, and Computer Science.

Keywords: social networks analysis, network topologies, graph theory, network similarity, diffusion modeling, simulation, software development.

Preface

"The world is governed by opinion."
Thomas Hobbes (philosopher, 1588 – 1679)

To my dear family and closest friends who give me the ambition to surprise myself every day.

Acknowledgement

Ever since I was a child, I remember myself being fascinated about the world of engineering. Even though I was barely touching the tip of the iceberg as a student, my intuition told me to follow a path of innovation and breakthroughs using all the knowledge I could acquire. This inner feeling made me curious to question nature using science, and more specifically using computer science. I remember clearly when I was greeted with a lot of warm friendship and respect into the ACSA research group. Since then, I have become a full member of the group, and aspiring for a greater contribution in science, I have decided to follow a PhD. It turned out to be a wonderful experience, and if I think back on the last three or more years, I have many people to give my appreciations to.

First, I want to thank my adviser Prof. Mircea Vladutiu who accepted me as his PhD student. I know how prestigious a position is in Prof. Vladutiu's team, so I want to express my full gratitude tenfold for having this opportunity to develop myself. He encouraged me to follow this new path of social networks analysis, and made sure I was keeping the right scientific track. Apart from the lucrative professional collaboration, he has always been there for me as a beacon of wisdom - he helped me overcome various issues and guided me through difficult decisions.

Second, I wish to express my deepest gratitude to Assoc. Prof. Mihai Udrescu, to whom I owe the whole topic of network science. He inspired me to take up the challenge of a first PhD in social networks in this university, so I remain perpetually grateful. Apart from a first inspiring spark, he provided innovative ideas to the project and lead me with an iron fist relentlessly. Also, he has proven to be a very good and dependable friend.

I would also like to thank the whole ACSA research group for the periodic brainstorming sessions which have proved valuable to my work. My colleagues Alexandru Iovanovici, Cristian Cosariu and Andreea Bozesan kept me motivated as they were one year ahead of me in the doctoral studies. We also studied and worked together, producing many valuable results. I want to thank Assoc. Prof. Lucian Prodan who is an admirable knowledge resource, and with his vast experience, he has also contributed to refining the ideas behind my work. Assist. Prof. Flavius Opritoiu has been always been a wise and rigorous member of the ACSA team, and he has inspired me to be as him, so I thank

you.

Finally, I want to especially thank Prof. Radu Marculescu from Carnegie Mellon University, USA, for monitoring my progress and being interested in my research. His world-class expertise proven invaluable for my doctoral studies, and I have to admit I learned a lot by collaborating on scientific papers and talks with him. He has an important role in keeping my research on the right track on several occasions.

Contents

List of Tables

List of Figures

1. Introduction

A major trend of modern science is the study and understanding of social opinion dynamics and individual opinion fluctuations, of how people influence each other and how they can be influenced. The benefit of understanding the complex processes behind how people adopt and form their own opinions about surrounding problems is a major concern for sciences like Psychology, Philosophy, Politics, Marketing, Finances and even Warfare [23, 146, 154, 82].

Financial sciences, for example, study the markets and consumers to improve profits. Marketing uses many techniques to understand the needs, the strengths and weaknesses of different social layers or groups. One of the key roles is to understand how any current marketing mix (product, price, place, and promotion) impacts consumer behavior [182]. This research focuses on understanding when, how, where and why a product may be bought by people and how these factors can be influenced. It models the buying process by combining elements from psychology, sociology, anthropology and economics [82].

Politics use social studies to study the political influence of parties and the means to create a consensus among voters. Whether an agreement, a cooperative or collaborative consensus is sought after, political parties are interested in an overall public opinion rather than the opinion of individuals [119]. Thus, a similarity to networks of computers can be seen, in which the overall throughput and correct packet delivery is important, rather than the performance and specifications of an individual node. Politics use diverse agents to ensure propaganda, especially before elections. These agents may include political representatives of parties giving speeches to groups or individuals, radio and television programs, written posters and leaflets and internet based propaganda. Propaganda is an appeal to emotion, it does imply intellect or any level of knowledge, and this is what makes it hard to quantify but also very effective in human opinion formation. There are numerous types of propaganda, based on persuasion, intimidation, ideological and national beliefs. One example of successful propaganda was the encouragement of women to take up men's jobs in factories to aid the war effort for the United States. This eventually contributed to the emancipation of women at the beginning of the 20th century by allowing more and more women to obtain better ranked jobs [115, 80].

Warfare has always used counter-intelligence to stop enemy propaganda and spies to influence the enemy's morale. Psychological warfare, even though older, was successfully reintroduced during the Second World War with the help of leaflet bombs. This type of propaganda had the purpose of turning civilians against their own forces through intimidation, promise of rewards or assistance [46]. One successful example of leaflet usage was during the First Gulf War when eighty thousand troops of the Iraqi forces surrendered to the Americans.

Social Science is a term used to encapsulate a large number of sciences branching from the study of society and human behavior. Its foundations are considered to have been laid down by E. Durkheim, K. Marx and M. Weber during the 19th century. There are three approaches in applying this science:

- Positivism, using studies based on natural science to explain social behavior [114]. This principle is based on empiricism and scientific methods are considered to provide a valid foundation for sociological research based on the fact that the only authentic knowledge is scientific knowledge. Quoting E. Durkheim: "Our main goal is to extend scientific rationalism to human conduct.... What has been called our positivism is but a consequence of this rationalism" [81].

- Interpretive sociology (or antipositivism), based on understanding how social actions affect the people, rather than explaining the feedback based on investigations of the natural world [100].

- Modern eclectic approach, combining multiple techniques.

Social research is conducted on understanding social phenomena, designing models or patterns and proving them with social evidences. Based on what type of evidence is used, there are two types of social research approaches [104, 77]:

- Quantitative design that rely on statistical data and offer reliable conclusions.

- Qualitative designs that rely on direct observations and tend to study subjective accuracy over statistical generality

The results obtained through social research are used to implement various crowd control techniques. Crowd manipulation is one such technique which is based on the research of crowd psychology. Scientists can prove that the psychology of an individual within the crowd is different from the overall crowd psychology, yet groups of people can act together for a common goal. It is this aspect that makes crowd manipulation a tool to influence groups of people to behave in a specific manner, to be directed towards a desired action [282, 285].

As a conclusion, social research is proving useful in understanding the mechanisms which transform individual opinion into a wide-spread social opinion; how opinions affect individuals and how they evolve in time as seen on a macroscopic scale. Modeling social behavior can be both a means of defending and boosting democratic rights as well as a means to impose and manipulate a society or a social layer [72]. From the riots organized during the French Revolution (1793) and the Boston Massacre (1770) to the opinion on global warming, ecology or modern warfare (London, 2009) [33], all these events are the result of social interconnection and social opinion building between people.

1.1. Thesis domain

New Network Science is receiving an increased interest from many fields of science since many empirical observations of our surrounding world show the same properties, regardless of whether they are of natural or synthetic origin [89]. There are topological models which describe geographical proximity, friendship distribution, neural networks in the brain, protein interaction mechanisms, natural food chains, the distribution of means of transportation, citation networks, sexual interaction patterns, the world wide web, power distribution networks, relationship of words in a language, interaction between ingredients in a recipe, the world markets, political structures [10, 276, 95, 250, 131, 82, 214].

As a branch of this network science, the major goal of *Social Networks Analysis* (SNA) is to analyze, understand and model real social networks. It can focus on the topological level of a network, i.e. how the nodes interconnect, or on the behavioral level, i.e. how the nodes interact. Both of these problems are approached through empirical studies (direct and indirect measurements, surveys, statistical analysis etc.) which lead to the proposal of a model for the observed real social network [116]. The interdisciplinarity of the *New Network Science* brings together many fields of science which process *big data* modeled as graphs [95, 89]. Regardless of the representation of nodes, edges, edge directions, and edge weights, this data often undergoes numerical comparison, sampling, and statistical analysis to extract relevant patterns from it. To that end, network scientists use diverse state of the art comparison techniques, but there is no single methodology to express similarity/dissimilarity in an objective and uniform manner. My proposal uses solely the topological properties of the underlying graph.

The scientific domains treated in this thesis are: social networks analysis, complex networks, network topologies, social opinion dynamics, graph theory, and statistics.

The presented work focuses on two main directions, equally important for social opinion modeling (see Figure 1.1):

- Creating networks which model the basic connection patterns of a real society, and
- A robust social interaction model, namely a set of rules which describe how agents behave in a society.

Both directions are important as the topology defines the physical interconnection rules, creating information saturation and hub nodes due to its layout, while the communication model defines the actual evolution of an agent's decision process, which forms its opinion. Thus, both directions have been researched a-priory for this thesis.

Regardless of the science in question, computational social science is still in its infancy. However, an increased interest is shown in this topic as several conferences in computer sciences are creating categories destined for research oriented towards social behavior. To better understand the social processes a better collaboration between natural sciences and applied sciences is needed, as both possess valuable knowledge [106]. Using the current computational power, computers can help researchers analyze interleaved mathematical and psychological models at a faster rate. Of course, validating results with empirical data is the final step in proving that a social process is understood. Recent mathematical research proposes new ways of modeling societies or clusters of individuals and present results of great theoretical value [112].

1.2. Motivation and impact

A noteworthy study states that our daily "social transactions" leave digital fingerprints which offer increasingly comprehensive pictures of both individuals and the groups we pertain to, with the potential of transforming the understanding of our lives, organizations, and societies in a fashion that was barely conceivable until recently [154]. The capacity to collect and analyze massive amounts of data is transforming fields like biology, economy and physics. However, the emergence of data-driven computational science has been much slower, carefully directed by a few intrepid computer

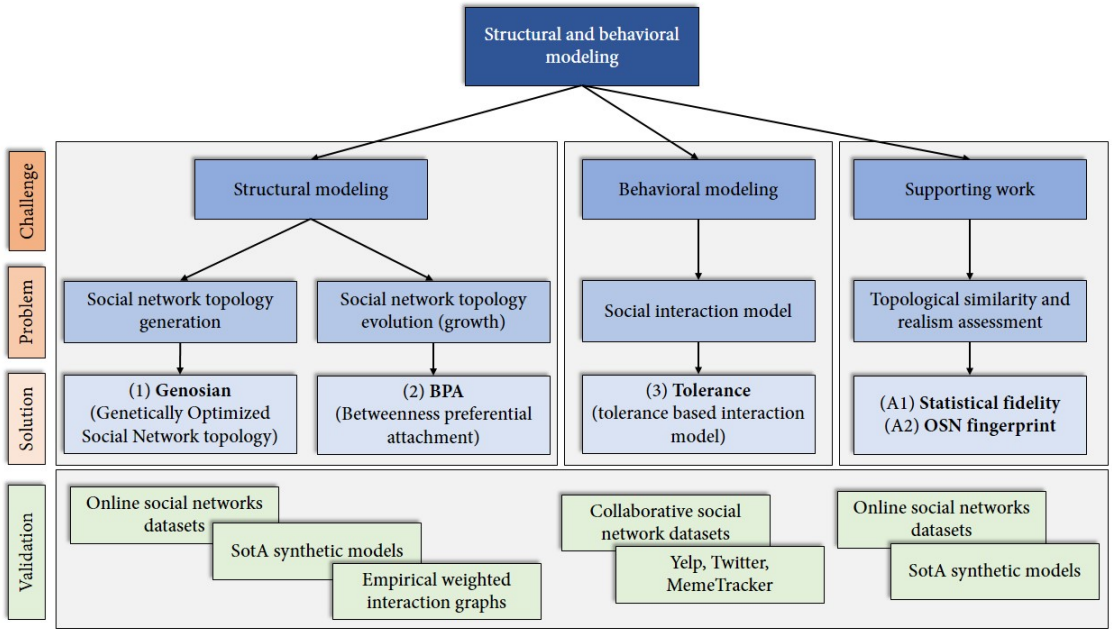


Figure 1.1.: The contributions of this thesis: structural and behavioral modeling.

scientists, physicists, and social scientists [187, 281, 25, 205, 154]. Current emergence of computational science impacts many research directions, and thus I consider my thesis to come and serve this movement at its most fundamental levels.

In light of the movement started by Alex Pentland - *reinventing society in the wake of big data* - current challenges include analysis, capture, data filtering, search, sharing, storage, transfer, visualization, and information privacy. The term *Big Data* casually refers simply to the use of predictive analytics or other certain advanced methods to extract value from data. While the size of the data is not always considered a decisive factor, of course, accuracy in big data may lead to better decision making, which lead to greater operational efficiency, cost reduction or reduced risk in social systems [176]. The envisioned analysis of data can find new correlations, to detect trends in business, prevent diseases, combat crime [71]. Researchers, marketing analysts, media practitioners, general advertising and governments regularly meet difficulties with large data sets in areas including Internet search, finance and business information. Scientists encounter limitations in computational science work, like predicting earthquakes, weather, genomics [270], connectomics, complex physics simulations [59], and biological and environmental research [227].

With Big Data we can now begin to actually look at the details of social interaction and how those play out, and are no longer limited to averages like market indices or election results. It could prove used for good or for ill, and so Big data brings us to interesting times. We're going to end up reinventing what it means to have a human society [177, 181].

The work presented in this thesis analyzes current social models and their relevance in opinion dynamics. The contribution is divided in two major sections: graph analysis and modeling using existing empirical data of social networks, and computational simulation and mathematical modeling

of the opinion evolution in a society.

A relevant survey made in the direction of improving diffusion models [112] creates a taxonomy which falls in line with the goals of my thesis, highlighting its scientific impact. Figure 1.2 shows the challenges and approaches that at of very high interest in current SNA.

The first and foremost challenge is to determine the topics which form the actual "opinion" in the society. Social media is an undeniable layer of our daily lives in the 21st century, and contains most of the communication undertaken by people. It is imperative that we are able to filter out those topics of interest for a certain research context (e.g. solely political, marketing context etc.). The two available approaches are based on term frequency and social interaction frequency. More specifically, relevant terms can be extracted through data mining, like analyzing tweets, emails, of Facebook messaging coming from all active users. Conversely, we can analyze the most intensive links between users and extract their common topics.

Once the topics which represent opinion are known, research can be oriented towards understanding and modeling the diffusion processes of those topics. The is an exploratory approach, which uses solely empirical data and try to reproduce similar theoretical models. Static networks can be used, e.g. a dataset of millions of tweets and re-tweets that has been data-mined in the past, to analyze messaging intensity and try to reproduce a similar dynamics phenomenon in a synthetic graph. Also, dynamic networks can be used, e.g. real-time mining of Twitter, to model the same synthetic processes. The latter approach is useful when no datasets are available, but is constricted by the usual API limitations of modern social platforms. Another approach is the predictive one. This can be graph-based, namely is uses mathematical models for agent interaction to try and reproduce empirical observations. Graph based approaches rely on the Independent Cascade (IC) [108] or Linear Threshold (LT) [110] diffusion models. The IC model requires a diffusion probability to be associated to each edge, whereas LT requires an influence degree to be defined on each edge and an influence threshold for each node. Both model proceed iteratively along a discrete time-axis, starting from a set of initially opinionated nodes, commonly named early adopters [231]. Non-graph-based approaches do not depend on the underlying topology, but rather divide nodes into several types. As such, there are two models pertaining to epidemiology, the SIR and SIS models [120].

Finally, once both opinion and diffusion are defined, it is important to determine which nodes act as spreaders of the opinion. This helps simulate and predict the outcome of the diffusion phenomena. There are topological approaches using centralities or motifs, as well as other types of approaches.

In the presented context, my thesis revolves around graph-based predictive models for modeling diffusion processes, and uses static empirical networks for validation purposes. Also, is offers a topological perspective of how social network structure emerges and evolves. This is achieved by creating an original topological model, introducing a personal growth model and doing empirical online social networks analysis using motifs.

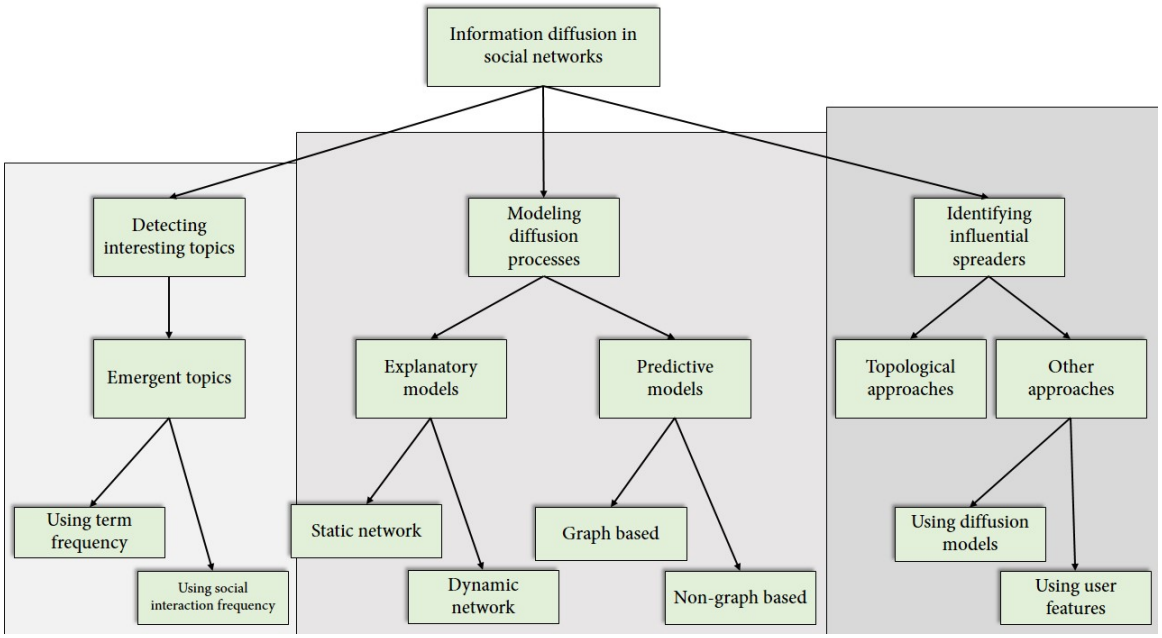


Figure 1.2.: The main scientific directions in better understanding information diffusion in social networks and the corresponding types of approaches.

2. Theoretical foundations

Complex networks cover an active area of scientific research inspired largely by the empirical study of real-world networks such as communication networks, economical networks and social networks. They are classified into four major types, based on the context which they model: biological networks (e.g., metabolic networks, transcription regulatory networks, protein-protein interaction networks, protein structure networks, neural networks, ecological networks, natural food chains) [10, 276, 82], social networks (e.g. friendship networks, citation networks, voter networks, world markets, political structures) [246, 276, 214], technological networks (e.g., computer networks, electrical circuits, road networks) [10], and semantic networks (e.g. word-net [188], recipe networks [250]). Without exception, all these networks can be represented as graphs, which include a wide variety of subgraphs.

This section carefully introduces every notion of complex networks, ranging from graph science to some of the most advanced features of social network analysis. The purpose of this section is to offer most readers with basic concepts of mathematics n overview of the topics discussed later in this thesis. It is written using both consistent mathematical notations, and an approachable scientific language.

Complex networks deal with a plethora of models, so apart from the common fundamentals of graphs, the section and thesis focus solely on social-context specific notions. As my thesis is elaborated for the degree of doctor in computer science, the first subsection introduces the reader to the relationship between computer science and SNA. Next, I present the most relevant metrics used in graph science (ranging from nodes and edges to centrality distributions), followed by social specific concepts (i.e. used for opinion modeling). The fourth subsection is dedicated to social network topologies, which represents one of the two main points of study for my thesis. I present the fundamental models which have inspired the advanced topologies that were an inspiration for my proposals. Next, I refer to the second part of my thesis, by introducing the state of the art social interaction models. Finally, I discuss the standards, trends and limitations in creating and comparing social models.

2.1. Social networks: An introduction to computer science

Social networks, in computer science, are a branch of complex networks, and their theory is based on network theory, graph theory and network science. The main purpose of social networks is to model the structure and relationships between persons in a real society [277]. The structure can be further generalized to groups of persons, clusters, layers, cities, states etc., each group with a particular set of defining characteristics. This area of science was proposed in the 1970s [197] and was based on empirical observations of computer networks and human networks, with many ideas coming from the distant field of sociology. Even though 40 years old, only recently (2010s) has this field started to attract great interest from universities and researchers around the world.

Even though the term social network used in conjunction with sites like Facebook or Twitter has

a different meaning than the term used in computer science, the similarities have offered a good perspective in science. As researchers C. Alt et al have explained, it is the evolution of companies like Facebook who have generated the interest of computer science in social networks [12]. Not only does the existing data offer social researchers valuable feedback on their work, presumptions and conclusions but the extended usage of social sites attracts more and more students into this area of science.

A social network is a construction with individuals (actors, agents) and bidirectional connections (relationships, friendships) between these individuals resembling a real social structure of people. The role of such a network is to provide information on how relationships evolve and how information is passed within the society as determined by the interactions. The two main aspects of a social network are the network topology and the agent interaction model. An important property of social networks is that they are self-organizing and emergent. Patterns present at a small scale, inside a small group of agents, replicate themselves at a greater scale. However, with increased network size, the information output becomes overwhelming. That is why studying a too general network (e.g. a country, the world) becomes unfeasible. Consequently, studies are done only on relevant groups with clearly defined properties so that the output information is unbiased. Usually there are three levels of social network composition and study: micro-level (studying an individual and its relationships), middle-level, and macro-level (studying effects on large populations, regardless of individual effects). The presented work manages micro-level social networks at the level of interactions, but analyzes the results at a macro-level as only the overall opinion distribution is relevant. The levels at which the research was done are [198]:

- Actor (Agent) level: the smallest unit analyzed in the network. It encapsulates metrics like tolerance, confidence, credibility, trust.
- Dyadic level: the relationship between two actors. All relationships are bidirectional (i.e. both actors are friends with each other) and transport one opinion from a source to a destination.
- Triadic level: represents the smallest social molecule in a society. It is formed by any two actors with a relationship between them. This relationship permits interconnections that alter the opinion, confidence and tolerance of both actors.

2.2. Metrics of complex networks

Several concepts specific to complex and social networks are presented in this section, as they will also be used over the course of the whole thesis. These concepts - as found in literature - are introduced as follows.

2.2.1. Graphs: nodes, edges, degrees and weights

As the building blocks of social networks consist of mathematical graphs, I start by defining this abstract data type which is commonly used in mathematics and computer science to model pairwise relations between objects. A graph $G = (V, E)$ consists of vertices (nodes) V which have connections between each other through the set of edges E. The graph may be undirected, meaning that

edges are symmetrical in terms of the two ends (there is no distinction between the two vertices associated with each edge), or its edges may be directed from one vertex V_i to another V_j. In this case, we can say that there is a path from V_i to V_j, but not vice-versa.

Nodes represent the abstraction of any natural or synthetic process for which network science may be used. At the most basic level, each node possesses an identity (name) and a set of edges through which it connects to other nodes. Often, nodes possess context-specific properties which are used in research to see how these properties cluster together - using bipartite graphs, and community detection methods [202, 205, 198].

Edges represent a relationship between two edges, connecting them, being either undirected or directed. If say, edge $e(i,\ j)$ connects nodes v_i and v_j and is undirected, then both nodes can be reached following e from the other end. If $e(i,\ j)$ is directed, then only a path from v_i to v_j exists in graph G. Moreover, in a directed context, it makes sense to have another edge $e^{'}(j,\ i)$ which creates a path in the reverse direction. Edges may also be weighted or unweighted. In the context of undirected graphs, each edge from E may be associated a weight equal to 1 for computing paths or costs. If the relationship between nodes implies different magnitudes, then weights may be assigned to edges. A special case of edge is the self-loop, in which a node redirects to itself (e.g. a web page has a link that redirects to itself).

The degree of a node is the number of nodes with which it is connected through graph edges. In directed graphs, a node has two degrees: an out-degree for edges exiting the node, and an in-degree for incoming edges.

A path in a graph is a sequence of edges that connect a starting node to a destination node. If there is no possibility to reach a certain node, then there is no path between that pair of nodes. A graph is called *connected* if any node may be reached from any other node of the graph following any path. For a connected undirected graph $G = (V,\ E)$ we have the following relationship between the minimum/maximum number of edges and the number of nodes:

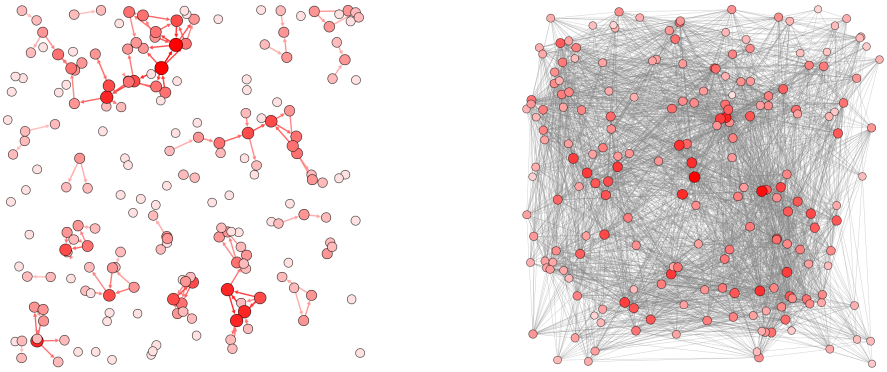
$$minimum\ |E| = |V| - 1 \tag{2.1}$$

$$maximum\ |E| = \frac{|V| \times |V-1|}{2} \sim |V|^2 \tag{2.2}$$

If the number of edges is maximal, then the graph is considered fully connected and there is a path of length 1 from any node to any other node. However, such graphs don't usually exists in nature, as connections are much sparser. To keep the graph connected (often a requirement in graph modeling meant to study natural processes) at least $|V| - 1$ edges have to be left in G.

2.2.2. Density and diameter

Based on the notions introduced in the previous subsection, we call G a dense graph when the number of edges E is close to the maximal number of edges. The opposite, a graph with only a few edges, is a sparse graph. The distinction between sparse and dense graphs is rather vague, and depends on the context. A different definition exists for density whether we refer to undirected or directed graphs. For undirected graphs, the graph density is defined as:



(a) (b)

Figure 2.1.: An illustration of a sparse graph (**a**) and a dense graph (**b**). The sparse graph has 200 nodes, 131 edges, is sparse, disconnected, with a density of $D = 0.0065$. The dense graph has 200 nodes, 3500 edges, an average degree of 17.5, may be considered dense, and is connected, with a density of $D = 0.1758$. All nodes are sized and colored in red direct proportional to their degree.

$$D = \frac{2\,|E|}{|V| \times (|V| - 1)} = \frac{|E|}{max(|E|)} \tag{2.3}$$

Using equation 2.2. the density D can be considered the number of edges which exist in the graph, divided by the maximum number of edges which can exist. For directed graphs, the graph density is slightly modified as:

$$D = \frac{|E|}{|V| \times (|V| - 1)} = \frac{|E|}{max(|E|)} \tag{2.4}$$

The lack of a $2\times$ modifier comes form the fact that directed graphs have a double number of maximum edges, that is, each pair of nodes can have two edges connecting them. The minimum density D is 0, and the maximum D is 1. An example of two graphs with opposing properties have been generated and depicted in Figure 2.1.

2.2.3. Degree distribution

The degree distribution of a network is a function describing the probable distribution of node degrees over that network. The basic characteristic of a single node in a network is its degree. The degree is the number of connections it has to other nodes and is denoted by $deg(v)$. Depending on the type of graph, there can be an in-degree ($deg - (v)$) and an out-degree ($deg + (v)$), for incoming respectively outgoing connections. Undirected graphs, like social networks, only have the degree characteristic. Nodes with a higher degree than other are called hubs, as they tend to facilitate

communication for distant nodes. Also, in scale-free terms, a more connected node has a higher chance of becoming even more connected. The degree distribution denoted $P(k)$ is defined as the ratio between the number of nodes with degree k and the total number of nodes [276]:

$$P(k) = \frac{N_k}{N} \tag{2.5}$$

where N_k is the number of nodes with degree k. The function describes the probability that a randomly selected node has degree k. For example, a regular mesh, with most nodes having degree eight, will have a distribution $P(k)$ with only a spike at $k = 8$. The more randomness is added to the network connections, the broader the spike becomes. On the other extreme, a fully random network will have a Poisson-like distribution of degrees. Empirical results however, show that many real networks follow a different distribution than the regular Poisson distribution. The nodes tend to be connected like in a scale-free network, thus they obey a **power-law distribution** [25]. The form of this distribution is:

$$P(k) \sim k^{-\gamma} \tag{2.6}$$

where γ is empirically observed to be between 2 and 3 for a power-law specific to social networks. As this form of distribution is not subject to network scale, it is characteristic for **scale-free networks**.

2.2.4. Power-law distributions

Many physical, biological, and synthetic phenomena tend to follow a power-law, as depicted in Figure 2.14. For example, these include fluctuations of financial markets [97], the sizes of earthquakes, craters on the moon, and of solar flares [204], the structure of the internet [87]. Also, in nature, the foraging pattern of various species has a similar distribution [127]. Moreover, the sizes of activity patterns of neuronal populations [143], the frequencies of words in most languages [188], frequencies of family names, the species richness on the tree of life of organisms [9], the sizes of power outages, criminal charges per convict, and many other quantities have been proven to follow a power-law [63].

Scientific interest in power-law relations derives from the ease with which certain classes of mechanisms generate them; the demonstration of a power-law relation in some data can point to specific kinds of mechanisms that might underlie the natural phenomenon in question, and can indicate a deep connection with other, seemingly unrelated systems. In physics, the presence of power-law relations is due to dimensional constraints, while in complex systems, power laws are often thought to be signatures of hierarchy or of specific stochastic processes.

Research on the origins of power-law relations, and efforts to observe and validate them in the real world, is an active topic of research in many fields of science, including Physics, Computer Science, Linguistics, Geophysics, Neuroscience, Sociology, Economics and others [59, 154, 82]. However much of the recent interest in power laws comes from the study of probability distributions. The behavior of these large events connects these quantities to the study of theory of large deviations (also called extreme value theory), which considers the frequency of extremely rare events like stock market crashes and large natural disasters. It is primarily in the study of statistical distributions that the name "power law" is used; in other areas, such as physics and engineering, a power-law functional form with a single term and a positive integer exponent is typically regarded as a polynomial function [64].

The identification of power-laws in data is often solved through graphical analysis. Although more sophisticated and robust methods have been proposed, the most frequently used graphical methods of identifying power-law probability distributions using random samples are Pareto plots, and log-log plots [64]. Log-log plots offer a way to graphically examine the tail of a distribution. This method consists of plotting the logarithm of an estimator of the probability that a particular number of the distribution occurs versus the logarithm of that particular number. Usually, this estimator is the proportion of times that the number occurs in the data set. If the points in the plot tend to converge to a straight line for large numbers on the OX axis, then we can conclude that the distribution has a power-law tail. Examples of the application of these types of plot have been published in scientific journals of the highest class [134].

2.2.5. Average path length

The average path length is one of the three basic measures of topologies. In a network with undirected edges, the minimum distance between two nodes, d_{ij}, is the minimum number of hops needed to reach node j from node i and vice-versa. The diameter of that network is the maximum distance between any two nodes. The average path L is the sum of all paths, divided by the number of paths in the network:

$$L = \frac{2}{n \times (n-1)} \sum_{i \neq j} d(v_i,\, v_j) \tag{2.7}$$

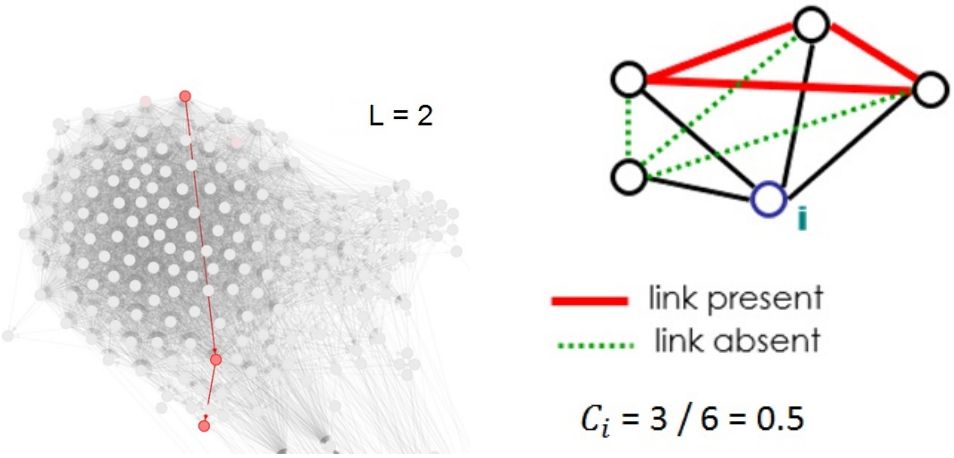
where n is the size of the given graph, and v are vertices. For example, in a network of friends, L is the average number of friends that form up the shortest way connecting any two friends [276]. In a road network, L is the average number of roads a driver has to change in order to get from one city to any other city. A particular aspect is that natural networks, even though having lesser edges than a computer network, still have a very small average path. This is the property known as small world effect found in **small-world networks** [281, 246].

2.2.6. Average clustering coefficient

The clustering coefficient is a measure of the nodes' tendency to cluster together. This can exemplified that in a friendship network, there is a great possibility that one of your friend's friend is also a direct friend. Or reinterpreted: it is very likely that two friends of a person are also friends with one another. Thus, the clustering coefficient can be defined as the ratio between the existing number of links between a node and his friends, and the total number of links that can exist. More precisely, if a node n_i has $deg(n_i) = d_i$ then it has d_i friends. The maximum number of links between all these noes is $d_i(d_i + 1)/2$.

$$C_i = 2\frac{E_i}{d_i \times (d_i + 1)} \tag{2.8}$$

where E_i is the existing number of links between the neighbors of n_i. The average of the coefficients of all nodes in the network is the clustering coefficient C of the network.



(a) (b)

Figure 2.2.: **a**. An example of the average path length L in a graph. **b**. An example of computing the average clustering coefficient C in a graph.

$$C = \frac{\sum_{i=1}^{N}(C_i)}{N} \tag{2.9}$$

or as defined by Luce and Perry in 1949 [171]:

$$C = \frac{3 \times number\,of\,triangles}{number\,of\,connected\,triplets} = \frac{number\,of\,closed\,triplets}{number\,of\,connected\,triplets} \tag{2.10}$$

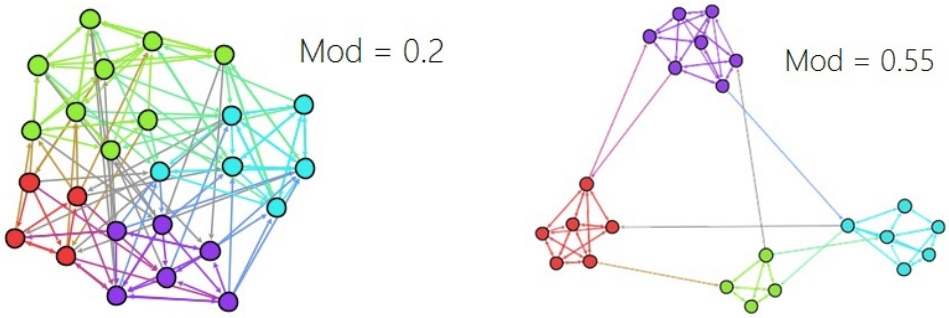
It can be concluded that the maximum clustering coefficient of a network is 1. A network with $C = 1$ is a fully connected graph with point-to-point connections, while a completely random network has $C \sim 1/N$. This is however very small compared to observable networks which have their clustering coefficient satisfy the following relationship:

$$\frac{1}{N} \ll C < 1 \tag{2.11}$$

This means that most networks are neither random, nor fully connected, and thus the triadic closure is a very important aspect of social networks. An illustration of the concepts of L and C is depicted in Figure 2.2.

2.2.7. Modularity and community structure

One of the main drives behind graph modeling of natural or man-made phenomena is to analyze how the different concepts (represented as nodes) connect and cluster together - both numerically and visually [205, 206]. Most network-based approaches yield a certain community structure that



(a) (b)

Figure 2.3.: An illustration of community structure in complex networks. **a**. A graph with a weak community structure, thus a modularity $Mod = 0.2$ **b**. A graph with a visibly strong community structure, thus a modularity of $Mod = 0.55$. All nodes are colored according to the community to which they belong. I have used the community detection algorithm [38] implemented in Gephi [30] for this purpose.

has substantial importance in building an understanding regarding the dynamics of the network. For instance, a closely connected social community will imply a faster rate of transmission of information or rumor among them than a loosely connected community [198]. Thus, if a network is represented by a number of individual nodes connected by edges which signify a certain degree of interaction between the nodes, communities are defined as groups of densely interconnected nodes that are only sparsely connected with the rest of the network. Hence, it may be imperative to identify the communities in networks since the communities may have quite different properties such as average node degree, clustering coefficient, and other centralities [206]. As such, community detection and analysis have received much attention in the last decade [200, 202, 205, 198, 197].

Modularity (*Mod*) is a measure of the structure of networks or graphs. It was designed to measure the strength of division of a network into modules (also called groups, clusters or communities). Networks with high modularity have dense connections between the nodes within communities but sparse connections between nodes in different communities. Numerically, it is defined as the fraction of edges that fall within the given communities minus the expected such fraction if edges were distributed at random. The maximum value for modularity is 1.

There are different methods for calculating modularity [205]; in the most common version of the concept, the randomization of the edges is done so as to preserve the degree of each vertex. An example of two graphs with different modularities have been generated and depicted in Figure 2.3.

2.2.8. Centralities of complex networks

When analyzing graphs, it is often required to extract the most *important* nodes. The so-called indicators of centrality identify such nodes in a graph. For example, applications may include identifying

the most influential persons in a social network, key infrastructure nodes in a computer network, urban intersections in which traffic may congest, or influential genes in transmitting disease. "Centrality concepts were first developed in social network analysis, and many of the terms used to measure centrality reflect their sociological origin" [197].

I have extensively used centrality distributions in my thesis to analyze networks in terms of structural properties and similarity. The most important centrality measures are defined below:

- **Degree centrality**. The simplest type of centrality, it refers to the degree of each node. The degree distribution $P < k >$ is an important aspect when studying empirical networks as they usually possess a uniform, normal or power-law degree distribution. The latter is relevant to scale-free networks which are discussed in section 2.4.5.

- **Closeness centrality**. A connected graph $G = (V, E)$ has a natural distance metric between all pairs of nodes, defined by the length of their shortest paths L. The *farness* of a node v_i is defined as the sum of its distances from all other nodes in V, and its closeness is defined as the reciprocal of the farness [32] as:

$$C(i) = \frac{1}{\sum_j d(j,\, i)} \tag{2.12}$$

- **Betweenness centrality**. It quantifies the control of a node over the communication between other nodes [94], by measuring the number of times a node acts as a bridge along the shortest path between two other nodes. The betweenness of node i, for all pairs of nodes a and b is defined as:

$$Btw(i) = \sum_{a \neq i \neq b} \frac{\sigma_{ab}(i)}{\sigma_{ab}} \tag{2.13}$$

- **Eigenvector centrality**. It is another measure of the influence of a node in a network. Similar to Google's PageRank, it assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

- **HITS (hyperlink-induced topic search)**. It identifies good authorities and hubs for a topic by assigning two numbers to a page: an authority and a hub weight. These weights are defined recursively. A higher authority weight occurs if the page is pointed to by pages with high hub weights. A higher hub weight occurs if the page points to many pages with high authority weights [144].

Figure 2.4 offers an overview of the most common centralities on a mesh network with 200 nodes and 450 edges. The data was generated in Gephi [30] using a plugin developed by the author, and the centralities are computed using existing facilities in Gephi.

2.3. Concepts of social networks

Moving further away from the mathematics and closer to the function of social interactions, there a some unique features which characterize social networks. While complex networks model any kind

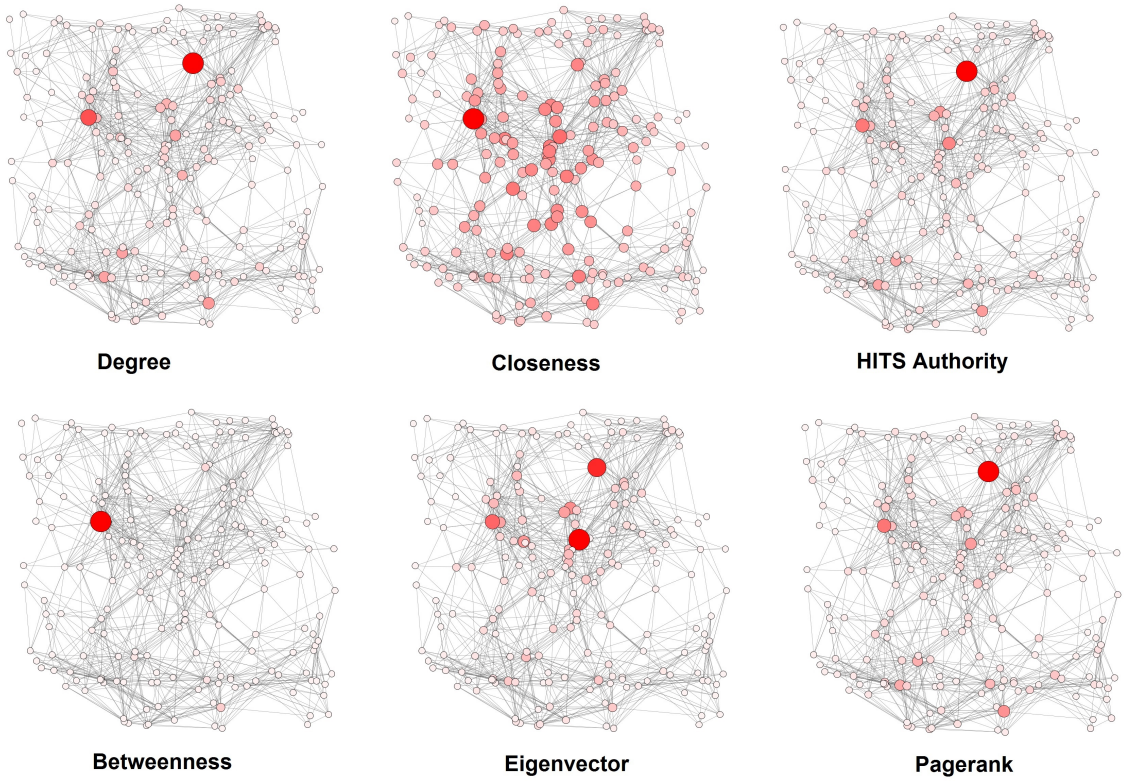


Figure 2.4.: An illustration of graph centralities in a graph with 200 nodes and 450 edges. All nodes are highlighted in red according to their increasing: **a.** degree centrality. **b.** closeness centrality. **c.** authority (HITS). **d.** betweenness centrality. **e.** eigenvector centrality. **f.** PageRank.

of biological, technological, semantic or social concepts into nodes, the branch of social networks deals with one recurring concept: nodes are viewed as actors. These are often individuals, online users, persons. They share knowledge, have an opinion, a public status and a private state, and interact with other actors. This interaction causes dynamics in opinion, or so-called processes.

2.3.1. Agent

An agent, or node, is the term used to describe the network node (Graph theory) or the social actor (Sociology). There are multiple types of agents in literature, and this thesis makes reference to five types of agents defined and/or proposed in the presented social models:

- Normal (regular) agents - this is the bulk of the population which is involved in the opinion dynamics process (i.e. has an opinion during simulation scenarios) and follows the rules imposed by the interaction model that is used. Regular agents interact with any other type of directly linked agents (neighbors), poll their neighbors for opinion, and update their state (opinion) in time. The distribution of the overall opinion in a society is that of the regular agents. All other agent types serve to increase the "realism" of the social mix.

- Unopinionated (null) agents - is a type of agent that does not take part in the interaction process. These agents are uniformly distributed across the society, are connected to all types of agents, but do not have, nor update an opinion. Their state is always set to *NONE*, and can be viewed as actual interruptions in the links of the social topology. If two regular agents are connected, they may influence each other. However, if two regular agents are separated by a null agent, then they cannot pass information from one to another. The null agents are used to increase the realism of simulations, as society rarely has all its individuals implied in one diffusion process.

- Stubborn agents - is an agent type whose opinion is defined before social simulation begins and who does not get influenced by others, i.e. his opinion cannot change. These agents are the sources of social opinion, while the rest of agents absorb and transmit these opinions. Such agents were defined by [2, 291, 4].

- Absurd (contrarians) agents - is an agent type similar to normal agents who builds his opinion by interconnecting with his neighbors, but reacts exactly the opposite in the process of opinion forming [164, 179]. Namely, all opinion influences have the opposite role: if the absurd agent talks to an agent who sustains opinion A over opinion B, the absurd agent will be more inclined to sustain opinion B, as if the normal agent would have sustained the other opinion.

- Random agents - are agents which, based on the interaction model context, interact with other neighbors like regular agents, but always take random opinion. Whether the actual opinion is updated in a random fashion after social interaction, or the tie strengths (e.g. trust, tolerance, weight on edges) is recomputed at random, these agents are used instead of regular agents in null-model simulations to demonstrate that the emergent behavior induced by the interaction model is not a cause of random events. In other words, random agents are used to show that an interaction model behaves substantially different if tested on regular agents, and thus the empirical observations are legit.

An agent can only be of one type throughout the simulation of a society. A small social networks consisting of different agent types is depicted in Figure 2.5.

Features of agent-based simulations

An agent-based social model is built of individual agents, commonly implemented in software as objects (instances of an Agent class). Agent objects have states and rules of behavior, as defined by a social interaction model. Running such a model simply amounts to instantiating an agent population, letting the agents interact, and monitoring what happens, using numerical and graphical (empirical) analysis. As such, solving the equations behind the interaction model simply means running the software simulation (forwarding it in time). Furthermore, when a particular instance of an agent-based simulation, call it S_i, produces result R_i, one has established a sufficiency theorem, that is, the formal statement R_i if S_i [196].

There are several advantages of agent-based computational modeling over conventional mathematical theorizing:

- It is easy to implement the intelligence of agents (through object-oriented programming and polymorphism) and limit their rationality in agent-based computational models.
- It is a simple task to make agents heterogeneous in agent-based models. One can instantiate a population having some distribution of initial states, e.g., opinion, tolerance. That is, there is no need to appeal to single, representative agents.
- Since solving means execution, the obtained results are not only the end equilibrium, but also the whole history of evolution. The dynamics are considered an important part of simulation analysis [279, 198].
- Social and physical processes are difficult to account for mathematically, except in highly simplified ways. However, in agent-based models it is usually quite easy to have the agent interactions mediated by networks [20].

Nonetheless, "the agent-based modeling methodology has one significant disadvantage regarding mathematical modeling. Despite the fact that each run of such a model yields is a sufficiency theorem, a single run does not provide any information on the robustness of such theorems" [20] . The question that arises is if an agent model S_i yields result R_i, how much change in S_i is necessary in order for R_i to become invalid? The only solutions to treat this problem in agent-based computing is through multiple runs, systematically varying initial conditions or parameters in order to assess the robustness of results.

2.3.2. Opinion

Opinion is the basic metric describing an agent's willingness to make a decision. This thesis focuses only on binary decisions [292, 4] like:

- Voting candidate A or B.
- Choosing product X or Y.

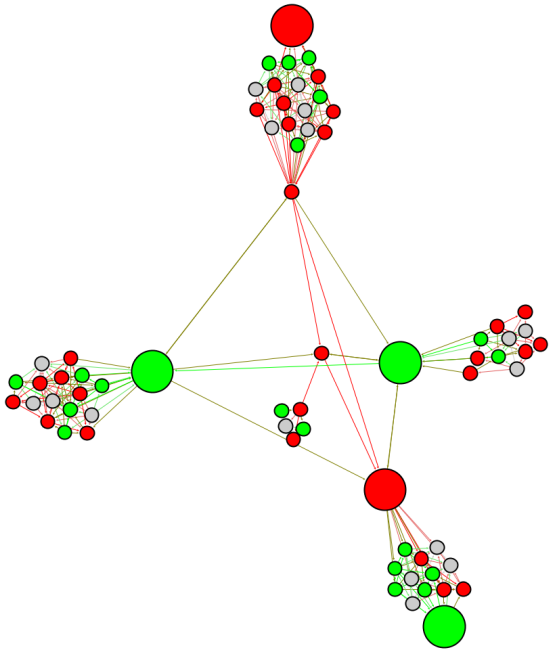


Figure 2.5.: An illustration of agent types in a graph with 68 nodes and 302 edges. All nodes are highlighted according to one of two possible opinions: red or green. Stubborn agents are depicted using larger nodes (3 green, 2 red), null agents are colored in gray, and regular agents have the color of their opinion.

- Buying or not buying a product: X-yes, or X-no.
- Approving with a person or not (court, politics, doctor etc.).
- Supporting a popular belief or not (propaganda, revolution).

Each agent has its own opinion. Personal opinion is changed in the process of communicating with a neighbor (friend), and depends on multiple parameters, as defined by the social interaction model. The way opinion is formed and changes defines the social interaction model of the society.

Opinion is usually a term coined for the information contained by an agent in simulation scenarios. It is the *currency* that is exchanged with other agents and the one which evolves in time. However, agents may hold other information as well, like a trust parameter in each neighbor, a tolerance (as porposed in this thesis), a public opinion, a credibility parameter etc. As a notation, an agent $v_i$will have opinion $x_i(t)$ in time. The value of $x_i(t)$ can be discrete (0 or 1) or continuous (between 0 and 1).

It is worth mentioning that while a node has a sinlge instance of an actor in a social network (one layer), there is recent work which studies multi-layer social networks[173, 39]. These on the other hand, treat each instance of an actor on every layer in an individual way. That is, an individual may be an actor in a particular topological setting on Facebook, another on Twitter, a completely different one at work or with his offline friends, In this case, *opinion* would become a composite vector of states of different context. This thesis does not explore the area of multi-layer networks.

2.3.3. Agent state and network state

The status of an agent is a binary descriptor of its opinion. While opinion is modeled as a fluctuating real number between 0 and 1, status is a measure quantifying what decision an agent would take at an exact moment in time. There are three possible states for any agent:

- **No**: 0 ≤ opinion < 0.5. Depending on the modeled decision, the agent is choosing opinion A or he is simply not choosing to vote, or buy a proposed product.
- **None (not decided/involved)**: opinion = 0.5. An agent will not express any opinion when polled.
- **Yes**: 0.5 < opinion ≤ 1. Depending on the modeled decision, the agent is choosing opinion B or he is simply choosing to vote, or buy a proposed product.

Stubborn agents will always have the same status during simulation. The status of the entire network is the average status of all nodes:

$$S = \frac{\sum_{i=1}^{N*}(s_i)}{N*} \tag{2.14}$$

where each node, regardless of type, accounts with $s_i = 0$ for state No and $s_i = 1$ for state Yes. The process of computing the network state is called *polling*. Undecided nodes are not counted in the polling process, thus the $N*$ symbol.

2.4. Topologies

As any graph based model, a social network can be described through its layout and connection patterns. A network topology is a term used to describe the interconnection pattern of the elements composing the network. Linking elements can be done physically or logically. As social networks describe combined human relationships, knowledge or emotion, the links are purely logical. Analyzing topologies is done with the help of graph theory, a mathematical theory used to describe relationships between objects.

For this thesis, a number of topologies have been reviewed in order to study their effect on social behavior. The used topologies can be divided in two major groups [276]:

- Regular (basic) topologies – the most wide-spread configurations found in technological networks, used mainly in computer science and communication. These networks have simple, symmetric layouts of nodes with clear patterns of interconnectivity. Moreover, regular networks are also called non-complex networks because of the reduced number of nodes (e.g. tens of nodes).

- Complex network topologies – a more comprehensive set of interconnections that bind a large number of nodes. Complex networks are characterized by a large to vast number of nodes (e.g. up to millions) which possess numerous links, both with local neighbors, as well as with distant nodes. Natural and man-made processes have but recently been modeled and studied as networks. In the context of my thesis, this category represents innovative topologies proposed to better model a real society.

The basic topologies, although too simple for social modeling, are used as control (null models) to highlight the impact of the complex topologies, which offer greater fidelity to reality. For example, regular topologies are used to implement computer-like networks to highlight the difference between packet sending and opinion flow; for power grids between cities; for telephone lines and cellular networks etc. Complex topologies are found, for example, in road and airline networks, world trade networks, gene interactions, collaboration between actors etc.

This section aims to present the characteristics of basic and complex topologies.

2.4.1. Lattice or regular mesh topology

The most commonly used technological topology, in which each node acts as a relay for neighboring nodes. This type of network allows routing of information (usually in the form of packets) along a path, from a destination through a source. The visual layout of this topology is not distinctive as any node will have a maximum of 4 (vertical and horizontal only) or 8 (diagonals also) neighbors. The number of edges in a lattice of N by N nodes is $2 \times N \times (N - 1)$. An example lattice, with only horizontal and vertical connections, of size 3x3, is depicted in Figure 2.6 and shows how nodes that are not on a margin have up to 8 possible neighbors.

The connection pattern depicted in Figure 2.6 is the regular mesh, but connections are also possible along the two diagonals, or just along the vertical (with a bounding box), or any other combination as shown in Figure 2.7.

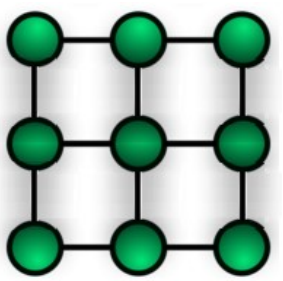


Figure 2.6.: Regular mesh topology: a 3 by 3 lattice.

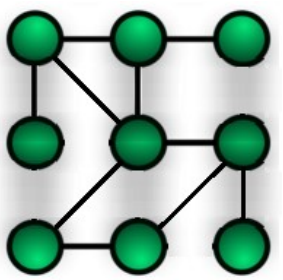


Figure 2.7.: Generic mesh topology.

2.4.2. Mesh topology

Meshes are widely spread topologies, in which each node may be connected to one or more neighboring nodes, within a close proximity. This type of network allows advanced routing of information along a path, from a destination through a source. Any node can have any number of connections (>0), with a total number of connections of $N \times (N-1)/2$.

The mesh is also used as a basis for more complex topologies. Another important mesh alternative is the wrapped mesh in Figure 2.8 which permits marginal nodes to communicate with their symmetrical opposites. This is important because it emulates the globe, so the western most nodes actually meet the eastern most nodes on the other side of the globe.

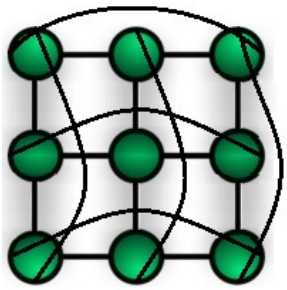


Figure 2.8.: Wrapped mesh topology.

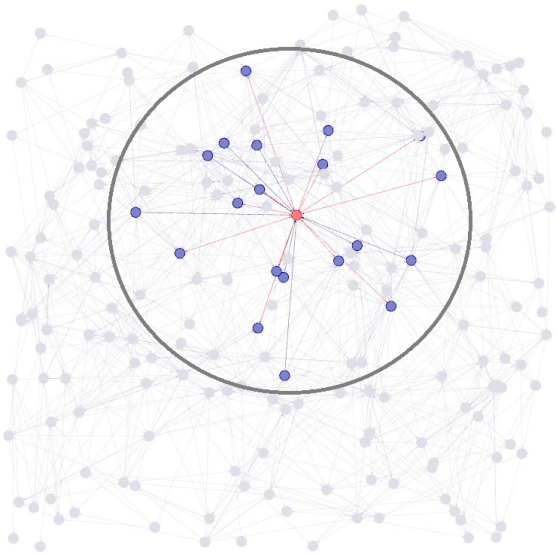


Figure 2.9.: Complex mesh topology highlighting how a node may connect to any number of his neighbors, but within a proximity threshold.

Another example, at a greater scale, is depicted in Figure 2.9 which shows a complex network based on a mesh topology.

The mesh topology is known to have a uniform degree distribution, a high average path length (as there are no ling range links), and a low clustering coefficient (as neighbors connect randomly in their vicinity).

2.4.3. Random topology

Random networks consist of vertices that are randomly connected with a given probability p, regardless of spatial localization. This phenomenon results in the creation of long range links across the network. The topology is constructed by placing nodes in a mesh configuration and then randomly adding a total of 8 random links to each node. Eight was chosen because it is the number of connections of inner nodes in a mesh. On inspection, random networks, as defined by Erdos and Renyi [86], show a dramatic decrease in the average path length, as long-range links are randomly inserted in the network. On the other hand, the clustering coefficient remains low as there is no rule implying that local nodes tie together. Such a network can be seen in Figure 2.10.

The algorithm to generate a random network is presented as follows:

```
Given N nodes and probability p:
for each pair of nodes (n_i, n_j):
  if generateFloat[0,1) < p then
    add edge e_ij between nodes
```

If $p \sim 0$ then the graph remains disconnected, if $p \sim 1$ the graph becomes very dense. In empirical

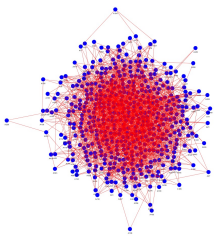


Figure 2.10.: Random Erdos-Renyi topology.

data, good values for p are considered between 0.01 and 0.2 [281, 279].

2.4.4. Small-world networks

Small worlds are social network specific topologies which possesses properties found in real societies. The topology is based on a graph with a generally low amount of interconnectivity, most nodes not being neighbors, but in which the average path length between any two nodes is small. More specifically, as the size N of the network grows, the length L only grows at a logarithmic rate relative N [279].

$$L \sim log\, N \tag{2.15}$$

This characteristic is called a small world property and is found in many empirical networks such as the internet, natural food-chains, business communities, sexual contacts, gene networks, the internet etc. [276, 202, 197]. The main two properties that define a network as being small-world are the average path length L and the clustering coefficient C. Some random networks have been identified as presenting a small-world property but both the average path length and the clustering coefficient are small. Empiric networks maintain a high clustering as the average path length decreases as shown in Figure 2.11. Generating a correct small-world network, as observed in nature, is done by randomizing a regular network's (e.g. mesh, ring) links, so that the clustering coefficient is kept high as the average path length drops. A middle region appears (pink) in which both properties are satisfied.

In order to create such networks Watts and Strogatz have proposed a network generation model that carries their name [281]. The reasoning behind creating such a network topology was that purely random networks do not have two important features observed in real networks.

First, real-world networks generate triadic closures. A triadic closure is a social characteristic between three individuals similar to transitivity in mathematics, in the sense that if A is related to B and B is related to C, then A and C must be linked by a relationship that is at most as strong as the previous two [280]. If this property is not present in a topology, the resulting clustering coefficient is low.

Second, real world networks also contain nodes that act as airport hubs. That is, there is a small portion of nodes within the network that have a much higher degree than others. Moreover, the degree distribution of all nodes must follow a power law distribution, as seen in nature. Instead, random networks, mesh networks or ring networks present an almost linear distribution of node degrees.

The Watts-Strogatz model generation can be described through the following steps:

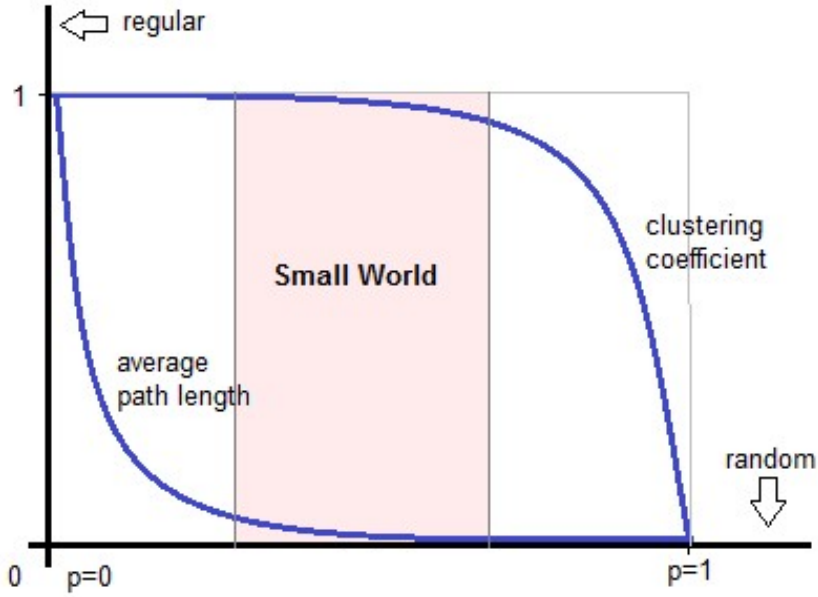


Figure 2.11.: The small-world effect positioned between the regular and random network properties.

1. The algorithm needs a starting ring topology which is iteratively restructured until a balance between the average paths and clustering is obtained. Each node along the ring is connected to the closest $K/2$ neighbors on the left side, and to the closest $K/2$ neighbors on the right side. K is an even variable parameter chosen so that $N \gg K$, where N is the number of nodes in the ring. Another parameter β is chosen so that $0 < \beta < 1$.

2. With the regular ring topology obtained, the reconstruction phase can begin:

---

for each node n_i in N:
  for each edge e_{ij} of n_i:
    remove edge e_{ij} and **create** edge e_{ik} with probability β so that the new edge e_{ik} is not cyclic and is not duplicate.

The resulting topology yields a network with a small average path length, a high clustering coefficient but the degree distribution shows a simple Poisson distribution. As seen in Figure 2.11, a regular topology (i.e. $\beta \to 0$, e.g. ring) has a very high average path length, denoted $L(0)$. A fully random network (i.e. $\beta \to 1$) has a very low average path length, denoted $L(1)$, where:
$L(0) = \frac{N}{2K}$, thus $L(0)$ is linear to N and $L(0) \gg 1$.
$L(1) = \frac{ln\, N}{ln\, K}$, thus $L(1) \ll N$.
The important aspect, however, is that the average path length of a topology between $\beta = 0$ and $\beta = 1$ shows a fast drop as seen in Figure 2.11. The clustering coefficient for a regular network is $C(0) = 3/4$, and for a random network it is $C(1) = K/N$. While the average path length

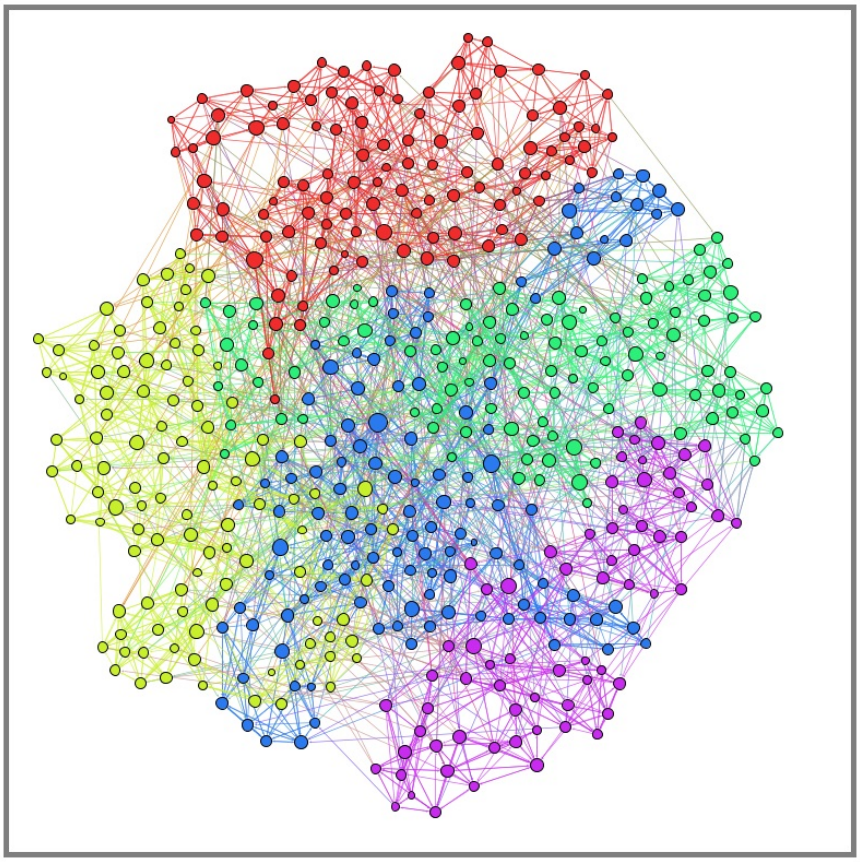


Figure 2.12.: A small-world network generated with the Watts Strogatz algorithm. Nodes are colored based on the detected community.

drops rapidly, the clustering coefficient's drop is delayed until β increases more. This resulting region presents a small-world property, as pictured in Figure 2.11 by the pink area.

As a conclusion, the Watts-Strogatz model proposes an advanced topology that encompasses one important real world property, namely the triadic closure. Such a generated network can be seen in Figure 2.12. Combined with the small average path length, it is an appropriate way to model some classes of real networks (e.g. road maps). However, as it does not create a heterogeneous degree distribution it cannot be used as a stand-alone solution for representing social networks.

2.4.5. Scale-free networks

A scale-free network is another social topology which describes many observable real world networks such as the Internet and relationships within groups. It is a topology based on preferential attachment . This process implies that nodes with a high degree will consequently increase their degree even more, while nodes with small degrees will stagnate in the process of creating new connections. The nodes of a scale-free network follow a power law distribution [25, 295]. A power law distribution

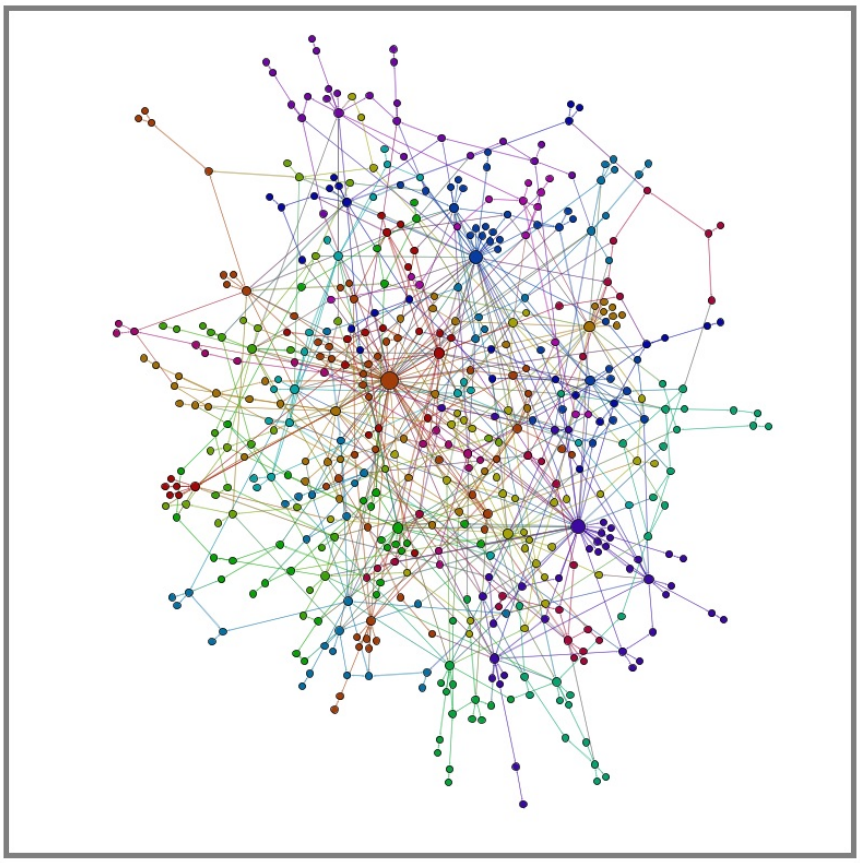


Figure 2.13.: A scale-free network generated with the preferential attachment algorithm of Barabasi-Albert. Nodes are colored based on the detected community.

means that a fraction of nodes, denoted $P(k)$, in the network that have degree k, can be described for large values of k, as $P(k) \sim ck^{-\gamma}$,where c is a normalization constant and γ is a parameter with typical values between 2 and 3.

As can be seen in Figure 2.13, a scale-free network is similar to airplane routes: most nodes have a very small degree, multiple nodes act as local hubs, and very few nodes are hubs for most clusters formed in the network. Although the degree distribution is as found in empirical networks and the average path length is small enough, the clustering coefficient is much smaller than needed. The distribution of the nodes can be seen in Figure 2.14. There a very few nodes with a degree over 10, several nodes with degrees 8-10, multiple nodes with degrees ranging from 6-8, and so on, as most nodes have a degree of 1-4. This kind of distribution is seen in nature (biological networks [276, 195], relationships between students in college, correspondence patterns [209], geographic constraints of groups [210]) as well as in industry (internet, router networks) [23]. A popular example demonstrating the occurrence of scale-free networks in real life is the collaboration of movie actors in films. A study has shown that all actors are linked by a small number of steps (thus a small average path) and also that some actors have been in contact with much more actor that others. One actor connected

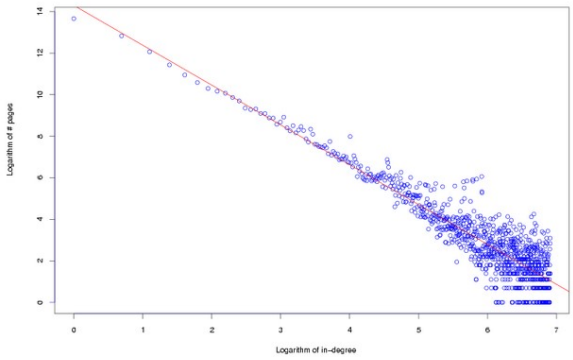


Figure 2.14.: Power-law distribution of node degrees. [42]

with an overwhelming amount of connections is Kevin Bacon, serving as a hub for connecting any other two actors with one another. A popular game named Six Degrees of Kevin Bacon has been released, focusing the fact that no actor is more than 6 steps (hops) away from Kevin Bacon [23]. Another example is the worldwide airline network with major hubs around the world connecting the continents, capital hubs connecting countries and local hubs for regional air traffic. Any two cities around the world can be reached by a small number of flights, but regional flights do not form clusters of interconnected cities.

An algorithm to create such a network was proposed by Albert-Laszlo Barabasi and Reka Albert at the University Of Notre Dame, and constructs the so called Barabasi-Albert model. The model centers around two important topics, namely network growth and preferential attachment. Growth is seen in most natural and synthetic networks, as food-chains evolve or as the Internet grows. The construction algorithm undergoes the following steps:

1. Begin with a random network of m nodes, with $m > 1$ node, in with each node has a degree of at least 1.

2. Each new node n_i is added by trying to connect it to every node m_j in the network with probability p_{ij}.

$$p_{ij} = \frac{k_j}{\sum_m k_m} \tag{2.16}$$

where k_j is the degree of node j in the network and the sum is composed of the degrees of all m nodes in the network. It the node is not successfully connected to at least one node in the network, the process is repeated. It is clear from this algorithm that hub nodes tend to attract more nodes faster, while nodes with low degrees will likely retain their degree, as their probability to be linked to new nodes is smaller.

By analyzing the properties of the resulting network it is clear that the degree distribution conforms itself to the social network requirement. The number of nodes with degree k, for large values of k, and denoted with $P(k)$ is:

$$P(k) \sim k^{-3} \tag{2.17}$$

The average path length of the Barabasi-Albert model increases logarithmic with the network size as [276]:

$$L \sim \frac{ln\, N}{ln\,(ln\, K)} \tag{2.18}$$

However, the clustering coefficient depends on the network size, as it scales with the degree of the node and with the network size:
$C \sim N^{-3/4}$ and $C(k) \sim k^{-1}$
Unlike scale-free networks, the clustering in small-world networks does not depend on the network size [10].

As a conclusion, the Albert-Barabasi model proposes an advanced topology that encompasses one important real world property, namely the power law distribution of its nodes. Combined with the small average path length, it is an appropriate way to model many classes of real networks (e.g. airplane networks, the Internet). However, as it creates a homogeneous clustering coefficient that scales with the degree and network size, it cannot be used as a stand-alone solution for representing social networks.

2.4.6. Advanced complex network topologies

Inspired by the small-world and scale-free topologies, a considerable amount of new networks have been added to literature in the past 10 years, and each of themmay still be classified into one of the two categories: small-world or scale-free. To recreate natural processes with a higher fidelity, there are proposals which add the small-world property to scale-free models [123, 251, 96, 166], or ones that add power-law degree distribution to the small-worlds [135, 57, 152, 274, 296].

A. Watts-Strogatz network with degree distribution (WSDD)

The WSDD [57] is designed by creating a small-world topology (short L and high C) but also modifying the degree distribution of nodes, from a normal distribution to a power law one. This is achieved by first generating a given number of N disconnected communities. Each community is built using the Watts-Strogatz small-world algorithm [281], and the sizes of the communities follow a power-law: there are very few very large communities, and many very small communities. Once each of the N communities are generated, they are connected by randomly selecting two nodes from two different communities, and adding an edge.

The resulting network is one with high C, a relatively low L (due to the inter-community links), an overall power-law $P < k >$ and a high modularity. Such a network is depicted in Figure 2.15a. The generated network has 280 nodes, 4527 edges, 10 distinguishable communities, and a modularity $Mod = 0.799$.

B. Cellular networks

Cellular networks have been proposed as a response to the need for large-scale multi-agent simulations [263]. They are based on the observation of covert networks, like the Al-Qaeda terrorist organization. Cellular networks consist of an arbitrary number of normal-distributed sized cells, with a high clustering, in which a node is chosen as a cell leader. The algorithm to generate such

networks begins by creating N independent mesh-like cells, with one node selected as a leader, per cell. All nodes connect to their cell leader. After this step, all cells are randomly connected with each other to create a small-world network of cell leaders only. This results in a super-network of leaders which are bridges for their respective cell nodes. Thus, any node may connect to another distant node in roughly 3-4 steps: link to local leader, link(s) between leader(s), link from distant leader to destination node.

The resulting network is one with high C, a relatively low L (due to the leader network), and a high modularity. Such a network is depicted in Figure 2.15b. The generated network has 118 nodes, 540 edges, 8 cells (communities) and a modularity $Mod = 0.747$.

C. Holme-Kim (HK) networks

The network model proposed by Holme and Kim [123] stems from the scale-free algorithm of Barabasi-Albert (BA) [25], but adds what the latter lacks: a tunable clstering coefficient. In order to obtain higher clustering, the following steps are followed: we start with a seed network consisting of m_0 nodes without edges; one new node v with m edges is added at every iteration; the edges are added using preferential attachment (as defined by the BA algorithm). In the BA model, the preferential attachment step is repeated for each edge m of each new node v. To solve the problem of clustering (i.e. to obtain a network with high C), the authors propose another step (triad formation step): if an edge is added between v and another node w, then v will also connect to a randomly chosen neighbor of w.

The resulting network is one with higher C (than scale-free networks), a low L (due to preferential attachment), and a low modularity. Such a network is depicted in Figure 2.15c. The generated network has 300 nodes, 959 edges, a non-distinguishable community structure (11 detected) with a modularity $Mod = 0.452$.

D. Toivonen networks

Similar in motivation and in structure to the HK model, Toivonen et al. propose a complex network model that encompasses more features found in empirical data[251]. The authors consider essential characteristics for social networks to include assortative mixing [201, 202], high clustering, short average path lengths, broad degree distributions [13], and the existence of community structure. The algorithm starts with a seed network of N_0nodes. At each step, to add a new node v, $m_r \geq 1$ random nodes are picked as initial contacts, and $m_s \geq 0$ random nodes are picked as secondary contacts for each initial contact. The newly added node v is them connected to the initial and secondary contacts. Identical steps are repeated until a certain network size is reached.

The resulting network is one with high C, a low L (due to random long-range links), and a high modularity. Such a network is depicted in Figure 2.15d. The generated network has 300 nodes, 925 edges, 9 communities, and a modularity $Mod = 0.69$.

E. LFR model

The network model of Lanchichinetti-Fortunato-Radicchi [152] is based around the idea that community structure is essential for synthetic networks realism, and can be used as a benchmark for such algorithms. The proposed algorithm wants to create heterogeneity in the community sizes and

degrees of nodes. Each node is initialyl given a degree taken from a power-law distribution with chosen exponent γ. Each node shares a fraction $1 - \mu$ of its edges with other nodes from the same community and a fraction μ of edges with other nodes outside the community. The sizes of the communities are also taken form a power-law distribution. Initially, are nodes are without community. THen, they are randomly assigned a community; if the size exceedes the node degree, the node enters the community; otherwise, a randomly chosen node is kicked from the community and becomes isolated. The process stops when all nodes have a community.

The resulting network is one with high C, a low L (due to random long-range links), and a low modularity. Such a network is depicted in Figure 2.15e. The generated network has 316 nodes, 2286 edges, and a very low modularity $Mod = 0.157$.

F. Tunable growing graphs

This network model was proposed by Pasta, Zaidi et al. [216, 296] and is based around the idea of community creation. The algorithm starts by initializing c triads which represent independent communities. The communities are connected with one edge between randomly selected nodes to form a connected graph. A new node n_1 is connected an existign node n_2 through preferential attachment; this results in n_1 belonging to the community of n_2. With probability p_t, node n_1 connects preferentially to other neighbors of n_2 forming triads. The same process is repeated for another pair of nodes n_3 (new node) and n_4 (existing node from another community). With probabilty p_c, an edge is added between two preferentially chosen nodes from the communities of n_2 and n_4.

The resulting network is one with high C, a low L (due to preferential links between communities), and a hierarchical community structure. Such a network is depicted in Figure 2.15e. The generated network has 400 nodes, 2221 edges, 8 communities, and a modularity $Mod = 0.441$.

These models have been considered as references for validation in this thesis, but of course, additional network models exists, and they are mentioned in works like [13, 40, 166].

2.5. Social interaction models

The interaction model of a society encompasses the rules which describe how an agent's opinion evolves after each interaction. Whether opinion is changed by factors depending on the communication initiator or on the receptor there are three classes of models [285]:

- **Egocentrism** (inner model): in which opinion is changed only by personal beliefs of that person. These parameters include trust, confidence and tolerance. All of these parameters are internal (hidden) from the outside and they also evolve with the person's opinion.

- **Exocentrism** (outer model): in which the new opinion only depends on the other person's parameters. This parameter is credibility. External parameters are public opinions of a person, shared by all agents initiating communication. These parameters evolve within the receptor but are publicly visible.

- **Hybrid** (combination of both models): which best describes how a person takes decisions. Everyone has personal subjective beliefs about another topic or person (inner), but they are also objective regarding the public opinion facing one topic (outer). There are situations in which decisions are taken by putting a greater accent on emotions (inner) rather than the truth, and

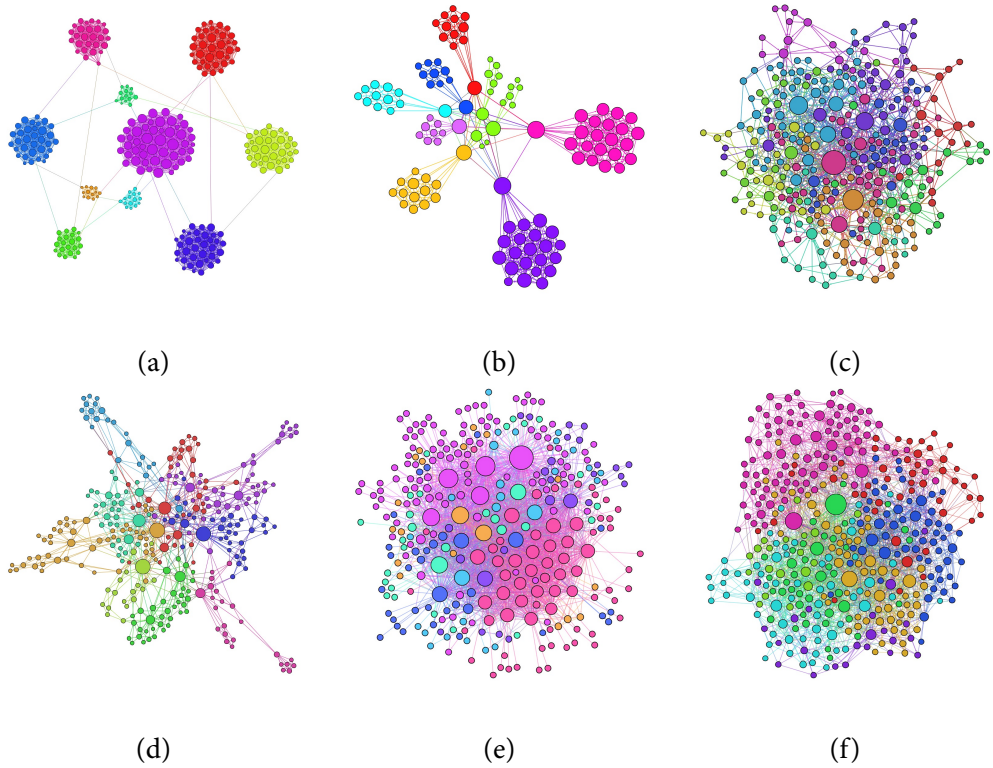


Figure 2.15.: An illustration of complex network topologies. **a**. A WSDD network with 280 nodes. **b**. A cellular network with 118 nodes. **c**. A Holme-Kim network with 300 nodes. **d**. A Toivonen network with 300 nodes. **e**. A LFR network with 316 nodes, **f**. A tunable growing graph with 400 nodes. All nodes are colored according to the community to which they belong, and sized proportional to their degree. I have used the community detection algorithm [38] implemented in Gephi [30] for this purpose.

decisions that take the accepted reality into account (outer) with little or no regard to personal feelings.

This classification is an original proposal based on different works in social psychology [19, 285]. Further parameters have an impact on personal opinion evolution but are currently only proposed for future research: age, education level, financial status, religious belief, health status, randomly appearing problems, risk etc.

While social interaction models mostly stem from studies in (social) psychology, social networks analysis has borrowed a simplified version of human interaction basics, one that can be mathematically modeled and parameterized. This thesis does not porpose to leave the field of computer science and social networks and enter that of psychology, as such, I limit myself to the validated proposals found in recent literature.

In accordance to a landmark survey by Guille et al. [112], the interaction model proposed in this thesis, as well as the state of the art revolves around so-called graph-based predictive models. There are three types of such models:

- With **static** thresholds (fixed from the start of the simulation): q-voter model [133], LCCC model [37], *hard-interaction* model [162], vector-based interaction [237], dual-threshold interaction [54], extended bounded confidence model [225], voter model with biased nodes [74], voter model with friends and foes [167]. All of the previoulsy mentioned models use uniformaly distributed threshold values, as they have no empirical backup. Additionally, some use thresholds from real-world data, but they are still static: information cascade using Twitter[99], asynchronous linear threshold model using social network datasets [234].

- With **pseudo-dynamic** thresholds (the thresholds adapt during simulation, but simply based on the evolution of the graph, not that of the interaction): diffusion model with early adopters [76], and diffusion model with trust [161].

- With **dynamic** thresholds (the thresholds adapt during simulation, based on agent interaction): diffusion model based on opinion evaluation [88] and the tolerance model proposed in this thesis.

As most of the presented literature is based on simple, static threshold models, this thesis wishes to bring an important contribution to literature by creating a truly dynamic (adaptive) interaction model. This original contribution is found in Chapter 6. Additionally, there are some interaction models which make use of special types of agents:

- Stubborn agents or blocked nodes [3, 2, 4, 292, 233] (used so that society never reaches consensus).

- Extremists [75] (the society may reach consensus).

- Contrarians [164] and non-conformists [133] (similar to absurd agents, and the society may reach consensus).

- Media nodes [225] (converge towards the opinion of most followers).

- Advisors [88] (similar to stubborn agents).

2.5.1. The q-voter model

The authors test the q-voter model [133] - a popular extension of the classic voter model [122, 292] - by adding conformist (regular) and non-conformist (exactly opposite behavior as regular) agents. They study the ratio p of conformists versus non-conformists and highlight the fact that opinion stabilizes at a time point t and value M that is dependent on that ratio p; this whole convergence happens as the result of a phase transition. The q-voter model is worth mentioning in line with Acemoglu et al [2, 4]. This work discusses a type of phase transition in opinion formation (stabilization) and opinion stabilization, that coincides with the observations over my proposed model, It shows that my interaction model also behaves in a way that is persistent in many other models.

2.5.2. The LCCC model

The authors analyze two models in their work [37]: the LCCC-model, based on a closed economy model where nodes are given an initial wealth w and they interact pairwise using a saving parameter λ, and the C-model, where nodes have a conviction probability c, that is, they adopt a random proportion c of a neighbor's opinion. This leads to either a symmetric phase or a broken symmetric phase as λ increases. The authors experiment on lattices and propose a modified model in which agents are only influential over the ones with weaker opinion, but agents who agree, do not influence each other.

2.5.3. The *hard-interaction* model

The authors introduce a threshold model [162] based on the degree of agreement d between any two nodes. That is, there exists an agreement d_{ij} for every pair of nodes. If $d_{ij} < \tau$ (how open minded a society is), then the two nodes i, j can interact and d_{ij} decreases (they converge on an average opinion). The authors show there is a phase transition triggered by an increasing τ (open-mindedness) that leads from a radicalized society to a society with consistent opinion. The value of τ is around 0.75 (75%).

2.5.4. The vector-based interaction model

The paper introduces a vector-based opinion model [237], in which every agent holds n opinions about n topics, and updates by coming in interaction with k other neighbors. The influence of each neighbor is a stochastic variable inside $[0, noise]$. The authors show that if $noise \sim 0$, then the outcome strongly depends on the initial state of the society and opinion clusters form; if $noise$ is large, the opinions mix "well", and no determinable clusters can take shape (i.e. a random mixing in time occurs).

2.5.5. The extended bounded confidence model

The paper considers the coexistence of media and of social influence as two separated but interdependent processes [225]. People interact with their neighbors or with the media using the Bounded Confidence Model [75, 19], if the distance between their opinions is below a given threshold σ (tolerance). In turn, the media aims to capture the highest number of followers, hence they change their message by moving toward the value of the media with the highest number of followers. A phase

transition results when increasing the tolerance for: the maximum distance d between two opinions in the network, and for the localization L, which denotes the inverse number of diverse opinions.

2.5.6. The voter model with biased nodes

The authors introduce a probabilistic model [74] based on selecting one of the neighbors' opinion with a probability that is proportional to the closeness of opinion. The paper builds upon the classic voter model by adding biased nodes (they pick one random opinion in their neighborhood). This decision was taken due to a set of real-world experiments that reproduce Ash's studies on conformity [15].

2.5.7. Pseudo-dynamic interaction models

These models have an opinion-triggering threshold value which does change during simulation, but it is simply based on the evolution of the topology, and not on the the interaction itself. There are two such models in the state of the art:

- The diffusion model with early adopters [76]: a threshold-based opinion diffusion model where the size of the relevant neighborhood varies over time. Therefore, the threshold here is represented by the size of neighborhood participating in opinion formation.
- The diffusion model with trust [161]: assumes that an agent is more likely to be influenced by opinions which are close to the present opinion of the agent. However, the fundamental difference from my proposed model is that the influence likelihood is not an internal state of the agent and does not evolve over time due to previous social interactions.

2.5.8. Dynamic interaction based on opinion evaluation

As a close competitor to my thesis proposal, there is work dealing with dynamic threshold values that evolves in time, with the simulation, and are based on node to node interaction.

The authors introduce a diffusion model [88] based on modeling the trustworthiness of the so-called advisors. This model is somewhat similar to those using stubborn agents. Complex factors are taken into account to model trust, and real-world data is used to validate the findings. The big difference is that this model aims at predicting trust such that an e-bay transaction is made or not, not on predicting opinion formation phases (like my work). Nonetheless, this is a model where the thresholds are dynamical, and serves as reference.

2.6. Caveat of creating realistic societies

Having evolved from basic computer network topologies, like the mesh and ring, complex networks have emerged by studying empirical networks in our world. Ranging from natural networks, like food-chains, actor's relationships, protein chains and correspondence patterns [23], to synthetic networks, like the World Wide Web and airplane traffic, these networks have generated interest in engineering around the world [287, 270, 241, 275]. However, better understanding of social networks,

fostered by social, economic and marketing research, has led to the proposal of newer and more advanced topologies which better resemble real networks [276, 63, 273].

The two fundamental networks which serve as a model for social topologies are the small-world network and scale-free network. Also, there are two properties a topology has to encompass in order to be considered a social network: creation of triadic closures [145, 34], in the sense that nodes tend to cluster locally, and a power law distribution of the nodes degrees in the network. Both of the topologies, however, only meet one criterion necessary for modeling a good representation of such networks. The small-world network creates triadic closures, measured by a high clustering coefficient C, along with a small average path length L, and the scale-free network creates the required power law distribution $P < k >$, measured by the degree distribution, along with a small average path length L.

Empirical studies done over a variety of natural and man-made networks have resulted in the definition of several metrics used to describe and measure these networks. Focusing on the metrics explained and measured in Section 2.2, I present an overview of the recent related work in regard to social modeling, as well as highlight why each current social model does not meet the required accuracy. Also in this section, I present an overview and discuss the state of the art statistical methods used in networks comparison in general, and particularly, the ones applied in social network analysis.

2.6.1. Related work

Current research to improve the accuracy of social topologies has been done by combining properties from the two fundamental models previously described with empirical data gathered from various contexts. A first notable study shows the impact of adding a power law degree distribution to small-worlds [57]. The *Watts-Strogatz model with degree distribution* (WSDD) is designed by creating a small-world topology (short L and high C) but also modifying the degree distribution of nodes, from a normal distribution to a power law one. *Cellular networks* have been proposed as a response to the need for large-scale multi-agent simulations [263]. They are based on the observation of covert networks, like the Al Qaeda terrorist organization. Cellular networks consist of an arbitrary number of normal-distributed sized cells, with a high clustering, in which a node is chosen as a cell leader. Further models exist that expand on the conclusions of Milgram's experiment [187]. The *static-geographic model* generates a social network in which links are added between nodes taking the actual distance into consideration: the greater the distance, the lower the wiring probability. The *introduction model* is similar to a small-world network, in the sense of recreating realistic triadic closures. Once a wiring is done between two nodes with probability p_1, one node tries to connect to as many friends of the other node as possible using probability p_2. The *random encounter model* is useful for modeling population dynamics. Each node receives a random 2D movement and connects with a probability *p* to any other node it collides with. *Growth models* are variations of the Albert-Barabasi [10] algorithm and model realistic network growth according to the "rich get richer" principle. Such a model is the *WIW* online platform started as an experiment in 2002 in Hungary [67]. Analyzing the edges between over 45,000 users, the study proves the existence of a high clustering in real social networks and the fundamental role of triadic closures in creating new friendships.

Similar work based on friendship formation proposes the creation of a synthetic network to be used to simulate social interactions in a population for a given geographic space [14]. It predicts

social travel focusing on friendship relationships and the results indicate that the model is able to generate networks that display the same structural properties as in the sample data.

2.6.2. Evaluating the related work

Basic network analysis is done by measuring fundamental graph metrics, and comparison of two or more networks, by doing an individual comparison of each metric independently. While such an approach is useful in trying to capture one specific feature of the network, it fails to create a general overview of the similarity between the analyzed networks [66]. Similar work aimed at comparing the importance of graph metrics concludes that each metric captures specific attributes of the network [35]. It states that further study on the effect of each metric is needed in order to be able to choose a fitting topology according to the requirements of an empirical model.

Comparing real systems is aimed at a deeper understanding of the interaction patterns between these systems [281, 25, 246], and extracting their common properties helps improve the models even further [281, 13, 141]. However, the predominant method of graph metric comparison suffers from limited information [165]. Some notable means of comparison are the distance ratio measure [49], used to compare individual mental models, a comparison from the data analysis perspective [165] and the study of the self-similarity of complex networks (Song, 2005). From a topological perspective there are studies done both in the direction of classifying social network models [141] and of structural pattern detection [215]. These methods however serve a higher level of meta-analysis rather than as measures of similarity.

The statistical methods with which network similarity can be measured are the cosine similarity [249], variance, covariance, Pearson correlation coefficient (PCC) [245], the Mahalanobis distance [174]. Other methods used in network analysis which are adopted from statistics include the T-test and the ANOVA test (analysis of variance). A recent study improves upon the T-test methodology by proposing an alternative geometrical approach called the Characteristic Direction, in order to identify differently expressed genes [61]. There is no single statistical approach used in current research because there is no unified metric that provides normalized values which are tailored specifically to network comparison. Yet, the most intuitive and thus used metrics are the Euclidean distance, Pearson correlation [244] and cosine similarity [249].

3. Network-based modeling of real-world data

Network science, with a little bit of imagination, can be used to model vastly complex phenomena into simple models which can be grasped by the human mind. This thesis revolves around the concepts of realism assessment of social networks, modeling of the underlying structure, and creating a better overview of agent based interaction and prediction. In order to achieve these goals, I have started with the analysis of multiple real-world datasets which I have modeled as graphs. Analyzing the emergent metrics, centralities and community structure that is formed on diverse empirical data, I was able to define reliable and fundamental models which target social networks. In this chapter, I present the essence of modeling social networks based on data collected from collaboration networks (from music industry, and fashion world), and compare it to non-social data from biological and technological networks (from sleep apnea patients, from patients with heart diseases, and also from road and sensor networks). All this serves so that I build a clear image of the different features which lie at the basis of social networks in general.

"Somewhere, something incredible is waiting to be known."

 Carl Sagan

This chapter represents an introductory set of original contributions which lie at the basis of my topological and behavioral modeling. In order to be able to define the parameters for realistic topologies, and understand the principles of social interaction, I have done multiple studies based on empirical datasets, all presented and published in conference proceedings or journals. To discern which graph properties of social networks are relevant, I have done two studies on social empirical data which represent a first of their kind, to the best of my knowledge. The first study models the network of all musicians from the music industry, which I call MuSeNet [258, 29]. In the second study, I have built a network of the female fashion world, which I call FMNet [254]. The innovative approach behind this second study has brought me a best-paper award at the prestigious conference at which it was presented.. Both these topologies have been generated and analyzed using state of the art techniques from SNA. Also, I use a network motif approach to extract topological features of online social networks [261]. I have elaborated this study in appendix B, using my original and validated network comparison approach [257, 256].

Further, to have a clear perspective from a non-social context, I have applied network modeling in network medicine, where I have obtained a plethora of publications, out of which I mention notable results in modeling the risk compatibility network of patients with sleep apnea [186, 265], and modeling the diagnosis compatibility network of patients with heart disease [247]. Finally, I have also conducted research in technological networks, namely to analyze and optimize urban traffic networks [259], and to develop an algorithm for placing relays and a central sink in a wireless sensor network, in order to balance cost versus latency in such networks where communication timing is essential [129]. These cross-discipline studies have brought a substantial overview on how social networks differentiate from other types of complex networks.

3.1. Collaboration in social networks

Starting with the research of Newman and Barabasi, oriented towards detecting community structures and collaborations, a wide variety of social interaction types have been bridged together by scientists in the last decade [200, 199, 26, 205, 207]. One of the incipient contributions which have set out the incentives for further studies shows the small-world organization of such collaborations, with an overall short path between any two nodes in the graph . The node degrees tend to follow a power-law distribution, and the emerging communities create clustering in the network [200]. While some topological aspects are deemed fundamental and present over all collaboration networks, there are many apparent differences in the patterns of collaboration depending on the studied fields. Additionally, it was shown that (optimal energy) force-directed layouts coincide with the modularity measure used for community detection [207]. To this end, I mention some of the latest and most noteworthy literature available in this field.

A fundamental study on large-scale collaborations studies the fast growth of international coauthorships and finds that such networks are self-organising and scale-free, with notable deviations from the ideal power-law[272]. Arching over to economics, a paper studies the structure of oligopolitical markets[109]. Based on the commitment of pair-wise collaborative firms the authors reach similar results in terms of topological properties.

One important property of social network clustering is homophily. Research like[83, 41] describes this phenomenon through the study of racial and ethnic segregation. The results of over 50 years of measuring the impact of segregation in education, housing, and the labor market are based on

friendship and collaboration network modeling. In all these models, homophily plays the crucial role which leads to segregation.

By analyzing the human desire to share information based on their interests, I find in literature the so-called recipe network [250], and even network of Marvel characters [8]. The results are dynamic databases that can be used to make recommendations.

Triggered by Milgram's experiment [187], and derived from the famous statement made by Kevin Bacon himself [90, 280], a whole science was dedicated to this, sparking an interesting concept in the domain of social networks - the Bacon number; this is defined as being the number of degrees of separation any given person has from Kevin Bacon (a particular application of the Erdős number [60, 200] to the Hollywood movie industry). The internet movie database grows yearly with each new movie release, thus, using all of the data may result in networks that are not transparent, and hard to analyze. Therefore, studying only a subset of the IMDB network, more specifically the adult collaboration network [98] has been proved to bring many benefits, for instance by removing any nodes characterized by long-spanning careers and focusing the resulted network more onto the time evolution.

In the field of music, a notable study analyzes the collaboration network of jazz musicians [105, 48, 102]. Some of the presented results include racial discrimination between musicians, and that the division into communities presents a strong correlation with the geographical locations where the bands have recorded – showing that the musicians and the bands network form a collaboration network of jazz musicians. Another study presents an overview of the professional collaborations of whole music industry, in the so-called MuSeNet [258]. The authors explain how the underlying topology of MuSeNet affects the flow of influence and yield for musicians. It is shown that the network fosters a topocratic environment in which the record houses have a bigger-than-expected impact due to their tight clustering and advantageous topological position.

3.2. MuSeNet: a social model of the music artists industry

Motivated by the constantly growing interest and real-world applicability shown in social networks, I model and analyze the network formed by music artists all around the world, which I call MuSeNet. Inspired by similar approaches, I compare the obtained analytic results with generic online friendship models and with the collaboration networks of actors. Together with my collaborators, we are the first to fully create such a network, and by using centrality measures and network motifs, we discover the most influential nodes in MuSeNet. In light of current advances in social networks, I highlight the importance of music producers in terms of meritocracy versus topological positioning, and discuss the differentiation between collaboration networks using a network motif approach. Finally, I show that MuSeNet has a characteristic sociability – a measure which is introduced in this section – in comparison with other empirical networks.

The motivation behind this study is to create incentives for studying the professional relationships of (music) artists around the world, how they form new links based on different attributes (common bands, music styles, genres etc.) and watching this collaboration network evolve with each new node (artist). Through intensive data mining from social media sources, through SNA methodologies, and motif distribution analysis, I have created MuSeNet (Musical Society Network), which represents - to the best of my knowledge - a state-of-the-art analysis of this kind. This study, along with its results allow us to elucidate the mechanisms of driving the emergence of this kind of social phenomenon, and

whether it shares dynamical and structural features with other natural, social processes. Additionally, on more generic scientific grounds, social phenomena like collaborations between musicians/bands or even new/old artists forming a new band (relationship) are an excellent opportunity to understand network formation processes and musical influence dynamics.

Additionally, this section presents a novel perspective on how different artists networks (movies, jazz, all music) can be differentiated using a network motif approach. Moreover, I compare these professional networks to usual online social networks (Facebook, Twitter, Google Plus) and quantify how much they differ using the network fidelity metric [257] (see appendix A). Even though similar in nature, it is shown in this study that all studied social networks have specific properties which make them unique in the real world. I coin this measure though the concept of *sociability* and discuss the real world effects these topologies have on the dynamics inside the networks.

3.2.1. Data acquisition

The database used in this study is obtained from the All Music Guide[1] online digital database. Together with my collaborators, I used this particular database, since at the time this case study was elaborated I considered it to be the most comprehensive [102]. The lack of an API or means of downloading raw data meant that we had to write a script to automatically parse each internal link, in order to retrieve the required information. After running for ~24 hours, it accessed 781 pages, resulting in 19,881 artists (15,501 after filtering) with the following data-set saved into an SQL database:

- ID - internal reference number
- Name - the name of the artist
- URL - the url of the artist, pointing to his/her profile on the All Music Guide website
- Genre - the conventional category that a particular artist identifies with
- Style - (a list of) style(s) an artist identifies with
- Member_of - a list of bands he/she was part of, if any
- Active_period - the time-period reported as active

After collecting the data, I needed to create a graph of musicians, similar to the state of the art methodology [105, 98]. I consider the artists as nodes in my graph and place the links based on compatibility. Particularly, compatibility is defined as the number of common bands two musicians have performed for. The more music bands two nodes have in common, the greater the weight of the link between them. It is to be noted that how one defines compatibility influences the structure of the resulting network. A different layout of ties (e.g. based not on common bands, but on overlapping activity years, gender, music genre, music style etc.) would offer different insights over the same dataset. This study only focuses on analyzing how common bands affect the clustering of artists in a complex network, the other mentioned insights are planned as future work.

Finally, I have created the MuSeNet social network of musicians as a *.gdf* file, a valid input file I can load up in Gephi [30], the leading tool in visualization and analysis of large networks. For the

[1] Cam be found online at http://www.allmusic.com/

purpose of this study I have truncated the weights on the resulting network into an unweighted graph, where a link denotes one/or more common bands between two musicians, and no link denotes no artistic interaction. The reason for using an unweighted graph instead of a weighted one comes as an optimization to balance the interaction phenomena. It is shown to yield more accurate results for the study in terms of determining whether it is a community based on merit (meritocracy) or position (topocracy) [202, 44]. To that end, I have left the attributes of genre, style and active year as parameters for doing the clustering of artists.

3.2.2. Network analysis of MuSeNet

In this section I present the graph metrics and visualizations obtained by applying social network analysis on MuSeNet. In Figure 3.1, I highlight the relevant communities that form over the musical network. Nodes are placed using ForceAtlas2 [132], a force-directed layout algorithm available in Gephi, and are colored according to the community they belong to. The communities are detected using the fast community detection algorithm implemented in Gephi [38]. Such an algorithm was chosen by the authors in light of the existing methodology to break down a social network into clusters and extract their representative features [205]. One of the analytical advantages of social networks analysis lies in the emergent community structure of the network it is applied to. The artists are grouped together by partially overlapping musical genres. The relevant communities that emerge, based on genre, are: pop-rock (24.56%), jazz (16.72%), blues (15.8%), classical (8%), country (5.35), and others. The proportion of music styles is a known fact, but what network analysis unveils are the spatial distribution as well as the overlapping of such styles. As such, the most popular genres are also the ones clustered together, as there are more collaborating artists. The topologically marginal genres are also the ones less popular, like avant-garde, reggae, vocal, or religious, so I can confirm there is a correlation between the communities' center of gravity and their real-world popularity. The further a genre-community is from the absolute center of MuSeNet, the less popular it is, and vice-versa.

As the most dominant music style, Pop/Rock (violet, in Figure 3.1) is very central and also tightly clustered, meaning that artists in this industry prefer to work together with others alike. On the contrary, the second most important genre highlighted by my analysis is Jazz (yellow) which tends to dissipate and overlap multiple styles. In my perspective, this is because of the very nature of Jazz artists to collaborate and create music with other genres. The same conclusion can be drawn for Classical music (green) which, in today's world, implies composing contributions for movie scores, commercials, and melodic lines for other genres. Finally, country music (cyan) shows a similarity to the Pop/Rock community, namely all artists are linked more with each other rather than with others. However, the community has a more eccentric position which I correlate with its popularity.

Figure 3.2 shows the distributions of centralities in MuSeNet. There is a power-law distribution of degree, betweenness, Eigenvector centrality and Pagerank, which is specific for social networks, both empirical and synthetic [10, 276]. Notable is the cluster visible in Figure 3.2c (larger red nodes) which shows that there is a small single dominant community of nodes with very high Eigenvector centrality. On inspection, this community is formed by mature artists who currently own a record studio. The fact that most published music goes through their studio makes them, as a whole, the central community in MuSeNet.

Referring to the idea of "meritocracy versus topocracy" discussed in a recent study by Borondo et

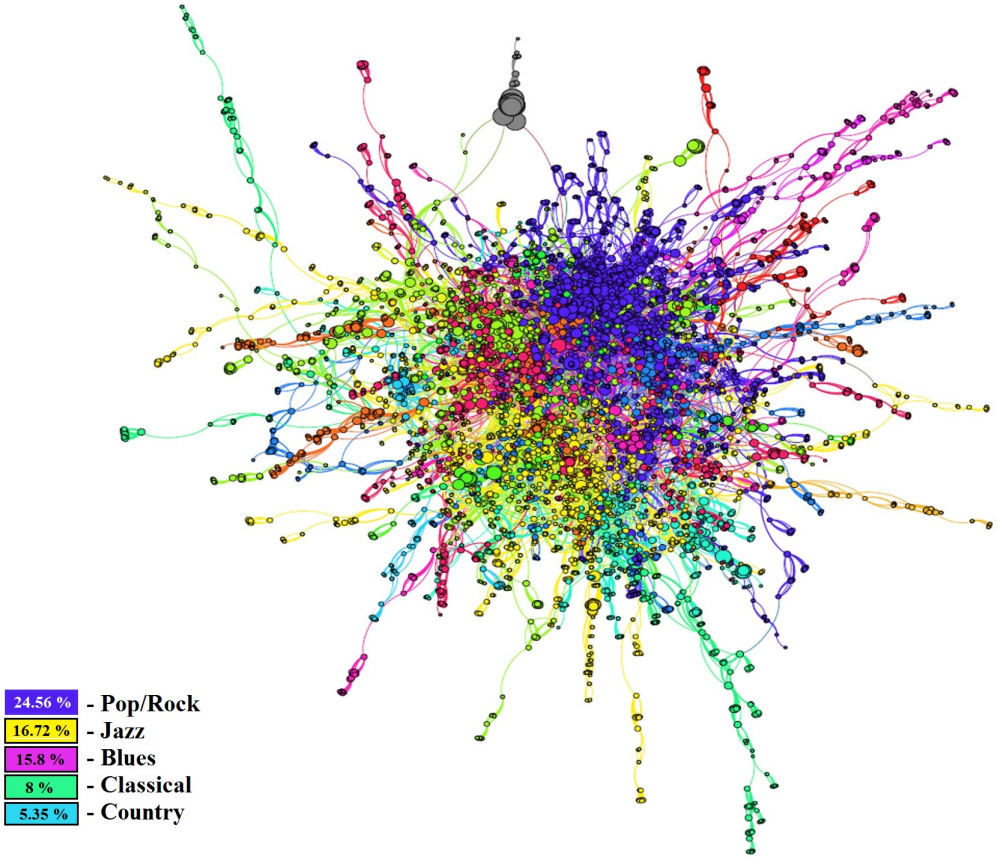


Figure 3.1.: Graphical overview of MuSeNet (generated in Gephi). Each musician is a node in the graph, connected with another node if there has been at least one artistic collaboration with that node. After applying the ForeAtlas2 [132] layout and community detection, nodes can be colored by highlighting the distinct musical genre-communities.

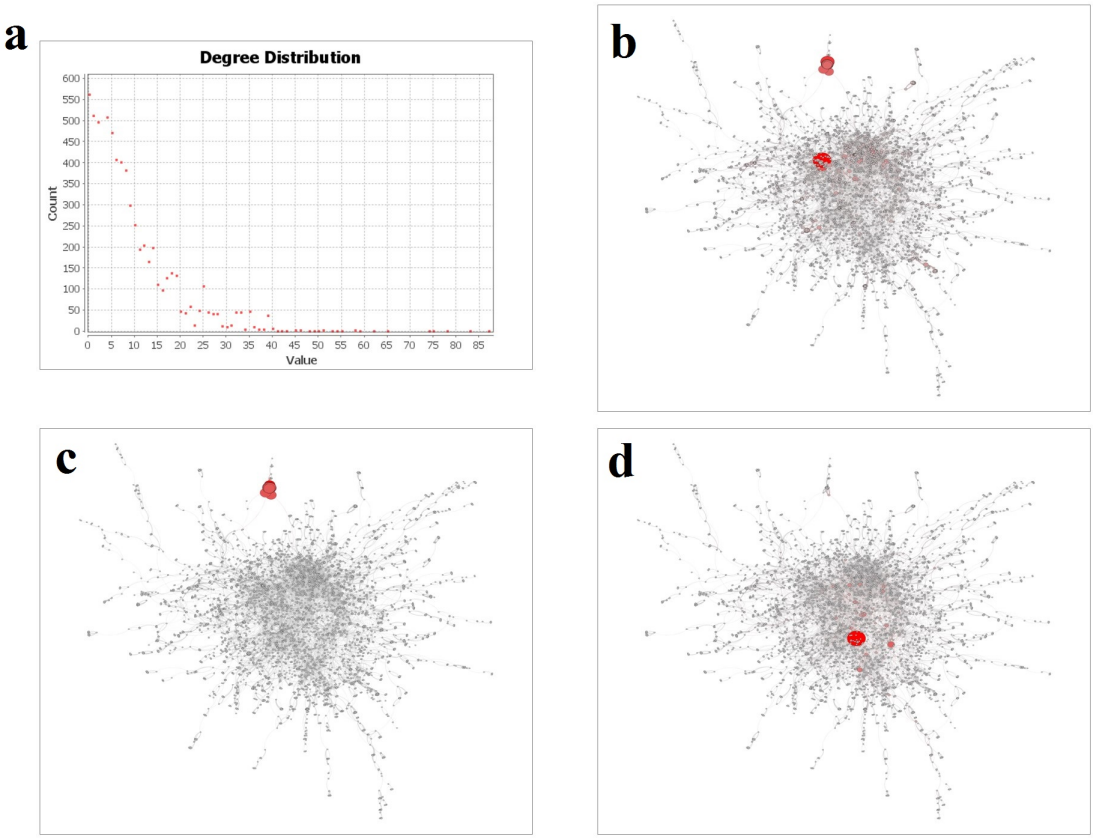


Figure 3.2.: Graphical overview of complex network measurements on MuSeNet. The nodes highlighted in red in each figure highlight one of the three measured centralities: **a.** Power-law degree distribution, **b.** Degree centrality, **c.** Eigenvector centrality, **d.** Betweenness centrality .

al. [44] this community (shown in gray in the upper part of Figure 3.1 and in red in Figure 3.2c) is the one that thrives mostly in the topocratic environment of the music industry, making the most out of its influence in the music industry. Moreover, this real-world influence is replicated in the graph.

Tables 3.1 and 3.2 show the top 5 artists with the highest centralities in the music industry. I have measured all four centralities since they highlight different aspects of importance in a network. The highest degree musician is Greg Errico, an artist and producer who's resume spans across the most important musical genres, until today. He was member of the "Sly and the Family Stone", and performed in Rock, Jazz, Fusion, along with David Bowie, Santana, Larry Graham and others.

On the other hand, betweenness depicts importance in terms of interaction control. Dave Grohl, a member of Foo Fighters and Nirvana, lies the crossroads of most collaboration paths between all other artists. Eigenvector centrality highlights members of the mentioned producer-cluster, with Greg Errico, Alphonso Johnson (etc).

In link analysis, where Pagerank is normally used, a web page will have high Pagerank if it has some

Table 3.1.: Musicians with highest degree and betweenness centralities.

Artist	Degree
Greg Errico	81
Alphonso Johnson	79
Dave Walker	67
Don Airey	65
John Wetton	62

Artist	Betweenness
Dave Grohl	.0124
Josh Freese	.0091
Chris Shiflett	.0084
Lu Edmonds	.0075
John Wetton	.0073

Table 3.2.: Musicians with highest Eigenvector and Pagerank centralities.

Artist	Eigenvector
Alphonso Johnson	.764
Greg Errico	.754
David Brown	.689
Graham Lear	.657
Neal Schon	.652

Artist	Pagerank
Greg Errico	2.925
John Wetton	2.777
Lu Edmonds	2.7
Jimmy DeGrasso	2.672
Alphonso Johnson	2.641

combination of high in-links, low out-links, and specific in-links from other high ranking pages. In the world of musicians, these artists with high Pagerank like Greg Errico, John Wetton, and Lu Edmonds have most likely been influenced by either a lot of people, a few very important people, or some combination of the two.

Finally, similar to the IMDB study which denotes Kevin Bacon as the most influential node in the Hollywood actor network, I find Dave Grohl as the "Kevin Bacon of the music industry". This aspect is clearly visible in Figure 3.2d, where I show the betweenness distribution, a classical method of computing influence. Dave Grohl is an American rock-musician, multi-instrumentalist, singer, songwriter, producer and film director. He is best known for being the lead vocalist, guitarist, main songwriter and founder of the band "Foo Fighters", drummer and song-writer of "Nirvana", "Them Crooked Vultures", "Queens of the Stone Age" etc. He has also performed session work as a drummer for a variety of other bands/artists, like "Garbage", "Nine Inch Nails", "David Bowie", "Paul McCartney", "The Prodigy", "Slash", "Iggy Pop", "Tenacious D", "Lemmy", "Stevie Nicks" etc.

MuSeNet can further be analyzed from different perspectives, a process I look forward to as a future direction.

3.2.3. Defining the sociability of complex networks

The comparison is done using the topological metrics which are specific for complex networks [246, 10, 276, 197]. These values are represented in Table 3.3: average degree (AD), average path length (L), average clustering coefficient (C), modularity (Mod), graph edge density (Dns) and graph diameter (Dmt). These metrics have the power of characterizing a complex network and offer valuable insight [276].

The numerical results from Table 3.3 show what I call the "sociability" difference between the three

Table 3.3.: Relevant measurements of average degree (*AD*), average path length (*L*), clustering coefficient (*C*), modularity (*Mod*), density (*Dns*) and diameter (*Dmt*) on each empirical network.

	AD	*L*	*C*	*Mod*	*Dns*	*Dmt*
Facebook	22.23	2.34	0.256	0.577	0.005	7
Twitter	12.39	2.68	0.239	0.28	0.054	7
Google Plus	12.15	3.9	0.404	0.44	0.035	12
Jazz	27.7	2.23	0.633	0.441	0.141	6
IMDB	113.5	1.55	0.996	0.476	0.062	4
MuSeNet	13.18	7.64	0.884	0.844	0.002	23

types of collaboration networks. Interestingly, the Facebook model is situated at an average level of sociability (i.e. metrics all centered on empirically representative values [257, 141]), while the IMDB actor network proves to be more sociable (i.e. significantly greater *AD,* shorter *L*, higher *C*, higher *Dns*, and shorter *Dmt*), and MuSeNet the least sociable. From a social perspective I explain the differences in the following way. Facebook users (i.e. usual persons) interact and create new friendships at what I call a normal rate. Actor's everyday job, however, relies on playing in movies with many other actors, and there are almost always different ones, as the casts for movies are very broad. This makes their network very clustered and thus seems more sociable, on my terms. On the other end, music artists do not usually create art (work) with many others. They rely on their own band of ~5 members, and not more then on the other artists from their own genre. This makes links in MuSeNet less dense, clustering very high and the community structure powerful. I consider this to be a "non-sociable" network. Twitter and Google Plus networks, like Facebook, also situate themselves around the moderate-sociable area, while Jazz musicians - interestingly - share the greater sociability of the actors. The explanation for this phenomenon can be seen in MuSeNet itself, as Jazz musicians work with many artists, and foremost with the majority from their own genre.

To quantify the discussed aspect concerning sociability I model the S-metric which expresses the so called sociability of any given complex network. It is imagined to take into consideration the basic graph metrics (also used in this study, e.g. Table 3.3) and compare them to a reference model. In this study I use the online social networks models distribution of metrics as the reference, and compare the metrics of each other collaboration network to them. First, I normalize the offset from the reference value of each metric, then I either add (direct proportional) or subtract (indirect proportional) the resulting normalized values. Thus, I define sociability as:

$$S_i^j = \sum_{i=1}^{6} \left[k_i \times \left(m_i - m_j \right) / m_j \right] \tag{3.1}$$

where S_i^j, the sociability of network i towards reference model j, is the sum of the six normalized metrics: average degree ($k_1 = +1$), average path length ($k_2 = -1$), average clustering coefficient ($k_3 = +1$), modularity ($k_4 = -1$), density ($k_5 = +1$) and network diameter ($k_6 = -1$). The signs (+/-) of the metrics reflect if the particular metric is direct (*AD, C, Dns*) of indirect (*L, Mod, Dmt*)

Table 3.4.: Sociability of the collaboration networks compared to Facebook, Twitter and Google Plus.

S	Reference models		
	Facebook	Twitter	Google Plus
Jazz	29.34	4.23	5.80
IMDB	19.33	11.62	11.76
MuSeNet	-3.56	-4.34	-2.46

Table 3.5.: Network fidelities φ of the three collaboration networks (rows) towards the six used references (columns). A higher value $0 \leq \varphi \leq 1$ denotes a higher similarity.

φ	Reference models					
	FB	TW	GP	Jazz	IMDB	MuSeNet
Jazz	.647	.595	.615	-	**.672**	.517
IMDB	.472	.535	.537	**.66**	-	.472
MuSeNet	.486	.451	.574	.491	.479	-

proportional to a more sociable network. As I have three elements in the sum contributing with +, and three with -, I can simplify equation 3.1 to:

$$S_i^j = \sum_{i=1}^{6} (k_i \times m_i/m_j) \tag{3.2}$$

$$k_1 = +1,\ k_2 = -1,\ k_3 = +1,\ k_4 = -1,\ k_5 = +1,\ k_6 = -1$$

Thus, the sociability of the collaboration networks using the Facebook model as a reference is given in Table 3.4. The Facebook model compared to itself will have a sociability $S = 0$. Any model that is considered as less sociable will have $S < 0$, and all models that are more sociable in terms of their graph metrics will have $S > 0$. In Table 3.4 I can see that MuSeNet is indeed on the "unsociable" side, while Jazz and IMDB are more sociable. Even though the S-values change once I change the reference model (Facebook, Twitter, Google Plus) the scale and signum of the values remain the same.

Table 3.5 presents the fidelity values of each collaboration network when compared to the online social networks (FB = Facebook, TW = Twitter, GP = Google Plus) and to themselves.

The results show a low similarity between all collaboration networks and each online network (45-65%). This can be explained because of the sociability difference - low and high, compared to the moderate one of the reference models. On the other hand, the metric comparison supports my sociability evaluation as it shows the IMDB and Jazz networks - both described a highly sociable - much more similar (67%) than compared to MuSeNet (<50%).

3.2.4. Discussion

This particular study has presented a state of the art analysis of the the whole music artists network. Similar to the study of IMDB actors, and Jazz musicians, I can conclude that certain artists have higher centrality indices. Like other complex networks, MuSeNet has the same properties: it is scale-free (meaning that artists' connectivity distributions are in a power-law form), and has a high degree of centrality [276]. I have highlighted the sociability of three networks through graph metrics. MuSeNet is a more closed network than IMDB and other usual friendships because music artists do not usually work with many others, since they rely on their on band and associated acts; links are also formed at a much slower rate, compared to the Facebook model.

In light of the study which finds Kevin Bacon as the most influential node in the Hollywood actor network, I find Dave Grohl as the "Kevin Bacon of the music industry". Moreover, I analyze MuSeNet from the perspective of other centralities as well, finding artists like Greg Errico to have the highest degree and Pagerank, and Alphonso Johnson to have the highest Eigenvector centrality. A second important empirical observation is the existence of a small single dominant community of nodes with very high Eigenvector centrality. This is the community formed by mature artists who currently own a record studio and through who's studios most music goes. This ecosystem mostly thrives because of the topocratic environment of the music industry.

With the broader perspective of social networks analysis - to better understand and model complex networks - in mind [56, 276, 82, 141], the obtained results pave the way for better understanding the particular concepts of social collaboration. Motif-based analysis has but recently been adopted from Systems Biology into social analysis and, in this study, I have shown how it can be used to numerically express the characteristic aspects of collaboration networks.

3.3. FMNet: modeling physical trait patterns in the fashion world

Driven by the ever-growing interest and real-world applicability shown in social collaboration networks, I have gathered data from Fashion Model Directory, the largest fashion model database. As such, I model and analyze the network formed by female fashion models all around the world, which I call FMNet. Inspired by similar approaches in the actors and music industry, I compare the empirical results with Facebook, Twitter, and Google Plus online friendship networks. As a first study of its kind in the fashion world, I create a network based on physical similarities, and by using centrality measures and network motifs, I prove that FMNet has all the properties of a social collaboration network. I discover and explain role of the most influential nodes (in terms of betweenness centrality) and communities (in terms of eigenvector centrality) in FMNet. The physical patterns found in this study offer a better understanding over the evolving trends in the fashion world.

3.3.1. Motivation and impact

The motivation behind this study is to create incentives for studying the professional relationships of (female) fashion models around the world, how they form new professional links, and how they correlate from the perspective of common physical traits using a set of common attributes (hair color, eye color, height, age). Through intensive data mining from social media sources, through SNA methodologies, and motif distribution analysis, I have created FMNet (Fashion Model Network),

which represents - to the best of my knowledge - a state-of-the-art analysis of this kind. This study, along with its results allow us to elucidate the mechanisms of driving the emergence of this kind of social phenomenon, and whether it shares dynamical and structural features with other natural, social processes. Additionally, on more generic scientific grounds, phenomena like similarity between fashion models forming particular physical trait clusters are an excellent opportunity to understand network formation processes and the influences pertaining to the fashion world. The results obtained through visualization of force-directed layouts [207] may pave the way for creating a recommender system which fashion creators and agencies may use to assign models into collaboratng with each ther for specific brands.

I start by presenting an analytical breakdown and interpretation of relevant graph metrics, centrality distributions, and community structure. Through this analysis I show that the proposed similarity network showcases typical properties of social collaboration networks. I then use the graph model the to highlight emerging trends in fashion.

Additionally, this paper presents a novel topological assessment on how different collaboration networks (movies, music, citations) can be differentiated using a network motif approach. Also, I compare these networks to reference online social networks (Facebook, Twitter, Google Plus) and quantify how much they differ using the network fidelity metric [257] (see appendix A).

3.3.2. Data acquisition

The database introduced in this paper is obtained from the Fashion Model Directory (FMD) online database[2]. FMD consists of information about fashion models, modeling agencies, fashion labels, fashion magazines, fashion designers, and editorials. It was first published online in the year 2000 and is currently considered the IMDb of the fashion industry, being the largest database of its kind. It includes over 10000 female fashion models, 1400 designers, 2000 fashion brands, 1700 magazines and many other fashion related information.

I used this particular database, since at the time this case study was elaborated I considered it to be the most comprehensive. The lack of an API or means of downloading raw data meant that I had to write a script to automatically parse each internal link, in order to retrieve the required information. Because of the delay introduced by the server response for each page access, I designed a multithreaded Java script that was able to complete the crawling in less than one hour, instead of >24h. My script accessed the profile pages of each fashion model and retained those entries with complete information, resulting in 9477 female models with the following dataset saved into a local database: *Name* - the name of the fashion model, *Nationality* - the model's current nationality, usually based on country of residence, *Birth_Year* - the year of birth, *Hair* - the model's natural hair color, *Eyes* - the model's natural eye color, *Height* - the model's height in centimetres, *Agency* - a list of agencies for which the model has worked, *Advertisement* - a list of fashion brands which the model has worked with, *Cover* - a list of magazine covers which the model has posed for.

3.3.3. Network analysis of FMNet

In this section I present the graph metrics,centrality distributions, community structure interpretation, and visualizations obtained by applying SNA on FMNet. In Figure 3.3, I highlight the relevant

[2]Available online at http://www.fashionmodeldirectory.com/

Table 3.6.: Relevant correlations (%) between eye color, hair color, fashion agencies, and fashion model origin. The acronyms for agency headquarters are: Milan (Mi), Barcelona (Ba), Sydney (Sy), Paris (Pa), New York (NY).

Eye color	Hair color	$corr\%$	Agency	Origin
Blue	Blonde	56	Mi, Ba	N-Europe
Green	Brown	59	Ba, Mi, Sy	N-, E-Europe
Brown	Brown	76	Mi, NY	E-Europe
Black	Black	53	Mi, Pa, NY	Asia

communities that form over the fashion model network. Nodes are placed using ForceAtlas2 [132], a force-directed layout algorithm available in Gephi, and are colored according to the community they belong to. The communities are detected using the fast community detection algorithm implemented in Gephi [38]. Such an algorithm was chosen by the authors in light of the existing methodology to break down a social network into clusters and extract their representative features [205, 207]. One of the analytical advantages of SNA lies in the emergent community structure of the network it is applied to. The fashion models are grouped together by partially overlapping physical traits: similar gradient of eye color, hair color, or similar height etc. I obtain a total of 9 communities. The relevant communities that emerge are mainly based on the clustering of similar eye color and hair color gradients. The proportion of models with particular physical features is a known statistic, but what network analysis reveals, is the spatial distribution as well as the overlapping of such features. As such, the most popular female model features are also the ones clustered together, as there are more *collaborating* nodes. The topologically marginal features are also the ones less sought after, like red hair, gray eyes, so I can confirm there is a correlation between the communities' center of gravity and their popularity in the fashion world.

By analyzing the layout of the 9 resulting communities, I can support the claims through the visual observation presented in Figures 3.4a and b. Namely, homophily plays such a role that each node is placed in the vicinity of other nodes with the same five chosen traits. The most obvious visual classifications are the ones of eye color: from black and brown, through green and dark blue, to light blue (see Figure 3.4a); and for hair color: from black and dark brown, through dark blonde and blonde, to light blonde (see Figure 3.4b). Through overlapping of the node properties I notice there are a few notable correlations between some physical features. These results are given in Table 3.6.

From the point of view of graph centralities I have measured the following: degree, betweenness, pagerank, and eigenvector centralities [276]. These offer different insights over the most influential nodes in a graph. However, in the context of a physical similarity graph, I interpret the influence of a node as to how impactful a certain combination of physical features is. Particularly on FMNet, a node with high centrality is a fashion model which holds a certain combination of physical traits which stand out as a reference for the fashion world.

I consider FMNet to be a representative collaboration network, as the metric distributions show. Figure 3.5 displays the power-law degree distribution which is representative for collaborations, and social networks in general [25, 246, 26, 24]. The scale-free property of FMNet is valid at community level as well, as can be seen in communities 1 and 5 in the same figure. This observation supports the argument that hub formation is present even in a non-social collaborating context, namely physical similarities between humans. Also, it shows that similarty networks, like FMNet, have emergent

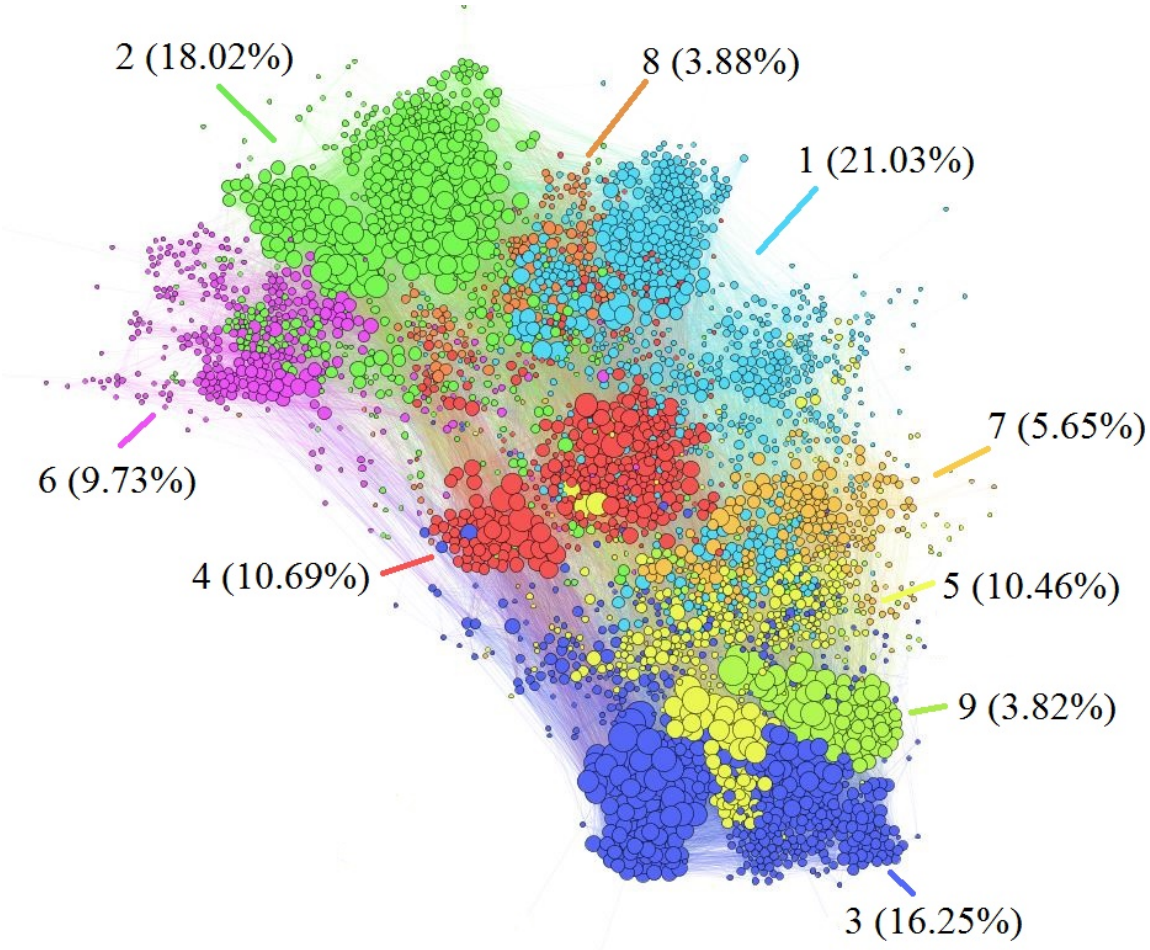


Figure 3.3.: Graphical overview of FMNet (generated in Gephi). Each fashion model is a node in the graph, connected with another node if there are at least three common physical traits with that node. After applying the ForeAtlas2 [132] layout and community detection, nodes can be colored by highlighting the distinct physical pattern communities.

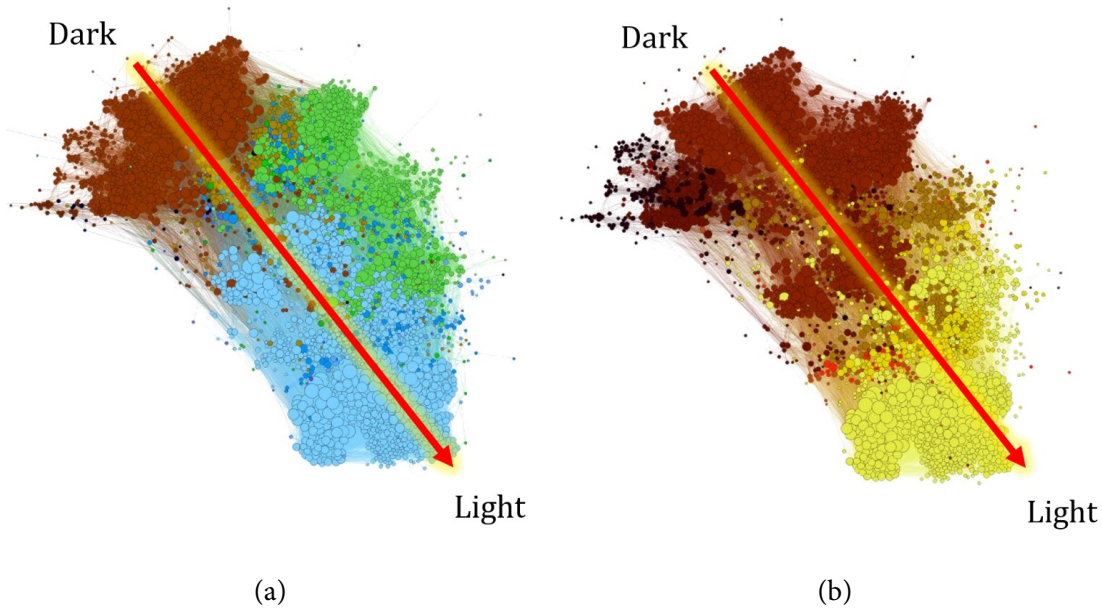


Figure 3.4.: The physical similarity network highlighting the two main physical traits: **a.** Dark brown to light blue eye color gradient . **b.** Black to light blonde hair color gradient. All node colors correspond to the eye and hair colors.

structures like social netwoks. I consider this to be valid because similarity networks over completely random feature vectors show complex network properties [242, 212].

Further, I highlight the distribution of nodes with high centrality in terms of degree (Figure 3.6a), betweenness (Figure 3.6b), pagerank (Figure 3.6c), and eigenvector (Figure 3.6d). From each measurement, the same highlighted regions appear to have a higher centrality: the lower region (communities 3, 5), the upper region (community 2), and a middle region (community 4). These three regions have the following trait patterns: blonde and blue eyes, brown hair and brown eyes, respectively blue eyes and brown hair. However, out of the 4 mentioned centralities, I focus on betweenness centrality to offer us insight regarding the most influential physical patterns in the fashion world. This metric is often used in SNA to measure influence of nodes in communication [276], and collaboration [258].

Betweenness centrality highlights a few fashion models which stand out as nodes with a much higher centrality than all others (see central region of Figure 3.6b). These are what I consider to be the most impactful fashion models in terms of setting a standard for the fashion world. The models are given in Table 3.7. Kelsey Gerry, the node with the highest betweenness centrality, is a typical fashion model in the sense that she has blue eyes (the most common eye color), brown hair (the most common), is born in 1989, and is 175cm tall. An interesting observation for most nodes with top centrality, is that they do not work at the most renowned or largest modeling agencies (i.e. Paris, Milan, New York).

In light of the studies done over the IMDb network, which show Kevin Bacon as the most influential node in the actors collaboration network, and over the music industry network, which consider Dave Grohl to be the most influential node in MuSeNet, I conclude that the female model Kelsey Gerry is

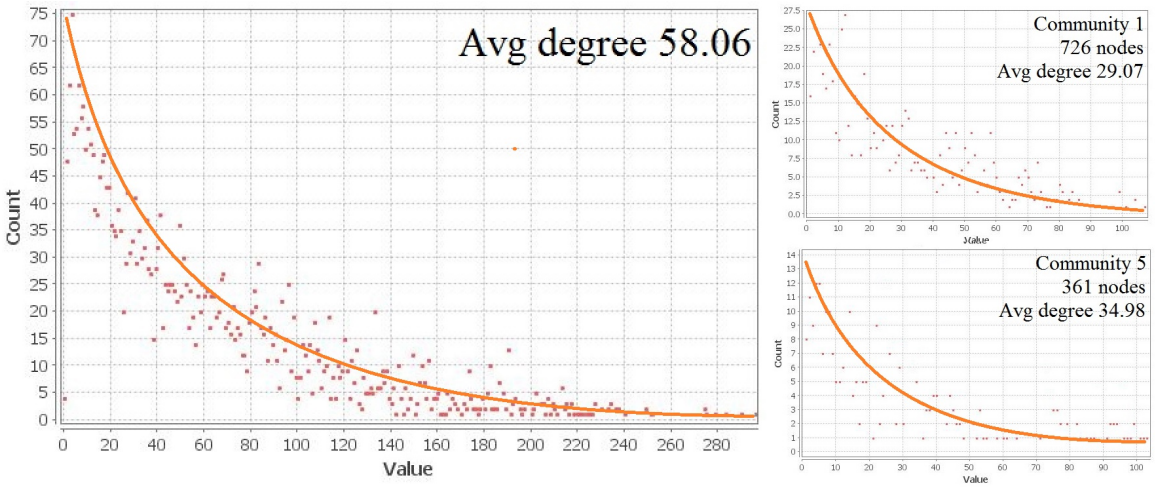


Figure 3.5.: Power-law distribution of node degrees. FMNet showcases the scale-free property specific to collaboration networks. This property is also present in each community.

Table 3.7.: Fashion models with the highest betweenness (Btw) centrality.

Model name	Hair color	Eye color	Agency	Btw
Kelsey Gerry	brown	blue	Berlin	39.8K
Kelley Havey	blonde	brown	Chicago	38.4K
Michelle Lombardo	brown	blue	LA	37.1K
Sherita Dehon	brown	brown	London	35.2K
Emily DiDonato	brown	blue	Sydney	35.2K

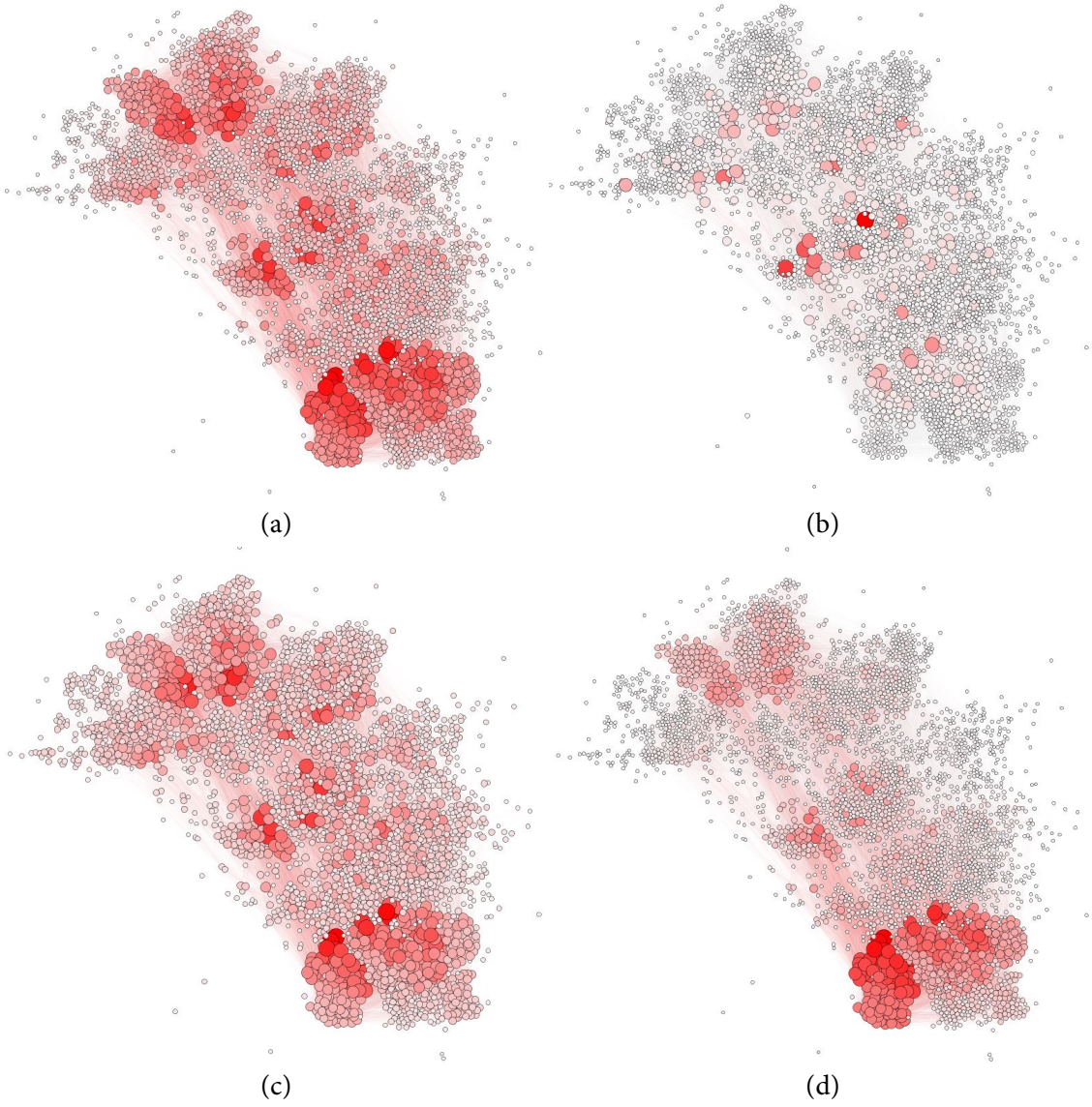


Figure 3.6.: The physical similarity network with each relevant graph metric highlighted through node color (red intensity) and node size. **a.** Degree distribution. **b.** Betweenness centrality. **c.** Pagerank. **d.** Eigenvector centrality.

the central node which sets a physical trait standard for FMNet.

Nonetheless, if I study the eigenvector centrality distribution over FMNet, I notice a string clustering of influence in the lower part of the graph. Community 5 holds an overwhelmingly high proportion of nodes with high eigenvector centrality, so I put this observation into the perspective of a landmark study which discusses the impact of topology versus merit. Referring to the idea of "meritocracy versus topocracy" discussed in a recent study by Borondo et al. [44] this community (shown in red in the lower part of Figure 3.6d) is the one that thrives mostly in the topocratic environment of the fashion model industry, making the most out of its influence in the fashion world. This community consists only of models with blue eyes and blonde hair; the fist model with a different physical feature is outside the top 100 in terms of eigenvector centrality. I consider this analytical observation to be well correlated with the real-world popularity of female models with these two mentioned physical features, and moreover, they also seem to attract other young models with the same features more than any other combination of traits.

FMNet can further be analyzed from different perspectives, a process I look forward to as a future direction.

3.3.4. Discussion

In this paper I have introduced a new empirical dataset that may be used by the SNA community - the female fashion model dataset obtained by online crawling from Fashion Model Directory. I use the dataset to presented a state-of-the-art analysis of the the whole fashion model industry. Because physical features are relevant in the fashion context, I construct FMNet, which is a network of physical similarities in terms of hair color, eye color, height etc. Similar to studies on actor [98] and musician [258] networks, I can conclude that certain fashion models have higher centrality indices. Like other collaboration networks, FMNet has the same properties: it is scale-free (meaning that fashion models' connectivity distributions are in a power-law form), and has a high degree of centrality [276].

In light of the study which finds Kevin Bacon as the most influential node in the Hollywood actor network, and Dave Grohl in the music industry, I find Kelsey Gerry as the "Kevin Bacon of the fashion model world". Her high betweenness centrality represents the reference in terms of physical trait patterns for other fashion models. A second important empirical observation is the existence of a single dominant community of nodes with very high eigenvector centrality. This is the community formed by another reference in the fashion world: models with blonde hair and blue eyes, a trademark for many modeling agencies and magazines. The presented work is aimed at improving the understanding of how fashion models collaborate and, the the best of my knowledge, create a first recommender system for fashion agencies.

3.4. Graph metric analysis in collaboration networks

The comparison is done using the topological metrics which are specific for complex networks [246, 10, 276, 197]. These values are represented in Table 3.8: average degree (AD), average path length (L), average clustering coefficient (C), modularity (Mod), graph edge density (Dns) and graph diameter (Dmt). These metrics have the power of characterizing a complex network and offer valuable insight [276].

Table 3.8.: Basic graph metrics for FMNet, MuSeNet, and three online social networks: Facebook, Twitter and Google Plus. The measured metrics are: average degree (AD), average path length (L), average clustering coefficient (C), modularity (Mod), density (Dns), and diameter (Dmt).

	AD	L	C	Mod	Dns	Dmt
FMNet	58.06	3.14	0.497	0.579	0.017	9
MuSeNet	13.18	7.64	0.884	0.844	0.002	23
IMDB	113.5	1.55	0.996	0.476	0.062	4
Facebook	22.23	2.34	0.256	0.577	0.005	7
Twitter	12.39	2.68	0.239	0.28	0.054	7
Google Plus	12.15	3.9	0.404	0.44	0.035	12

Table 3.9.: Fidelity measured against the three online social networks: Facebook (φ_{FB}), Twitter (φ_{TW}), and Google Plus (φ_{GP}). A higher φ value means a higher similarity between the collaboration network and the online network.

	φ_{FB}	φ_{TW}	φ_{GP}
FMNet	0.619	0.567	0.68
MuSeNet	0.486	0.451	0.574
IMDB	0.472	0.535	0.537

The numerical results from Table 3.8 show that FMNet has a good resemblance towards online collaboration networks. This is supported by the scale-free property which results in a small average path length L. The clustering is higher than for internet users but lower than that of the IMDB and MuSeNet networks. This can be explained through the fact that musicians and actors work in much tighter collaboration than average internet users who interact with many other diverse users. The physical similarities of FMNet however, tend to be in the middle, showing that the average similarity between fashion models is somewhat higher than expected in a normal population. In other words, there are few physical feature combinations present in the fashion world, and this high redundancy context yields a high clustering coefficient. The community structure is thus also relatively high, as each physical pattern is clearly delimited from the other.

Finally, I compare the three collaboration networks with the three online social networks, as presented in Table 3.9. As FMNet has the highest resemblance to any of the three online networks.

3.5. Motif distribution analysis in collaboration networks

I propose a two step approach for motif-based comparison on MuSeNet and FMNet. First, I measure the distributions of motifs of size 3 (subgraphs with 3 nodes) on each empirical network using FANMOD [284]. FANMOD is a light-weight tool for fast motif detection designed using one of the fastest detection algorithms available, RAND-ESU [283]. To that end, I obtain the distribution depicted in Figure 3.7.

The results of the motif analysis offer a different perspective over the already reached conclusions.

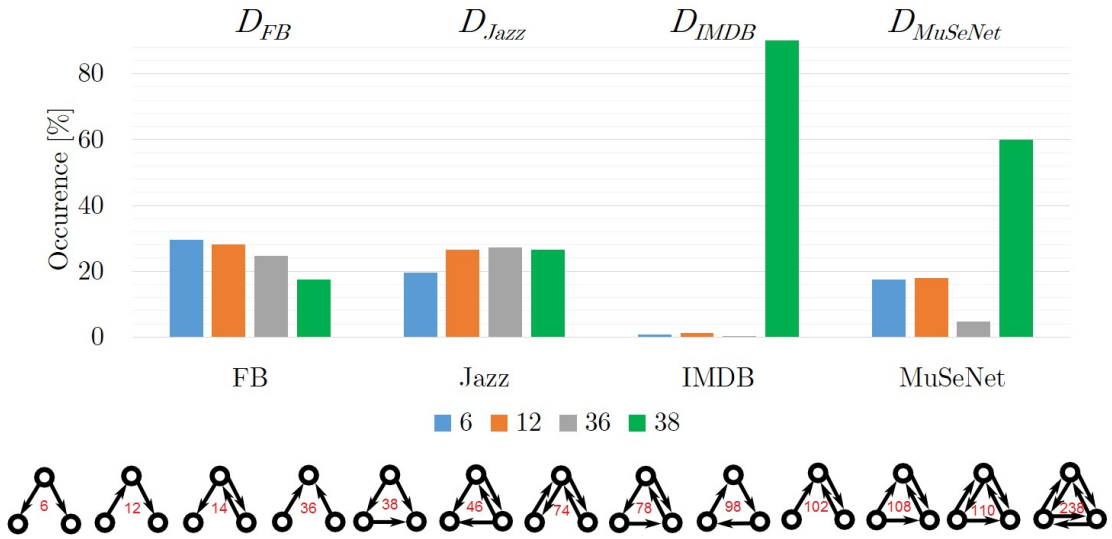


Figure 3.7.: The resulting motif distributions for the chosen empirical network topologies. The occurrence of each motif is expressed in percentage in the central histogram. As can be seen, only distinct motifs (not all) characterize each network. All 13 motifs of size 3 are depicted at the bottom of the figure.

The Jazz network behaves more like an online social network - with a uniform distribution of motifs - while the IMDB and MuSeNet networks have a predominant motif characterizing them.

The motif size used in this study is fixed to 3, that is subgraphs with 3 nodes are quantified, not larger ones. While there are approaches in literature studying network functionality using motifs of sizes 4-6, I rely only on the size 3 motifs since there are few such distinct patterns, they are much more numerous to be found in graphs, and thus substantially more relevant [11]. There are a total of 13 combinations of motifs with three nodes using directed edges. The motifs can be seen in the lower part of Fig. 3.7.

Finally, I also apply the fidelity metric to compare the motif distribution vectors with one another. The obtained values are given in Table 3.10. A value of 1 means complete similarity, while a value of 0 means complete dissimilarity. The data is interpreted as, for example: the Jazz network has a similarity of 81.8% towards the Facebook model etc.

The results presented here show a different perspective: both MuSeNet and IMDB are relatively distant to a normal online social networks (FB), while the Jazz musicians network is more similar. This supports the idea o a high sociability for the Jazz network. Also, the results support the fact that FMNet is the closest resemblance to a real-world online social network (i.e. 92% similarity to FB). All the obtained results support the fact that FMNet, even though a physical similarity based network, has all the properties of a social collaboration network.

Table 3.10.: Network fidelities φ of the three collaboration networks (rows) towards the four used references (columns) in terms of motif distributions.

φ	Reference models				
	FB	FMNet	Jazz	MuSeNet	IMDB
FMNet	**.92**	-	N/A	.525	.163
Jazz	**.818**	N/A	-	.662	.36
MuSeNet	.572	.595	.595	-	.231
IMDB	.341	.346	.171	.433	-

3.6. A perspective from non-social complex networks

In order to better understand the role that each graph metric, and centrality have in the context of social networks, I have done research also in the adjacent fields of technological and biological networks.

To that end, Appendix C details the results obtained in modeling patients datasets as graphs using the network medicine approach. This means that patients form a graph (i.e. nodes) and are connected based on their anthropometric compatibility, as specific for a certain medical scenario. Specifically, I present two sets of studies in which I have scientific results, namely improving the diagnosis accuracy for sleep apnea patients through phenotyping [186, 265], and improving patient treatment schemes taking into the consideration different types of medication administered to patients with cardiovascular disease [247].

This set of studies showcases the importance of community detection and the algorithms available to detect communities of nodes (i.e. patients, in this case). Moreover, in the process of building complex networks from ground-up it is important to determine a correct method for adding edges, and to limit the number of edges to such a value that community sizes and number are in balance. Having a *too weak* condition for adding edges between two nodes leads to a very connected graph in which communities are overlapped and hard to distinguish from a functional point of view. Having a *too string* condition for adding edges between two nodes leads to a disconnected graph in which there are too many small communities to be able to distinguish patterns, similarity, compatibility of any nature. In both these directions I have devised step-wise refined methods for adding a correct number of edges when creating graphs, in such a way that the resulting number of communities is relevant for analysis.

Finally, Appendix D details the results obtained in modeling technological communication infrastructures using technological networks as a theoretical support..Specifically, I present two sets of studies in which I have scientific results, namely understanding the formation of traffic hot-spots (e.g. traffic jams) in urban road networks [259], and creating an algorithm for improving the effectiveness of communication in a wireless sensor network, by taking into consideration the trade-off between cost and latency [129].

This set of studies showcases the importance of the betweenness centrality, eigenvector centrality, and the power of motifs to compare topologies. The detected hot-spots in urban networks are fully highlighted by betweenness. Analyzing their distribution, I noticed a power-law distribution, which is specific for social networks. On the other hand, degree is not distributed in a power-law manner, but rather normal or uniform. Thus, urban road networks tend to have a social component,

which is intuitive, since they serve our daily social needs. The second study presented in Appendix D [129] also shows the importance of community formation, this time in the geographical context of wireless sensors. Furthermore, eigenvector centrality is shown to capture the most important node in each community, being correlated with position as well. As opposed to a social network, where links are non-physical, the aspect of aligning centrality with the 2-dimensional center of weight of a community is essential in optimizing communication.

3.7. Discussion

In light of the results presented in this chapter, I have assembled a wide and valuable perspective over social networks analysis and modeling. The two studies regarding the collaboration of musicians (MuSeNet) [258] and fashion models (FMNet) [254] represent very innovative applicative approaches which bring novelty to literature. In both studies I have used graph metrics and centralities analysis to showcase the importance of the emergent communities which develop and explain real-world particularities of the two artistic fields. Together with a theoretical study of the impact of the underlying topology [261], all these studies have helped me understand the importance of graph metrics like average degree, path length, clustering coefficient, diameter, graph density and modularity, as well as the role of centralities like degree, eigenvector and betweenness.

Additionally, I have presented two studies of complex networks applied in medical science, following the so-called path of network medicine. The results obtained in predicting central sleep apnea [265, 186] bring landmark novelty and improvement in the field of sleep medicine, The study undertaken of assessing the treatment response of patient with hypertension [247] also showcases a new, useful, perspective for medical doctors. These experiences outside the field of social networks analysis have helped me greatly to understand the role of different metrics to take into consideration when modeling empirical data using graph analysis.

The observations obtained in all these studies have helped me pave the way for the next three chapters which deal with the essence of my thesis namely understanding social structures and creating mathematical models which can reproduce the topology, dynamicity and interactivity within social networks.

4. Generating realistic social network topologies

Social network analysis is receiving an increased interest from multiple fields of science since more and more natural and synthetic networks are found to share similar features which help us understand their underlying topological properties. One desire is to create a model of the human society, however, the complexity of such a model is increased by the nature of human interaction, and present studies fail to create a fully realistic model of the societies we live in. My approach is inspired from studies of online social networking and the ability of genetic algorithms (GA) to optimize topological data in a natural manner. I combine the properties of the small-world and scale-free models to create a community-based social network, which is then rearranged using empirically obtained data from Facebook friendship networks, and optimized using GAs. As a result, my synthetically generated social network topologies are more realistic, with a proposed realism fidelity metric that is with 63% closer to the observed real-world parameters than the best existing model

"If I have seen further it is by standing on the shoulders of Giants."

 Bernard of Chartres

4.1. Motivation

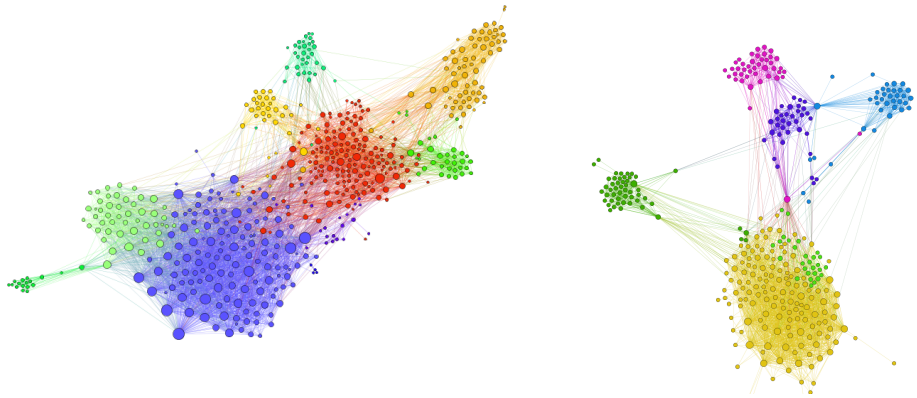
The effort to mathematically model an accurate and realistic society has been triggered by the observation of the three fundamental properties of social networks: average path length, clustering coefficient and degree distribution [246]. The well-known models of small-world [281] and scale-free [25] networks both present these network properties but they fail in creating fully realistic models of the societies we live in. Over the years, many attempts have been made to merge as many empirically-observed properties as possible into a single social model. There are topological models which describe geographical proximity, friendship distribution, neural networks in the brain, protein interaction mechanisms, natural food chains, the distribution of means of transportation, citation networks, sexual interaction patterns, the world wide web, power distribution networks, relationship of words in a language, interaction between ingredients in a recipe, the world markets [276, 121, 82] etc. However, an abstract and generic, yet flexible and realistic model that describes how people interconnect in society has not yet been described. The benefit of having such a model is the capability of simulating custom social scenarios of interest on very large virtual data sets. It can help medical science predict the spreading of diseases [82]; sociology and politics to understand the flow of information, opinion etc.; and combined with diffusion models, help predict the outcome of elections, polls, surveys [82]. As there is not enough real world data to research on, this can only be simulated if we have dependable, realistic social models [121]. If we only rely on mining after existing real data found in online databases it becomes hard to find topologies with a certain distribution of properties. For example, if one needs to simulate the behavior of a realistic social network with a certain centrality distribution of nodes, my proposed algorithm can create such a desired network on demand, at the same time assuring that it is as realistic as possible.

This chapter tackles the problem of synthetically generating realistic social network topologies. Unlike existing social models like the small-world [281], scale-free model [10], cellular model [263], static geographic model, Watts-Strogatz model with degree distribution [57], introduction model, random encounter model etc., I propose a methodology based on creating realistic models, inspired from accurate empirical data, which are then optimized using genetic algorithms. Previous methods fail in creating models of the human society; therefore I have used empirically obtained data from Facebook friendship networks and have concluded that, although diverse in shape and size, all share common metrics. Furthermore, I optimize the metrics on my propsoed model until it reaches a desired state of realistic accuracy.

The original model - entitled Genosian - proves to be more efficient by replicating all network metrics measured on the empirical data set. I quantify this efficiency through the proposed *fidelity* metric φ (see Appendix A) which can measure the realism of any network model. The algorithm is highly parametrized, flexible for multi-purpose social scenarios, and is also being integrated into Gephi [30], the leading tool in visualization and analysis of large networks.

4.2. The real-world reference data

My research is based on the empirical study of Facebook friendship graphs which have been extracted from Facebook using an application named *netvizz* [228]. The data set consists of 93 different friendship graphs of subjects aged 16 to 35, with sizes ranging from 177 to 1030 nodes. Even though different at first glance, under a closer numerical analysis all measured metrics vary only slightly between



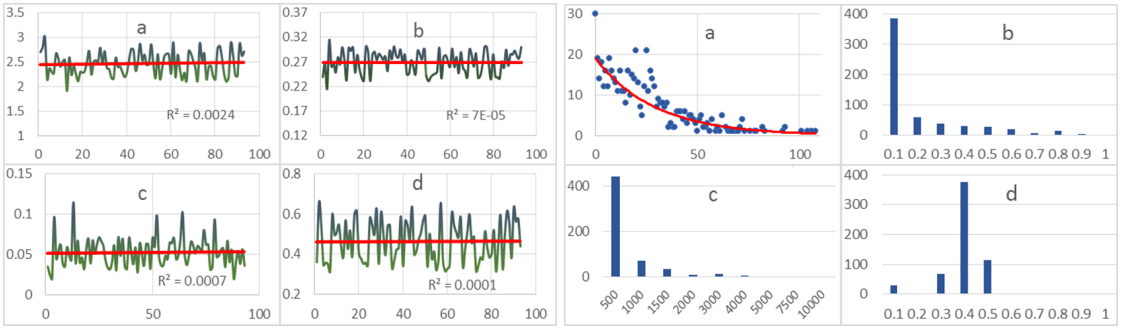
(a) (b)

Figure 4.1.: Two online friendship networks: a Facebook network **(a)** of 590 nodes and a GooglePlus **(b)** network of 344 nodes. The size of each node is proportional to its degree and the coloring is done according to the community it belongs to (*i.e.* done by running a community finding algorithm first). This Facebook network is chosen as an example because it lies nearest to the overall average metric distribution from the data set.

the multitude of networks. Figure 4.1a shows such a friendship network extracted from Facebook. I start by measuring the basic network metrics: network size (nodes and edges), *average path length* (L), *clustering coefficient* (C) and *average degree* ($<k>$), and also network *diameter*, *density* and *modularity* [276]. Additionally, I analyze the distributions of the *degrees* ($P(k)$), *betweenness*, *closeness* and (eigenvector) *centrality* [276, 150] .

Figure 4.2a highlights the narrowness of the convergence intervals for average path length, clustering coefficient, density and modularity as measured for all networks extracted with *netvizz*. This strengthens the argument that all realistic topologies have metrics which fall inside these thresholds. Figure 4.2b shows the degree and centrality distributions for a representative friendship network. The chosen network (Figure 4.1a) lies nearest to all average values for each metric. Despite intuition and the inherent diversity of humans, it is clear that the measured values pertain to a social pattern which seems to be found in the underlying nature of the human interaction model. I have evaluated these parameters because they characterize a realistic social topology. Following similar reasoning, a study demonstrates that even a completely synthetic network – the Marvel characters universe – has evolved into a real-like social network [8]. Therefore the proposed synthetic topology generation process has to meet these demands.

The study presented in this chapter is explicitly tailored for recreating Facebook (online) friendship networks, as these networks best capture the aspects of social interconnectivity. As for a dataset repository, I rely on the Stanford Large Network Dataset Collection (SNAP) [157], which contains many medium and large sized networks for this study. Additionally, I rely on a privately collected dataset of ~100 egonetworks of smaller sizes. As a future step, the study may be extended to extract the characteristics of Twitter, Google Plus, peer-to-peer networks (technological) and redefine the algorithm to suit additional specific cases.



(a) (b)

Figure 4.2.: **(a)** The distribution of measurements over the data set: **a.** Average path length (L) with an aver-age value of 2.48, a minimum of 1.92 and a maximum of 3.0; **b.** The clustering coefficient (C) with an average value of 0.26, a minimum of 0.21 and a maximum of 0.31; **c.** Network density with an average value of 0.052, a minimum of 0.02 and a maximum of 0.11; **d.** Network modularity with an average of 0.462, a minimum of 0.31 and a maximum of 0.65. Degree and centrality distributions for one representative network (represented in Figure 4.1a). **(b)** The distributions for a representative network: **a.** Power law degree distribution; **b.** Power law eigenvector centrality distribution; **c.** Power law betweenness centrality distribution; **d.** Closeness centrality distribution. It presents a particular Gaussian distribution with a cutoff value (0.5). This is a specific feature for friendship networks.

4.3. Evaluating the related work

Basic network analysis is done by measuring fundamental graph metrics, and comparison of two or more networks, by doing an individual comparison of each metric independently. While such an approach is useful in trying to capture one specific feature of the network, it fails to create a general overview of the similarity between the analyzed networks [66]. Similar work aimed at comparing the importance of graph metrics concludes that each metric captures specific attributes of the network [35]. It states that further study on the effect of each metric is needed in order to be able to choose a fitting topology according to the requirements of an empirical model.

Comparing real systems is aimed at a deeper understanding of the interaction patterns between these systems [281, 25, 246], and extracting their common properties helps improve the models even further [281, 13, 141]. However, the predominant method of graph metric comparison suffers from limited information [165]. Some notable means of comparison are the distance ratio measure [49], used to compare individual mental models, a comparison from the data analysis perspective [165] and the study of the self-similarity of complex networks (Song, 2005). From a topological perspective there are studies done both in the direction of classifying social network models [141] and of structural pattern detection [215]. These methods however serve a higher level of meta-analysis rather than as measures of similarity.

The statistical methods with which network similarity can be measured are the cosine similarity [249], variance, covariance, Pearson correlation coefficient (PCC) [245], the Mahalanobis distance [174]. Other methods used in network analysis which are adopted from statistics include the T-test and the ANOVA test (analysis of variance) [70]. A recent study improves upon the T-test methodology by proposing an alternative geometrical approach called the Characteristic Direction, in order to identify differentially expressed genes [61]. There is no single statistical approach used in current research because there is no unified metric that provides normalized values which are tailored specifically to network comparison. Yet, the most intuitive and thus used metrics are the Euclidean distance, Pearson correlation [244] and cosine similarity [249].

Analyzing the results from Table 4.1 and Figure 4.3 one can highlight the limitations of the current related work. No single state of the art model manages to replicate more than two or three empirical measurements. As there is no suitable metric to compare and quantify the realism of social networks, this thesis makes use of the network fidelity metric φ (phi) which was previously introduced in literature by the author [257]. For full motivation and mathemtical validation of this metric, see Appendix A. By measuring the φ of any two networks represented with the same metrics, it can be concluded which model offers the greatest realism compared to my empirical data set. Also, the algorithms for generating all the analyzed networks, as originally described by their respective authors, are implemented as Gephi plug-ins by the author.

Because none of the presented work manages to fully model a realistic friendship network, as modeled on Facebook, I further propose my own topological model which encompasses all the metrics and distributions.

4.4. The genetic-optimized social network (GenOSiaN)

The goal of my proposed social network model (called *Genosian*) is to create an accurate replica of the friendship models gathered from Facebook. Figure 4.4 presents the overview of my proposed

Table 4.1.: Measurements for average degree (*AvgD*), average path length (*L*), average clustering coefficient (*C*), modularity (*Mod*), diameter (*Dmt*) and density (*Dns*) are done on synthetically generated networks of 1000 nodes. The presented models are: Facebook, small-world (SW), scale-free (SF), cellular, static-geographic, WSDD and the proposed model (Genosian). A lower φ-value shows the realism of the models, with the geographic model being the most accurate state-of-the-art model ($\varphi = 0.27$), but with the Genosian (proposed) model being 122% more realistic ($\varphi = 0.125$).

	AvgD	*L*	*C*	*Mod*	*Dmt*	*Dns*	φ
Facebook	19.822	2.4815	0.2659	0.4677	8.5	0.0496	0
SW	3.994	5.614	0.321	0.726	11	0.005	0.371
SF	3.12	4.598	0.015	0.622	10	0.003	0.38
Cellular	11.388	3.786	0.599	0.908	7	0.02	0.367
Geographic	6.628	3.34	0.065	0.52	8	0.013	**0.277**
WSDD	21.583	4.589	0.738	0.897	9	0.041	**0.31**
Genosian	20.02	2.404	0.308	0.65	5	0.05	**0.125**

methodology. As there is no direct algorithm of ensuring the creation of a graph with the desired predefined metrics, I adopt a step wise refinement approach. By combining the two fundamental models, small-world and scale-free, a community based network is created. This approach is also validated by similar research, namely the cellular network model [263], the Watts-Strogatz model with degree distribution [57], and scalable virtual communities [226]. The resulting communities are connected with initially random friendship links which are then re-wired (optimized) using a genetic algorithm (GA) approach. Related work shows that reciprocity and sibling bias are shown to have a considerable effect over the creation of friendship ties [194], thus our friendships tend to form around the dominant nodes in each community. Using a custom GA the centrality distributions are repeatedly measured and optimized until they correspond to the empirical distributions.

To overcome the problem of searching in an infinite solution space I propose the usage of heuristic methods, namely genetic algorithms (GAs), to solve this computational limitation. GAs are used to generate a representative mixture of solutions to optimization and search problems [191].

The initial set of random solutions which build up the population is, in this particular case, composed out of edges (candidate solutions) which need repeated rewiring, as in a genetic manner, to produce better and better solutions. Finally, after a predefined number of repetitions (generations) the GA stops and the best solution is chosen from the population [191]. Solutions are ordered using a fitness function, namely the betweenness and/or the eigenvector centrality of each edge target. Figure 4.5 shows the chromosome representation used by the algorithm. Each solution (candidate) consists of en edge: a pair node source - node target on which the genetic operators are applied.

A. Initialize Parameters

The algorithm takes as input the following: the number of communities, the average community size, the two rewiring probabilities p_1 and p_2, and the genetic algorithm parameters. Each community is individually built using the same principle, so the number of communities is used to model diverse real situations, like *e.g.* number of college groups. By default, the average community size determines a power law distributed size around the given value; alternatively, each communities' size can be manually set.

B. Create communities

The creation of each community is an independent task and is inspired by the Watts-Strogatz algorithm. The proposed difference consists in how the value *k* is individually chosen for each node, as inspired from the WSDD approach. It is aimed at creating a small world network with a power law degree distribution. After the regular network ring is created, local edges are rewired to long range edges, within the community, using probability p_1. At this step, I obtain a given number of communities, each with realistic *L*, *C*, density and degree distribution.

```
Create community:
create size nodes with id = community index | global node index
for each node n_i:
  connect n_i to k neighbors on left and right sides (2*k edges)
for each edge e_i, with probability p_1:
  choose a new random/preferential edge target from the community
```

C. Connect communities

The last initialization step consists of connecting the obtained communities. Using probability p_2 each node is connected with another random or preferentially chosen node from a different community. The preferential selection is done by choosing higher degree nodes in favor of lower degree ones. This step stabilizes the graph density, diameter and modularity, but the centralities remain normally distributed. After this step, the list of added inter-community edges is kept for the next iterations of the algorithm.

```
Connect communities:
for each community c_a:
  for each node n_i from c_a, with probability p_2:
    choose another random community c_b
    choose a random/preferential node n_j from c_b
    create edge e_k between (n_i, n_j) and save it to list E
```

D. Measure fitness of edges

The list of newly created edges is sorted in descending order of the betweenness and/or Eigenvector centrality of the target of each edge. For this, I run the corresponding centrality measurement algorithms and then order the edges. The idea is to rewire the edges by keeping the source node, but selecting a better target node. Better targets represent nodes with higher centrality, as my empirical

observations suggest. The GA can be repeated for either a given number of steps (iterations) or until the measured centralities resemble the empirical ones. Experiments show that running a higher number of iterations (>5) makes the network organize itself in a perfect manner, which actually decreases the realistic accuracy. Consequently, I suggest using the algorithm with a fixed number of steps (2-5).

```
Measure fitness:
for each edge e_i in E:
  fitness f_i ← centrality(target n_i of e_i)
sort E in descending order of f_i
```

E. Rewire edges between communities

Considering I sort the population in *descending* order of the fitness after every iteration, I evolve the solutions from one generation to the next using three methods:

1. *Best solutions*: the first percentage $pBest$ of the current generation is copied to the next generation. That is, the edges with the most central target nodes (best fitness) are kept in the graph (there are $sBest = pBest \times N$ individuals chosen).

2. *Crossover*: a second percentage $pCross$ of the next generation is composed out of edges from the current population on which a custom crossover is applied (there are $sCross = pCross \times N$ individuals chosen for crossover). It is applied on the edge target using a second random target node chosen from the same community. The local IDs from the chromosome of the original target and the second random target are combined through binary concatenation using the first c bits from one node's ID and the remaining bits from the other node's ID. The crossover threshold *c* is a random number: $0 \leq c \leq n$, where *n* is the number of bits used to represent the IDs. The resulting index is guaranteed to belong to a specific node in the community, which is then set as the new target for the particular edge on which the crossover is applied.

3. *Mutation*: a third percentage $pMutation$ is composed out of the remaining edges from the current population on which genetic mutation is applied (there are $sMutation = pMutation \times N = N - sBest - sCross$ individuals chosen fro mutation). It is applied by changing the edge target node with another random or preferential node from the same community as the target, as given by the fitness function.

The three percentages add up to 100%. Finding the best values for them is an experimental study and may differ from one need to another. Fig. 6 explains how the algorithm is applied step by step.

```
Apply genetic operators (one step):
E' ← empty
while i++ < sBest:
  E'_i ← E_i
while i++ < sBest + sCross:
  n_j ← random node from community of target of (E_j)
  new_idx ← crossover (getNodeId(n_i), getNodeId(n_j))
```

```
  n_new ← getNode (new_idx) from same community
  target (E_i) ← n_new, add to E_i
while i++ < size(E):
  n_r ← target of random/preferential edge from E
  target (E_i) ← n_r, add to E_i
E ← E'
```

Once the algorithm stops, it produces a graph of size $N \simeq$ *number of communities* $\times$ *average community size*, which significantly resembles the presented Facebook friendship networks. The basic metrics (average degree, *L*, *C*, diameter, density, and modularity) are realistically obtained through steps A-C of the algorithm, while steps D-E ensure the centralities are distributed as needed. Finally, in Figure 4.6 I exemplify the algorithm in a step-by-step manner and explain how the rewiring is done.

4.5. Results and discussion

Conceptually, the Genosian algorithm is designed to recreate realistic online social networks topologies, but it differs from the state of the art models. I consider the algorithm to be a multi-objective optimization method, through the fact that it targets a specific set of graph metrics which have to be *optimized*. As a consequence, I use the developed fidelity metric [256], which is designed for a suitable comparison manner for multi-variable entities, to compare to the chosen state of the art models and empirical data presented in Table 4.1.

In this section I present and compare the similarities between the Genosian network, the real Facebook data set, and the best two models presented in the related work chapter, both visually (Figure 4.7) and numerically (Table 4.2). Comparing the results in Table 4.2 with the ones in Table 4.1, it is clear that the Genosian networks manage to accurately replicate the original Facebook network. No other model reproduces more than 3 basic metrics ($0.27 < \delta < 0.38$), while the proposed model reproduces 5 out of 6 ($0.125 < \delta < 0.2$). Thus my model is, on average, 63% more accurate than the best previous model (Geographic); the best network is however 2.21 times more accurate than the Geographic model, and 2.47 times more accurate than the WSDD model. It is worth mentioning that my synthetic networks do not create the random leaf nodes which increase the diameter, as they are statistically insignificant, thus the real diameter is lower. The six networks highlight the impact of the wiring probabilities. A low p_2 creates a very modular community structure which is not desired. Increasing p_2 decreases the clustering and the modularity, and increasing p_1 increases the density. Figure 4.7, along with Figure 4.1, highlight the results in a visual manner. One can observe the similarity between the Facebook friendship networks (Figure 4.1a, Figure 4.7a) and the proposed Genosian network (Figure 4.7b). Nevertheless, the numerical comparison is the one that makes us conclude how realistic a model is, and both Table 4.1 and Table 4.2 reinforce the fact that my proposal produces the best reproduction of real friendship network topologies. In Figure 4.8 I also show the distributions of centralities for the Genosian model. This further stresses the superior realism of my proposed model because, in comparison with the other models, it represents a better match for the actual Facebook distributions from Figure 4.2. Throughout this chapter I rely on Facebook friendship networks as a basis for comparison and validation of realism. This is argumented by the popularity of Facebook, which offers large, diverse and real-like networks, as explained in Section 2.

However, other platforms, like GooglePlus (Figure 4.1b), Twitter, Wikipedia *etc.*, offer empirical data for validation. In principle, there is no problem for Genosian in replicating topologies that resemble GooglePlus or Twitter, provided that similar studies, as those pertaining for Facebook (as provided in Section 2), are performed.

Table 4.2.: The basic metrics for a representative Facebook friendship network and for six Genosian networks of sizes 500-1000 nodes. The two columns on the right represent the wiring probabilities p_1 and p_2 (see steps A, B, C of the algorithm) used to create each of the distinct synthetic networks. The values used for the genetic percentages are: *pBest*=50%, *pCross*=30%, *pMutation*=20%.

	AvgD	*L*	*C*	*Mod*	*Dmt*	*Dns*	p_1	p_2	φ
FBook	19.82	2.481	0.266	0.47	8.5	0.05	-	-	0
G1	22.26	2.514	0.343	0.62	4	0.032	0.1	0.1	0.2
G2	22.17	2.426	0.212	0.5	4	0.029	0.1	0.2	0.166
G3	21.18	2.185	0.142	0.36	3	0.046	0.1	0.3	1.89
G4	21.45	2.416	0.324	0.65	4	0.044	0.2	0.1	0.167
G5	21.88	2.353	0.182	0.49	4	0.033	0.2	0.2	0.172
G6	20.02	2.404	0.308	0.65	5	0.05	0.3	0.1	**0.125**

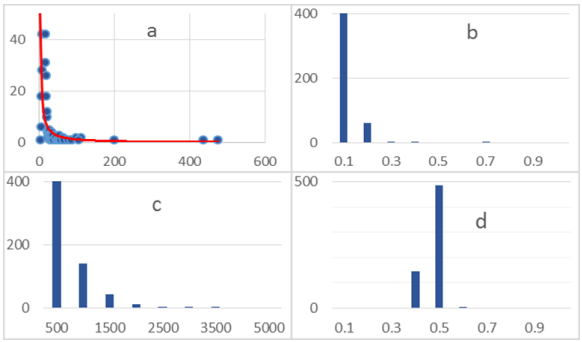


Figure 4.8.: The centrality distributions for a representative Genosian network. See Figure 4.2b for comparison. **a.** Power law degree distribution **b.** Power law Eigenvector centrality distribution, exactly as in Facebook networks, and unlike many other social network models. **c.** Power law betweenness distribution **d.** Closeness distribution with the same particular Gaussian distribution as in real Facebook topologies.

Modeling the societies we live in is one of the goals of social network analysis. This endeavor stretches out in multiple directions: defining a mathematical model of the topology, modeling real-time growth, adding an opinion diffusion model etc. Many natural and synthetic networks have

already been analyzed and their underlying models understood, documented and reproduced. However, the quest remains open to propose an accurate model of the society. While it encompasses the most fundamental properties – small-world and scale-free – it also brings a lot of complexity due to the nature of human interaction. My proposal – the Genosian network – is an innovative solution combining a realistic empirical data set with social network analysis and genetic algorithm optimization.

This first contribution to my thesis begins by showing that Facebook friendship networks are accurate at reproducing the real friendship networks between humans. The data set of over 100 such networks, with sizes ranging from a few hundreds to thousands of nodes, is then analyzed and it is concluded that, although very diverse in shape and size, all these networks share very strict metrics. Further, I present the work related to newer proposals in this direction. It is shown that none of these manage to replicate the properties of the empirical friendship networks. Thus, I propose the Genosian network model and explain how it is algorithmically generated. I finalize by offering both a visual and a numerical comparison and discussion between the proposed, empirical and other related models.

In conclusion, I believe that my work has achieved its goal and manages to replicate realistic societies very accurately. The achieved accuracy shows a 63% improvement with respect to the best previous model (static-geographic model), in terms of the realism fidelity metric φ. The inspiration from Facebook is nothing but natural, as more and more virtual data is modeled by human interaction. My future work is aimed at extending the real social network topology study for Google Plus, Twitter, *etc.* in order to refine the Genosian algorithm accordingly. I also plan on using this model as a basis for social network growth.

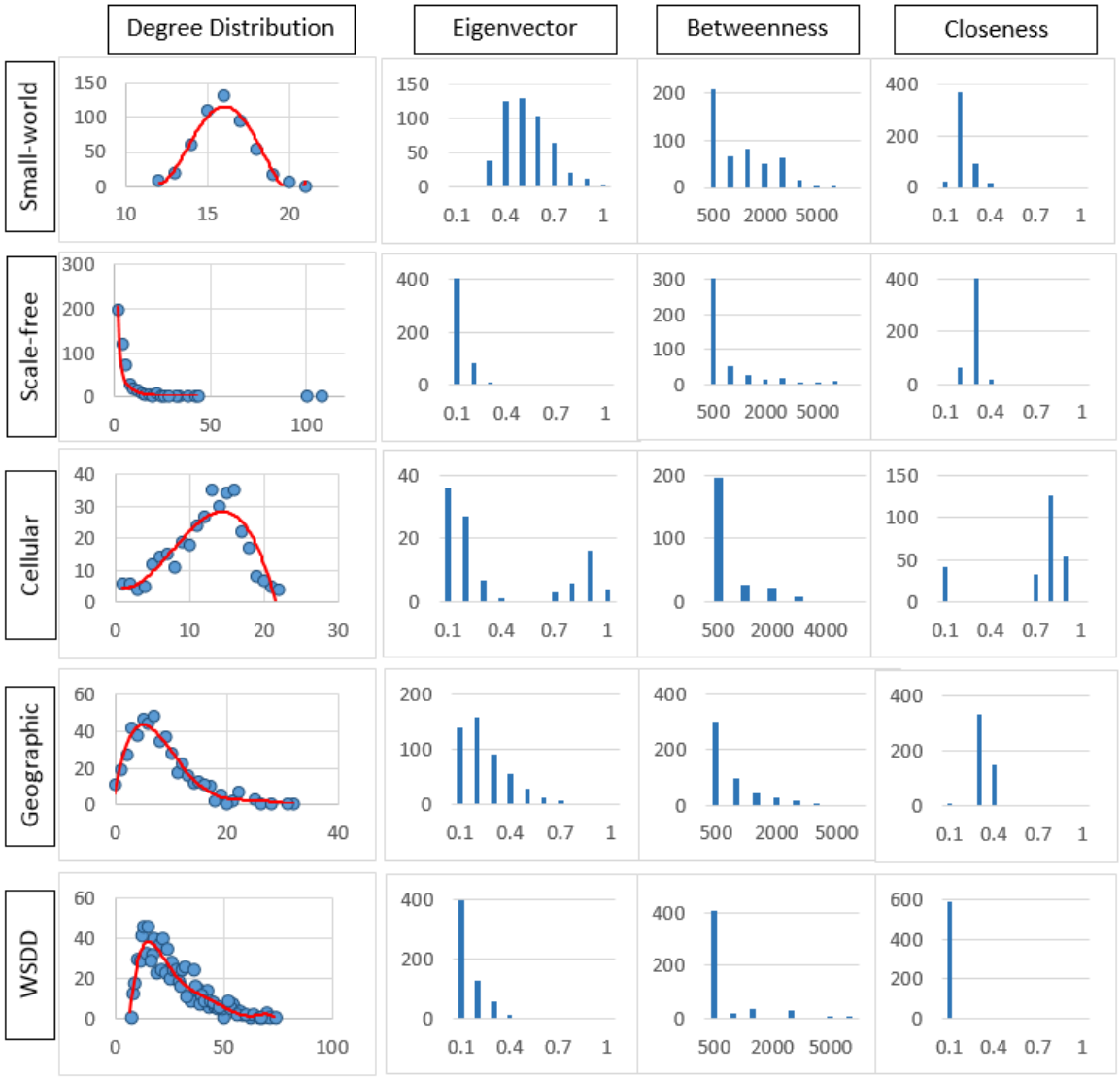


Figure 4.3.: The degree and centrality distributions over a selection of five relevant social network models (the same ones as described in Table 1).

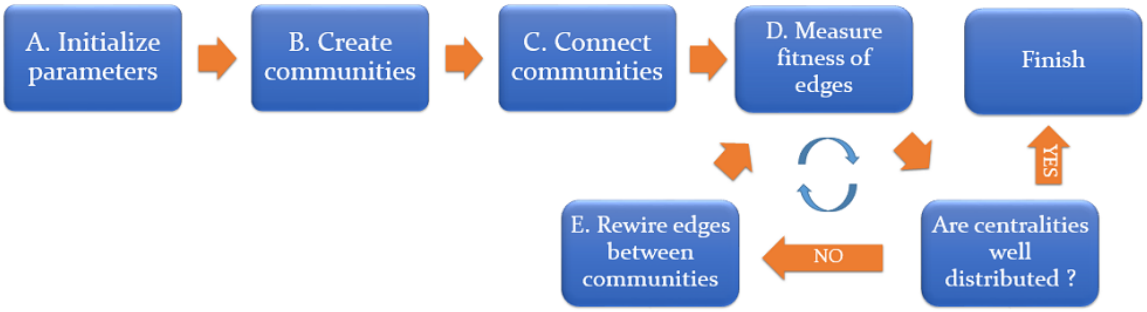


Figure 4.4.: Flowchart describing the steps of the Genosian algorithm.

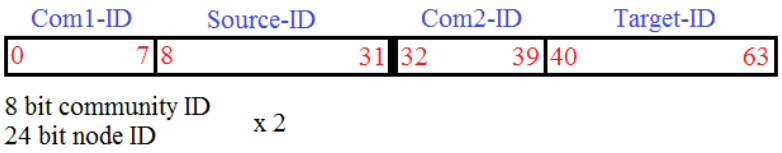


Figure 4.5.: The genetic chromosome representation. Each solution is composed out of two 32bit-represented IDs. The source node is represented by concatenating the community ID of the node (8bit) with the actual node ID (24bit). The same rule applies for the edge target.

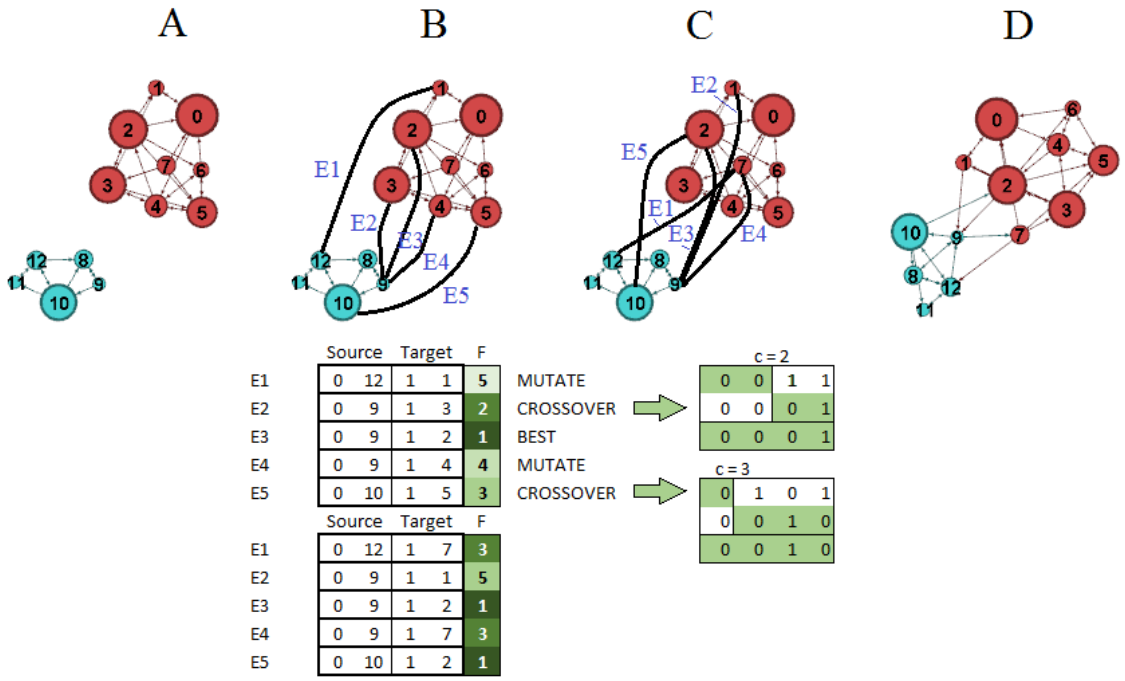


Figure 4.6.: Step by step explanation of the *Genosian* algorithm. The table shows the evolution of the chromosomes, their fitness ranking (green), and exemplifies the crossover. **A.** Communities C0 (cyan) and C1 (red) are created [steps A, B of the algorithm]. **B.** Five random edges are drawn be-tween the two communities: E1 to E5 [step C of the algorithm]. The fitness F of the edges is computed and the edges are ordered (1-5 in the green column). F is given by the centrality of each edge's target, *i.e.* which is proportional to the size of the nodes in the figure [step D of the algorithm]. **C.** Apply the genetic operators and rewire the five edges. Thus, in order of the fitness, the best solution is copied over (E3), crossover is applied on the next two solutions (E2, E5), and mutation on the last two solutions (E4, E1). Crossover on E2 is applied by combining the target "3" with a random target "1", with c=1, which results in the new target "1". Mutation on E1 is applied by choosing a new random target from the same community, namely node "7" [step E of the algorithm]. **D.** Considering the algorithm is finished after one step, the ForceAtlas2 layout algorithm is reapplied on the graph [132].

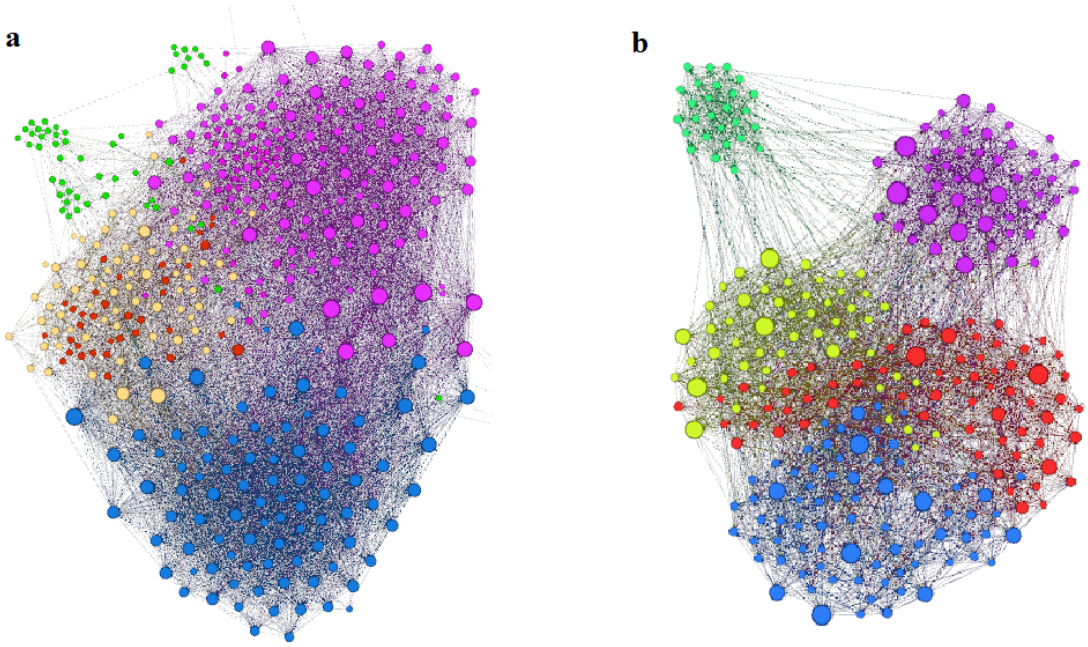


Figure 4.7.: A visual comparison between a Facebook friendship network **(a)** with 457 nodes, and a Genosian network **(b)** with 269 nodes. The synthetic network in the figure is the one corresponding to G2 in Table 4.2. The coloring of all nodes is done according to the community they belong to, and their size is proportional to their degree.

5. Betweenness as the driving force behind social networks emergence and evolution

Existing models of social network topologies are based on the principle of degree-driven preferential attachment. However, by using a precise fidelity metric, I argue that a paramount feature of realistic social network topologies is that both node and link betweenness need to be power-law distributed. As a second theoretical contribution, my analysis reveals that, in real-world social networks, the link weights are correlated with nodes betweenness. Consequently, I propose a new social network model and generation algorithm driven by the principle of node-betweenness preferential attachment. The experimental results show that my betweenness preferential attachment (BPA) algorithm is more accurate than the state-of-the-art models. Indeed, besides reproducing the power-law distribution of node and link betweennesses, my generation algorithm makes all other social network parameters and centralities fall into place naturally within the validated realistic thresholds. Finally, in this chapter I propose a new socio-psychological interpretation which transcends the mere topological view by offering a deeper understanding of how the social ties evolve and develop. Taken together, these three contributions represent a major step towards a deeper understanding of mechanisms behind social network emergence and evolution.

"Imagination is more important than knowledge."

 Albert Einstein

5.1. Motivation

Despite the widespread occurrence of the normal (Gaussian) distribution in nature, many social, biological, as well as technological networks can be described by a power-law (Zipf) distribution of nodes degree [276]. Therefore, the Barabasi-Albert (BA) model, based on the degree-driven preferential attachment (i.e. the "rich gets richer" principle), has been proposed to model such networks [25]. Still, it was argued that using the BA model to describe or analyze social networks can be cumbersome or even inappropriate [6]. This is because a node degree driven model may not be a good predictor of how people connect and how their social ties evolve [6, 47, 1]. The most compelling arguments are:

- People are physically and psychologically limited to a maximum number of real-world friendships, namely there is a maximum degree to be reached [79, 45]. However, in the BA model there is no upper limit for nodes degree.
- People have weighted relationships, that is, not all friends are equally important. Studies have actually shown that the average person knows roughly 350 persons, can actively befriend no more than 150 people (Dunbar's number) [79], but actually has only a few very strong ties [148]. Obviously, the distribution of weights in one's ego-network affects the evolution of social networks in a significant manner.

Starting from these overarching ideas, this contribution presents a systematic approach that can explain the formation of realistic social networks, underpinned by the pivotal role of node betweenness [264]. Indeed, while existing literature provides only some case studies on the importance of betweenness [160, 1], my main objective is to define an accurate social network topological model based on:

1. Introducing a new betweenness preferential attachment (BPA) algorithm that employs the betweenness-driven preferential attachment mechanism and node betweenness – link weight type of correlations.
2. Providing a socio-psychological interpretation of betweenness centrality's role; this offers a deeper understanding on social networks evolution and development.

As a result, it is found that:

- The synthetic topologies generated with BPA have a much higher fidelity towards reproducing the real-world social networks compared to previous models reported in the literature.
- Although the newly proposed algorithm uses only node betweenness for ties formation and links weight allocation, all other social network parameters and centralities fall into place naturally within the validated realistic thresholds [257].

5.2. Background

In my representation, a social network is a graph $G = (V, E)$ with nodes $v \in V$ (individuals, agents) and edges $e \in E$ (relationships, friendships), that can be directed and weighted to represent

relationships among various individuals, according to a real social structure. The role of such a graph is to help us get insight on how relationships evolve and how information is passed within the society, as determined by the social interactions among people [276, 82].

This section describes the state-of-the-art models for social networks representation, along with their defining topological features [10, 276]. Therefore, I rely on the most relevant network parameters such as average path length (L), clustering coefficient (CC), modularity (Mod), density (Dns), diameter (Dmt), as well as the most important centralities like degree, betweenness, closeness, and eigenvector centrality [276, 203, 205, 150].

Extensive empirical research has defined three fundamental features of real-world social networks, namely the small-world effect[246, 281], the scale free property [25, 276], and the emergence of community structures [202, 91]. Recent research aimed at improving the accuracy of social network topologies mainly consists of attempting to combine the properties from the two previously described fundamental models with empirical data gathered from various contexts. As such, there exist proposals which either add the small-world property to scale-free models [123, 96, 166] or approaches that add power-law distribution to small-worlds [135, 57, 274, 296].

In this study, I rely on the *Watts-Strogatz model with degree distribution* (WSDD) is designed by creating a small world topology (short average path length *L* and high clustering coefficient *CC*), and then modifying the degree distribution of nodes, from normal into a power law [57]. *Cellular networks* have been proposed as a response to the need for large-scale multi-agent simulations [263], and are based on the observation of covert networks like terrorist organizations. Cellular networks consist of an arbitrary number of normally-distributed sized cells, having a high clustering, in which a node is chosen as the cell leader. The cells are interconnected only through their leaders, and have a high tolerance to attack and infiltration. Further models exist that extend the conclusions of Milgram's experiment [187]. For instance, the *static-geographic model* [154] generates a network in which links are added between nodes by taking the actual spatial distance into consideration: the greater the distance, the lower the wiring probability.

5.3. Dataset analysis

In this section, I analyze diverse data sets available from online social networks, namely Facebook, Twitter and Google Plus friendship networks. Prior studies confirm that data mining from sources such as Facebook is reliable for realistic social network research [125, 91]. Another reference also indicates a strong correlation between the real world social capital and virtual friendships of people [268]. This conclusion is also supported by my own previous studies of Facebook datasets [257].

I argue that considering link weights is paramount for defining realistic social network models. On the other hand, because the availability of weighted social network datasets is still problematic, I present real-life social network parameters both in unweighted and weighted contexts. This way, I am still able to underline the characteristics which act as a "signature" in differentiating social networks from other types of complex networks.

5.3.1. Unweighted social network parameters

I use the Facebook data sets to study the distribution of other relevant centralities: eigenvector, pagerank, and betweenness (Figure 5.1). Node betweenness is therefore defined as:

$$b(v) = \sum \frac{\sigma_{ij}(v)}{\sigma_{ij}} \tag{5.1}$$

where $i \neq j \neq v$, σ_{ij} is the total number of shortest paths from node i to node j, and $\sigma_{ij}(v)$ is the number of those paths that pass through node v [1][94, 203] .

Figures 5.1a-c show specific power-law distributions for degree, eigenvector, and betweenness. On the other hand, the closeness centrality (Figure 5.1d) has a normal distribution with a clear cut-off on the right side.

In order to present an illustrative example of how betweenness is distributed in real-world networks, I present the Jazz musicians collaboration network, which consists of 198 nodes and 2742 edges [48]. This network (see Figure 5.1e) represents the professional collaboration of many Jazz musicians throughout their career, linking two musicians (i.e. nodes) if they have collaborated at least once in the past. Even though the analyzed networks are not weighted, I find that the power-law distribution of betweenness is a naturally occurring pattern in many real world networks which involve a social relationship.

5.3.2. Weighted social networks characteristics

When assuming that the social network links are weighted in accordance to the strength of the social ties among different people, a very important aspect is the distribution of these weights. Most current research on social topics like opinion dynamics [291, 4], influence mining [128], political preferences [214, 107] is based on unweighted data; however, I have gathered a collection of diverse weighted social networks in order to study the distribution of weights and their correlation to the degree and betweenness distributions.

For instance, I consider the Les Miserables co-appearance network [147] with 77 nodes (nodes are characters; edge weights are number of co-appearances of two characters in the same scene), an online social network [211] with 1899 nodes (nodes are users, while edge weights represent number of online interactions), and a large Twitter network of 37,366 nodes [290] (nodes are users, edge weights are number of retweets between users).

Edge and node related betweenness distributions

Using the graph visualization tool Gephi [30], I show that the empirical social networks are characterized by a power-law distribution of their weights (see Figure 5.2a-c). Furthermore, as also revealed by the analysis of unweighted social networks, the betweenness centrality seems to share a similar type of power-law variation, but with a clear cut-off value. This aspect is shown in Figure 5.2d-f. According to the organization of nodes, the edge betweenness follows a similar power-law distribution, as depicted in Figure 5.2g-i.

[1]The same rationale can be applied for link betweenness.

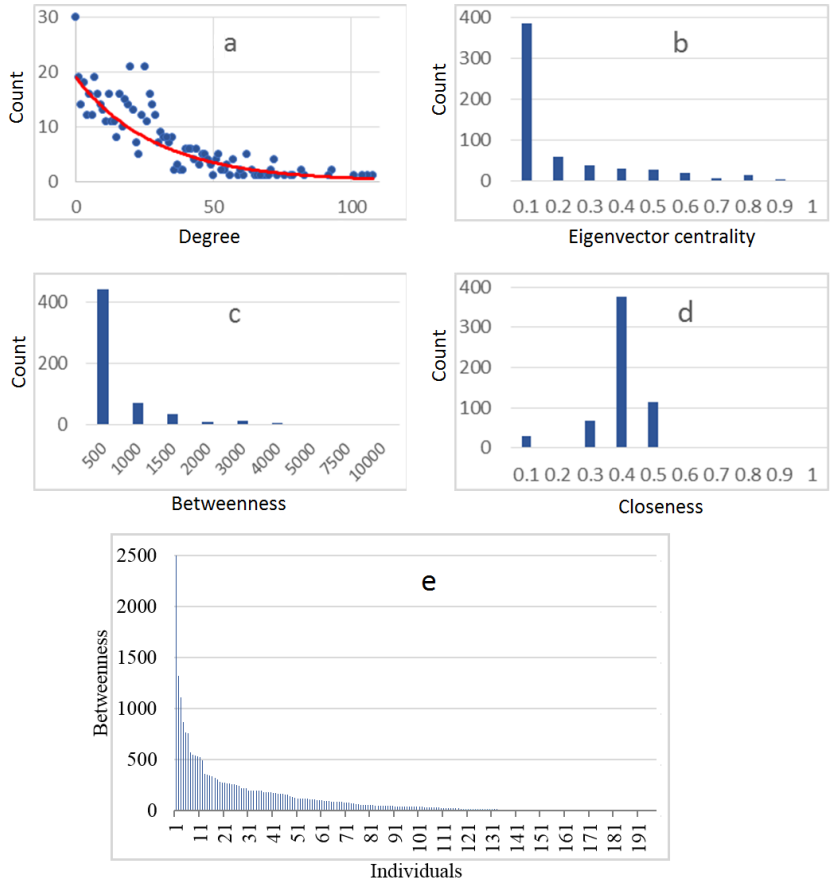


Figure 5.1.: **a.** Degree distribution for one representative Facebook network; the power law distribution of degrees is representative for such social networks, i.e. most persons have a low degree (left side), some persons have a moderately high degree (middle section), while only a few people have a very high degree (right side). **b.** Eigenvector centrality distribution for the same Facebook network; this metric shows a power law distribution, a specific feature of social networks [159]. **c.** Betweenness centrality in the Facebook network showing a power law distribution. **d.** Closeness centrality distribution a representative Facebook network which follows a particular Gaussian distribution with a cutoff value of 0.5; this is an empirically observed feature for friendship networks [260]. **e.** Illustrative example of a collaborative social network (Jazz musicians network [48]) which is characterized by a power-law distribution of betweenness.

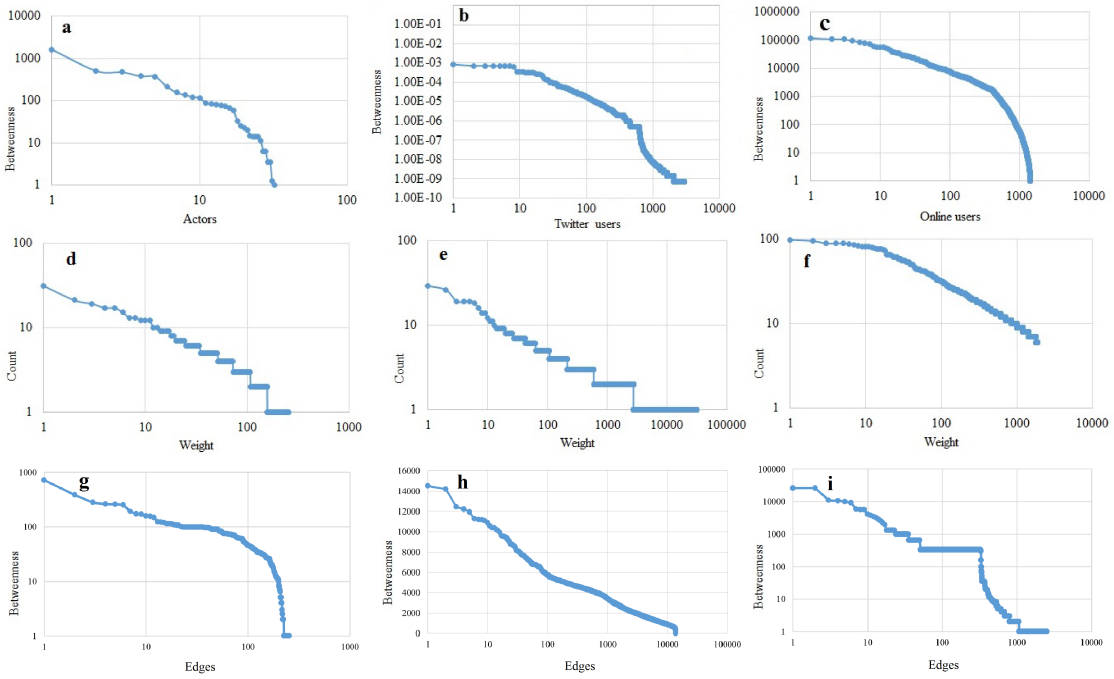


Figure 5.2.: Power-law distributions in log-log scale of: edge weights (**a**), node (**d**) and edge (**g**) betweenness in the Les Miserables actor network [147]; edge weights (**b**), node (**e**) and edge (**h**) betweenness in a Twitter network [157]; and edge weights (**c**), node (**f**) and edge (**i**) betweenness in an online social network [211].

Node betweenness – edge weight correlation

In order to investigate if there is a correlation between centralities distributions and the edge weights, the first step consists of filtering out all edges with small weights from the initial network G. Filtering means obtaining network G^* by eliminating all disconnected nodes, and using their initially measured betweenness to highlight the remaining nodes. Using this methodology, we are being left with the top 10% of edges because of the power-law distribution of weights. Subsequently, I apply a second filter by eliminating the nodes with low betweenness. Figure 5.3b highlights the strong correlation between nodes with high betweenness which are also linked through strong ties. The same conclusions are reached for the online network in Figure 5.4b.

Node-edge quantitative correlations

Motivated by the visual interpretation in Figure 5.2, the second step is to quantitatively measure the edge weight-node centrality and edge weight-edge centrality correlations. To this end, I define a correlation function as:

- **Definition 1 (Node fitness–edge weight correlation).** Given a weighted graph G and a filtered graph G^* when only top 10% of the weighted edges are kept, the correlation function c between the node fitness f and edge weight w is defined as:

$$c(f, w) = \frac{\sum_{v_i \in G^*} f(v_i)}{\sum_{v_j \in G} f(v_j)} \tag{5.2}$$

Therefore, in Equation 5.2, the correlation is defined as the ratio between the sum of fitnesses of each node v_i from the filtered network G^* and the sum of fitnesses of each node v_j from the original (unfiltered) network G. In my investigation, I consider the fitness f as either node degree or node betweenness. Therefore, I measure the sum of all degree and all betweenness values for all nodes in the original unfiltered network G, and then apply the mentioned methodology on the resulting filtered network G^*. The ratio between the remaining fitness and the initial total fitness represents the resulting correlation.

The full results displayed in Table 5.1 show the following correlations: edge weights – betweenness centrality, and edge weights – degree centrality. All the empirical data sets show a (much) stronger correlation to betweenness. The correlations are 80%, 65%, 67%, for Les Miserables, Twitter, and online network, respectively. In contrast, the nodes degree yields a much lower correlation for the same networks, respectively: 28%, 15%, and 8%.

By doing the same analysis for edge weight – edge betweenness correlation, I find that there is a weak implication of both properties on the same edge. Namely, as this method of comparison can be done directly - each edge has a weight and a fitness - I use the classic Pearson correlation and find that there is a correlation of: 0.02 for Les Miserables, 0.01 for Twitter, and -0.063 for the online social network.

5.4. Betweenness preferential attachment (BPA)

Using this betweenness-driven perspective, I define now the preferential attachment model as a node fitness for bridging with new nodes. Then, I propose a new fitness-based model which exploits the

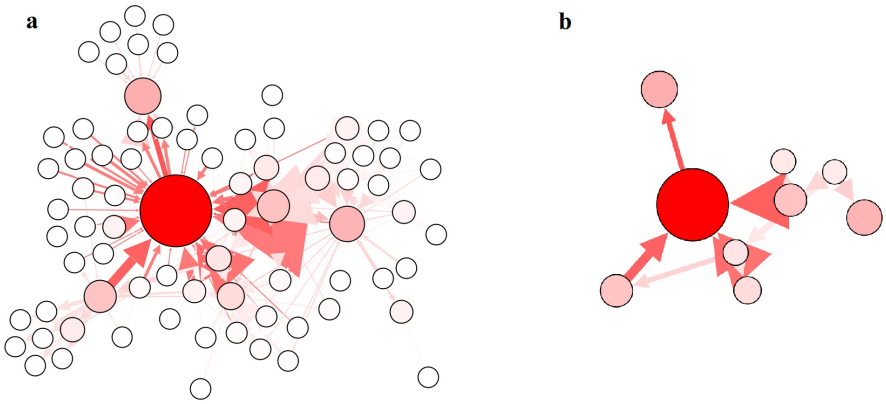


Figure 5.3.: Correlation between edge weights and node betweenness in the Les Miserables network [147]. **a.** All 77 nodes (actors) from unfiltered network G have their color and size highlighting betweenness. **b.** The filtered network G^* after keeping only the top 10% edges (in terms of weight). All the remaining connected nodes in G^* have high betweenness.

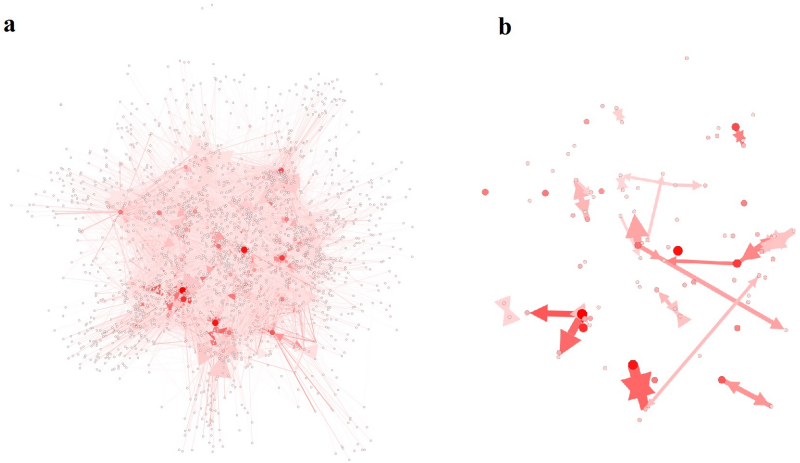


Figure 5.4.: Correlation between edge weights and node betweenness in the empirical weighted social online network [211]. **a.** All 1899 nodes (online users) in the unfiltered network G, which have their color and size highlighting betweenness. **b.** The corresponding filtered network G^* after keeping only the top 10% edges (in terms of weight). All remaining nodes have high betweenness values.

Table 5.1.: Correlation of degree centrality with edge weights, as well as correlation of betweenness centrality with edge weights in complex networks: Les Miserables [147], a Twitter network, and an online network [211]. The total fitness is obtained by summing up the fitnesses of all nodes (e.g. degrees) in the initial graph G, while the filtered fitness is obtained by summing up the fitnesses of the nodes remaining in G^* after the filtering procedure, as explained below and illustrated in Figures 5.3, 5.4.

		Filtered fitness	Total fitness	Ratio		Correlation
Les Miserables	Betweenness	3850	4802	0.801	→	**80%**
	Degree	142	508	0.279	→	28%
Twitter	Betweenness	0.009854	0.01512	0.652	→	**65%**
	Degree	9566	61672	0.155	→	15%
Online network	Betweenness	2.49E+06	3.68E+06	0.675	→	**67%**
	Degree	3228	40592	0.079	→	8%

existing correlation of betweenness and edge weights.

5.4.1. Unweighted BPA model

Originally, the "rich gets richer" concept is based on the idea that the probability of a newly added node to connect to an existing node is proportional to the degree of the target node [25, 10]. In this section, I generalize this concept by using the more generic term of (node) fitness. Consequently, if the fitness is considered to be based solely on node degree, then I obtain the original Barabasi-Albert model.

In order to explore the extended node fitness concept, I can generate networks which use the following measures for fitness: degree (*D*), betweenness (*B*), eigenvector centrality (*EC*), closeness (*C*), clustering coefficient (*CC*), as well as all combinations of two and three such measures. The BPA model algorithm is formally defined as follows:

BPA Model Algorithm
1) ***Growth***: Begin with an arbitrary connected graph G with V nodes. At every step, a new node u is introduced and connected to the v_i ($1 \leq i \leq V$) existing nodes in G.
2) ***Preferential attachment:*** The probability p_i that the new node u will be connected to any of the existing nodes v_i is proportional to the fitness f_i of node v_i, so that $p_i = f_i / \sum_j f_j$, where the sum is made over all nodes in the graph.

Fitness f_i can be a single node metric (e.g. degree, betweenness, clustering etc.) or a combination of two or more such metrics. In the latter case, the fitness becomes an equally weighted composite value, like e.g. $f_i = \frac{1}{2} f_i^{'} + \frac{1}{2} f_i^{"}$, where $f_i^{'}$ and $f_i^{"}$ are two distinct node metrics.

Similar to the Barabasi-Albert model [25], the nodes with high fitness f_i tend to become hubs and thus increase their degree at a faster rate. However, as the BPA model focuses on betweenness, the

initial hubs with high fitness will increase their degree, but this will trigger an increase in fitness of other (new) nodes with a lower degree. My algorithm can generate situations where newly added nodes are characterized by high fitness values; this is because they can create new shortcuts in the graph, and thus become new social hubs.

5.4.2. Weighted BPA model

Following the observations presented in Section *weighted social networks characteristics*, and after assessing the power of betweenness in an unweighted context, I set out to recreate a betweenness-fitness (*B*-fitness) model consisting of the empirically observed patterns. Namely, I add power-law distributed weights on those edges that are adjacent to nodes with higher fitness. As such, I try to correlate the observed match between weights and betweenness centrality. The algorithm for edge weight assignment according to the power-law fitness-weight correlation is presented as follows:

Dynamic Weighted BPA Model Algorithm (DWBPA)
1) ***Redistribute graph weights***: Begin with an arbitrary connected graph G with V nodes and bidirectional edges. A weight w_{ij} is added on all directed edges e_{ij} in the graph, so that w_{ij} is proportional to fitness f_j of the target node v_j. For each node v_i, all existing weights w_{ij} are normalized so that the outgoing weighted degree is 1.
2) ***Growth:*** At every step, a new node v_k is introduced and connected to v_i ($1 \leq i \leq V$) existing nodes in G. The probability p_i that v_k is connected to any of the existing nodes v_i is proportional to its fitness f_i, so that $p_i = f_i / \sum_j f_j$, where the sum is made over all nodes in the graph.
3) ***Dynamic redistribution:*** Once a new node v_k is connected to an existing node v_i, weights w_{ki} and w_{ik} are initialized with the normalized fitnesses f_i and f_k respectively. As the weighted outgoing degree of node v_i increases by w_{ik}, every other weight w_{ij} is rescaled with $-w_{ik}/n$, where n is the previous number of neighbors of node v_i.

Edge weights are updated in time, and whenever a weight decreases below 0, the edge is removed from the graph. The explained steps are illustrated in Figure 5.5, where a new edge with weight w_{1-7} is added from node 1 to node 7; this results in the rescaling of all edges from node 1 to the neighboring nodes 2, 3, 4, and 5 with $-w_{1-7}/4$.

5.5. Social network model assessment

Measuring the similarity between the social network models and real-world social networks is essential to correctly assess the state-of-the art synthetic topological models, as well as my proposed model. However, this job becomes rather complex, as it has to take into account all the relevant metrics [276, 260]. To this end, I use the network fidelity metric φ (phi); this metric was first introduced in [257], and has already been used for the analysis of other empirical datasets [29, 260]. A thorough description of this metric is presented in the supporting information section S1.

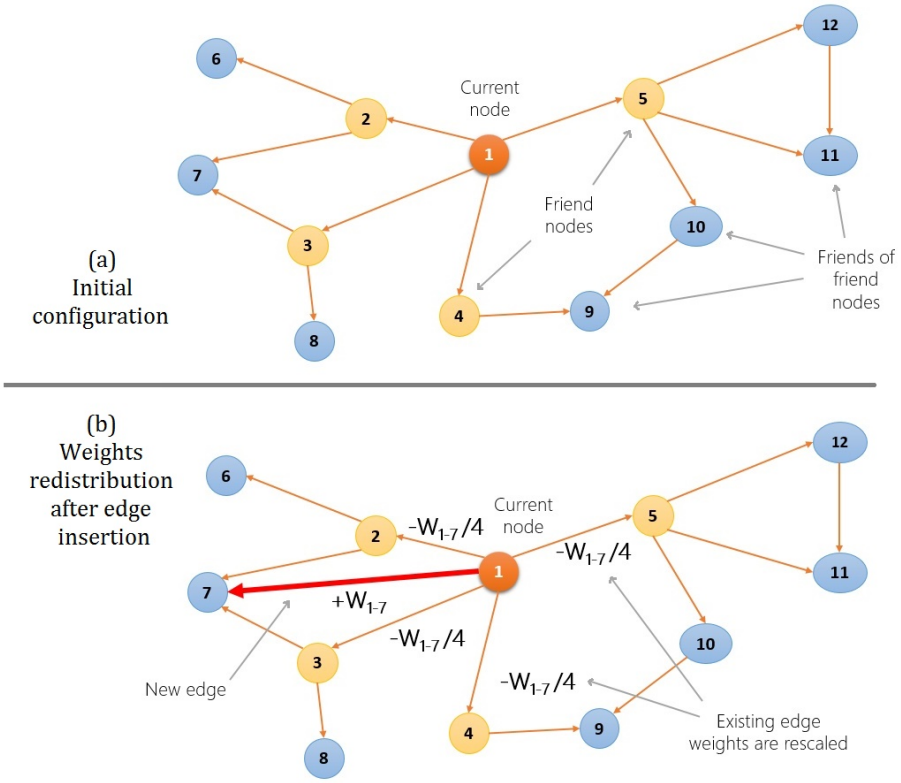


Figure 5.5.: Dynamic weights redistribution for the BPA growth algorithm. When a new edge (red) is added between nodes 1 and 7, it is assigned a weight (w_{1-7}) that must be proportionally subtracted from the other neighbors of node 1 ($w_{1-7}/4$). If an edge weight falls below 0, it is removed. The sum of weights for all outgoing edges of a node is always 1.

5.5.1. Real-world reference models

The analyzed data is obtained from the Stanford Large Dataset Collection [157], from H. Makse's dataset collection [271], as well as from my own Facebook data collection with over 100 different networks. Therefore, I rely on a diverse range of networks; I use a number of 100 Facebook datasets, with sizes ranging from 150 to 63,000 nodes, 10 Google Plus networks with a maximum size of 1600 nodes, one Twitter network of 35,000 nodes, two Slashdot networks (editor evaluated technology related news) of roughly 80,000 nodes, one Epinions network (*who-trusts-who* online social network) of 75,000 nodes, and a very large Pokec network (Slovakian social network) of 1.6 million nodes.

Measuring the representative graph metrics over the acquired data gives conclusive results for average degree (*AD*), average path length (*L*), average clustering coefficient (*CC*), modularity (*Mod*), network diameter (*Dmt*), and network density (*Dns*) [257, 260]. Regardless of network size, the measurements of all metrics for my Facebook datasets fall within representative intervals [260], with the mean specific values given in Table 5.2. Even though few systems are more diverse than the social

networks of humans, I can conclude that there is an underlying pattern in which all these networks fall. The same conclusions are observed with Twitter and Google Plus networks.

Table 5.2.: Specific values for average degree (*AD*), average path length (*L*), average clustering coefficient (*CC*), modularity (*Mod*), diameter (*Dmt*), and density (*Dns*) averaged for each of the 6 data sets: Facebook, Google Plus, Twitter, Slashdot, Epinions, and Pokec.

	AD	*L*	*CC*	*Mod*	*Dmt*	*Dns*
Facebook	19.82	2.481	0.266	0.468	8.5	0.05
Google Plus	12.15	3.9	0.404	0.44	12	0.035
Twitter	12.39	2.685	0.239	0.28	7	0.054
Slashdot	23.08	4.7	0.092	0.343	11	0
Epinions	13.41	5	0.261	0.445	14	0
Pokec	18.75	5.2	0.109	0.3	11	0

5.5.2. Assessing state-of-the-art models

Out of mentioned state-of-the-art networks in the *Background* section, I have chosen five models to use in my further discussion: geographic (regular mesh topology), small-world, scale-free, WSDD and cellular. The motivation for this choice lies within the topological diversity of each: the first three are fundamental models for network science, while the latter two combine properties of the the previous.

I start by generating a 10,000 node representative network for each of the main five models[2], and present the resulting graph metrics in Table 5.3. I need these parameters to compare the state-of-the-art social network models with the real-world datasets. The measured fidelity φ towards the real-world datasets of these synthetic models quantifies their accuracy in modeling realistic social networks. This way, I also create a comparison basis for my betweenness-centric model in terms of fidelity towards the real-world datasets.

The fidelity results relative to real-life social networks are presented in Table 5.4. Upon inspection, I learn that the fidelity achieved by the Barabasi-Albert model towards the empirical social networks it should approximate is quite low. Therefore, preferential attachment based on node degree (as node fitness) *is not very effective* in creating realistic social network models. This conclusion is also suggested by Abbasi *et al.* [1], Adamic and Huberman [126, 6], and it also extends to all other previous models (small-world, cellular, geographic, WSDD). Indeed, none of these social network (synthetic) models are able to accurately replicate the distinctive characteristics of real-world social networks. As revealed by my dataset analysis, these relevant characteristics are determined by the fact that node betweenness is power-law distributed and correlated with link weights. Moreover, the existing models do not account for edge weights and their power-law distribution, which is of major significance for the accuracy of social network analysis.

[2]The algorithms for generating all the analyzed networks have been implemented as Gephi plug-ins by the authors.

Table 5.3.: The basic metrics for five representative social network models. The numerical values for average degree (*AD*), average path length (*L*), average clustering coefficient (*CC*), modularity (*Mod*), diameter (*Dmt*), and density (*Dns*) are measured on the synthetically generated networks: small-world, scale-free, cellular, static-geographic, and Watts-Strogatz model with degree distribution. The size of each network is 10,000 nodes.

	AD	*L*	*CC*	*Mod*	*Dmt*	*Dns*
S-World	3.99	5.61	0.321	0.73	11	0.005
S-Free	3.12	4.60	0.015	0.62	10	0.003
Cellular	11.39	3.79	0.599	0.91	7	0.02
Geographic	6.63	3.34	0.065	0.52	8	0.013
WSDD	21.58	4.59	0.738	0.9	9	0.041

Table 5.4.: Networks fidelity φ of the synthetic networks towards the six empirical social network models: φ_{FB}, φ_{TW}, φ_{GP}, φ_{SL}, φ_{EP}, φ_{PK}. A higher φ-value means a higher fidelity towards the reference empirical model ($\varphi \to 1$), while a lower value means more dissimilarity ($\varphi \to 0$).

Model	Facebook	Twitter	Google Plus	Slashdot	Epinions	Pokec
Fidelity	φ_{FB}	φ_{TW}	φ_{GP}	φ_{SL}	φ_{EP}	φ_{PK}
S-World	0.628	0.561	0.698	0.628	0.745	0.647
S-Free	0.619	0.556	0.67	0.706	0.7	0.676
Cellular	0.632	0.659	0.746	0.553	0.653	0.55
Geographic	0.722	0.674	0.719	0.715	0.708	0.684
WSDD	0.688	0.563	0.684	0.653	0.626	0.618

5.5.3. Assessing the realism of BPA and DWBPA

BPA results for simple fitness

In order to obtain the results in Table 5.5, I have generated five synthetic scale-free networks using the BPA algorithm, with the fitnesses based on degree, betweenness, eigenvector, closeness, and clustering coefficient. Each result column in Table 5.5 represents the averaged values for these synthetic networks. Since the average size of my empirical datasets is in the order of thousands, I also generate 10,000-node scale-free networks, the same size as the state of the art models in Table 5.3. However, to extend my observations, I validate these measurements on larger networks, ranging up to 100,000 nodes.

The same six graph metrics are measured on all generated networks: AD, L, CC, Mod, Dmt, and Dns, but I express the fidelity in terms of the most relevant metrics: L, CC, and Mod. The interpretation of the results is facilitated through the fidelity metric as I compare each resulting column to Facebook (φ_{FB}), Twitter (φ_{TW}), and Google Plus (φ_{GP}) empirical references.

As it can be observed in Table 5.5, using betweenness as a fitness function recreates the most faithful social model, with $\varphi_{FB} = 0.83$. The other fitnesses prove to create much less accurate models, with

Table 5.5.: Experimental values for average degree (*AD*), average path length (*L*), average clustering coefficient (*CC*), modularity (*Mod*), diameter (*Dmt*), and density (*Dns*) obtained for synthetically generated scale-free networks using five different fitnesses: degree (*D*), betweenness (*B*), eigenvector centrality (*EC*), closeness (*C*), clustering coefficient (*CC*). The bottom lines contain the fidelity of each model towards the empirical Facebook, Twitter, and Google Plus references (φ_{FB}, φ_{TW}, and φ_{GP} respectively) using *L*, *CC* an *Mod* as comparison criteria.

	D	***B***	*EC*	*C*	*CC*
AD	3.128	2.948	3.044	3.032	3.156
L	4.311	3.256	4.124	5.623	3.865
CC	0.03	0.253	0.025	0.006	0.008
Mod	0.604	0.592	0.613	0.621	0.588
Dmt	9	6	9	13	9
Dns	0.006	0.006	0.006	0.006	0.006
φ_{FB}	0.656	**0.83**	0.658	0.584	0.671
φ_{TW}	0.539	**0.747**	0.545	0.478	0.559
φ_{GP}	0.717	**0.776**	0.726	0.635	0.748

fidelities around 60-65% in the case of the Facebook model. The other empirical reference models merely strengthen the same observation, with the *B*-model being roughly 30% more similar to the empirical references than the other metrics used as fitness. Upon a careful inspection, I note that the *B*-fitness (i.e. betweenness-based fitness) model is the only one to also optimize the clustering coefficient. To uphold the numerical results for the *B*-fitness model of 10,000 nodes from Table 5.5, I show in Table 5.6 how the high fidelity scales with increasing network size. To that end, Table 5.6 contains fidelity results for *B*-fitness networks of sizes 10,000-100,000.

Because all six empirical models prove the same conclusion, and because of the wider acceptance of the Facebook model as being a realistic replica of real-world social networks [125], I further compare my synthetic results with the real-world FB datasets solely.

Table 5.6.: Fidelity of proposed unweighted model (BPA) for network sizes: 1,000-100,000 nodes. The proposed model is compared against the Facebook (φ_{FB}), Twitter (φ_{TW}), Google Plus (φ_{GP}), Slashdot (φ_{SL}), Epinions (φ_{EP}), and Pokec (φ_{PK}) unweighted networks. The fidelity is calculated by taking into consideration the three columns: *L*, *CC* and *Mod*.

	AD	*L*	*CC*	*Mod*	*Dmt*	*Dns*	φ_{FB}	φ_{TW}	φ_{GP}	φ_{SL}	φ_{EP}	φ_{PK}
BPA-1K	*2.948*	3.256	0.253	0.592	*6*	*0.006*	0.83	0.75	0.78	0.57	0.82	0.55
BPA-10K	*3.724*	4.385	0.181	0.757	*8*	*0*	0.65	0.60	0.70	0.63	0.75	0.62
BPA-50K	*3.382*	6.203	0.168	0.785	*8*	*0*	0.58	0.52	0.61	0.58	0.70	0.62
BPA-100K	*1.902*	24.923	0.167	0.788	*8*	*0*	0.47	0.41	0.45	0.39	0.50	0.41

BPA results for complex fitness

The results corresponding to other synthetically-generated scale-free networks using combinations of two fitnesses (*D-B*, *D-EC*, *D-C* etc.) and three fitnesses (*D-B-EC*, *D-B-C* etc.) are shown in Tables 5.7 and 5.8 respectively. By inspecting the results for single fitness and composite fitness I strengthen my claim: the model which uses betweenness as a fitness function is the most accurate (Table 5.5, in bold), and the most faithful in terms of φ. When analyzing the composite fitness scenarios only, I find that the highest yielded fidelities correspond to the combinations in which betweenness is used (bolded values in Tables 5.7 and 5.8); while all other cases decrease fidelity.

Table 5.7.: Experimental values for graph metrics obtained for synthetically generated scale-free networks using composite fitnesses based on two metrics with equal weights (50-50%): *D-B* (degree-betweenness), *D-EC* (degree-eigenvector centrality) etc., using all combinations with similar notations. The bottom line contains the fidelity of each model towards the empirical Facebook reference using *L*, *CC* an *Mod* as comparison criteria; the highest fidelity values are bolded.

	D				***B***			*EC*		*C*
	B	*EC*	*C*	*CC*	*EC*	*C*	*CC*	*C*	*CC*	*CC*
AD	3.028	3.076	3.104	3.224	3.1	3.052	3.1	3.128	3.08	3.08
L	3.667	4.401	5.059	4.053	3.723	4.589	3.735	4.771	4.095	4.384
CC	0.113	0.029	0.021	0.01	0.098	0.012	0.018	0.014	0.008	0.002
Mod	0.604	0.616	0.615	0.582	0.604	0.613	0.59	0.601	0.607	0.603
Dmt	8	9	11	8	8	11	7	12	8	10
Dns	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
φ	**0.695**	0.617	0.59	0.641	**0.684**	0.605	0.658	0.604	0.628	0.614

The presented synthetic results strengthen the claim that betweenness alone is the key factor to realistic modeling of social emergence in social networks. The visualizations for the obtained synthetic networks are given in Figure 5.6. This visual rendering illustrates that the *B*-fitness network has local clustering as well as assortativity between the important nodes, unlike the other three networks.

The high fidelity of the *B*-fitness model, and of all models involving betweenness, show that there is an inherent pattern of natural metric-alignment, as all other synthetic network metrics fall within their realistic thresholds. This effect is not visible for any other single-metric-fitness model, nor is it that obvious for multiple-metrics-fitness models. Betweenness alone proves to be the sufficient requirement to model and generate realistic social networks.

DWBPA results

By applying the dynamic weighted algorithm, I generate synthetic weighted networks based on BPA. Table 5.9 shows representative values for the synthetic model, the Facebook reference, and the Les Miserables weighted network. As can be seen, the fidelities are increased even further, reaching a 93% in the case of the second empirical network reference. To validate my results, I extend the same analysis on DWBPA networks of 10,000 - 100,000 nodes (see Table 5.9). Due to the larger sizes, the

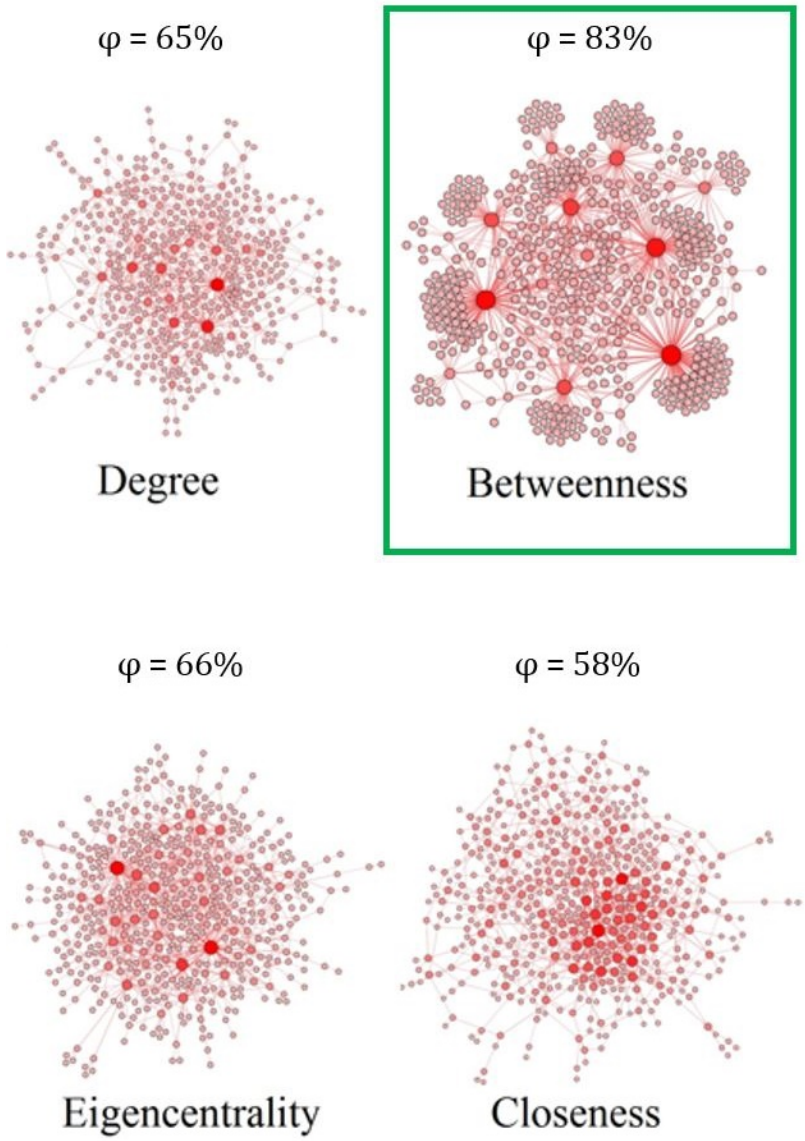


Figure 5.6.: Unweighted scale-free networks synthetically generated using four different centrality measures as node fitness for preferential attachment: degree, betweenness, eigencentrality, and closeness, along with the corresponding fidelity values towards Facebook empirical reference. The nodes are colored and sized proportionally to their fitness in each respective network.

Table 5.8.: Experimental values for graph metrics obtained for synthetically generated scale-free networks using composite fitnesses based on three metrics with equal weights (33-33-33%): *D-B-EC* (degree-betweenness-eigenvector centrality), *D-B-C* (degree-betweenness-closeness) etc., using the introduced notations. The bottom line contains the fidelity of each model towards the empirical Facebook reference using *L*, *CC* an *Mod* as comparison criteria. The highest fidelity values are bolded.

	D						***B***			*EC*
		B		*EC*		*C*	*EC*		*C*	*C*
	EC	*C*	*CC*	*C*	*CC*	*CC*	*C*	*CC*	*CC*	*CC*
AD	3.14	3.1	3.128	3.28	2.992	3.204	3.164	3.128	3.188	3.128
L	3.884	4.234	3.938	4.535	4.275	4.446	4.245	3.856	4.039	4.447
CC	0.064	0.035	0.015	0.007	0.007	0.003	0.025	0.009	0.009	0.003
Mod	0.595	0.607	0.59	0.582	0.625	0.589	0.595	0.589	0.579	0.603
Dmt	8	9	9	10	9	10	9	7	9	11
Dns	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
φ	**0.664**	0.63	**0.646**	0.619	0.612	0.618	0.631	**0.649**	**0.643**	0.612

fidelities decrease proportionally, from 93% to 67% (100K nodes), compared to Les Miserables. The respective fidelity results from Table 5.9 demonstrate the increase in realism of my model compared to the existing models in literature, shown in Table 5.4.

Table 5.9.: Fidelity of proposed model with power-law distributed weights correlated with high fitness nodes (DWBPA) for network sizes: 1,000-100,000 nodes. The proposed model is compared against the Facebook reference model and (φ_{FB}) the Les Miserables (φ_{LesM}) and Twitter (φ_{TW}) weighted networks. The fidelity is calculated by taking into consideration the three columns: *L*, *CC* and *Mod*.

	AD	*L*	*CC*	*Mod*	*Dmt*	*Dns*	φ_{FB}	φ_{LesM}	φ_{TW}
DWBPA-1K	*7.06*	2.19	0.275	0.547	*4*	*0.071*	0.90	0.93	0.74
DWBPA-10K	*8.33*	3.29	0.254	0.514	*7*	*0.001*	0.87	0.87	0.77
DWBPA-50K	*8.71*	4.507	0.249	0.556	*9*	*0*	0.78	0.82	0.68
DWBPA-100K	*8.74*	18.66	0.249	0.571	*7*	*0*	0.63	0.67	0.53
Les Miserables	*3.3*	2.641	0.287	0.565	*5*	*0.087*			

Having illustrated the increase in realism brought by correlating node betweenness with edge weights, I further illustrate the situation in which edges are artificially weighted using other distributions than power-law. I believe it is paramount to prove that power-law distribution of edges weights is what best corresponds to real world social networks. Towards this end, I consider three different scenarios of placing weights on these edges:

- Unweighted (all weights = 1, uniform distribution): similar to most state of the art research,

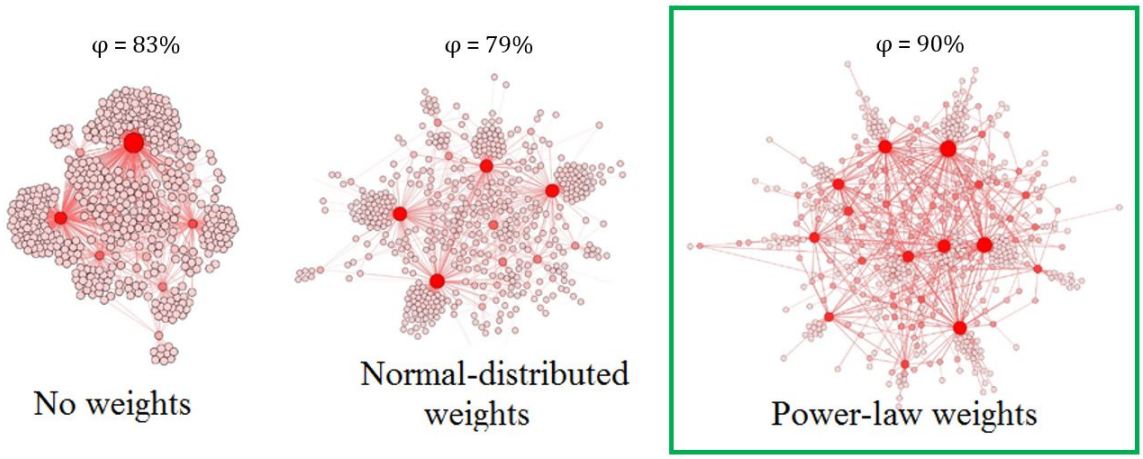


Figure 5.7.: Scale-free networks with *B*-fitness using three weight models: no weights (uniform), normal distributed, and power-law, along with corresponding fidelity values towards the Facebook empirical reference. The nodes are colored and sized proportional to their betweenness in each network.

which lacks realistic accuracy;

- Randomly placed and normally distributed weights (0 <weights $\leq$ 1, mean weight = 0.5);
- Randomly placed and power-law distributed weights (0 < weights $\leq$ 1, mostly small weights, with very few large weights).

Figure 5.7 displays the graphical results for the three described scenarios. From left to right, the unweighted model has the already measured fidelity $\varphi = 0.83$, the normal-distributed model has a lower fidelity of $\varphi = 0.79$, and the power-law model has the best fidelity of $\varphi = 0.90$ towards the Facebook empirical reference. Although the differences in fidelity may not seem significant, they are well outside the noise margin [260].

The weighted model is compared to both the Les Miserables weighted and Facebook unweighted references. I bring together the unweighted and weighted networks in order to prove the leap in realism granted by adding weights properly. Indeed, this is the limitation of most state-of-the-art studies which consider unweighted social networks. This proves that not only normal-distributed edge weights are non-realistic for social networks, but they also lower the realism of the model in terms of fidelity φ. Having a normal distribution of weights balances out the preferential attachment that lies underneath the network. Second, even if the placement is not correlated with node betweenness (i.e. random), adding power-law distributed weights to edges increases fidelity by 7%, up to a significant 90% when compared to the Facebook model.

5.5.4. Summary of experimental results

To summarize the experimental procedure of my study, I revisit its main results:

- Starting from the current state of the art, I compute the fidelity of the Barabasi-Albert scale-free model $\varphi = 0.65$, in comparison with the Facebook model.
- I then introduce the BPA model based on betweenness preferential attachment which increases the fidelity to $\varphi = 0.83$; applying this model to a power-law distributed weighted network further increases the fidelity to $\varphi = 0.90$.
- Finally, by adding an algorithm for dynamic correlation between the node fitness (betweenness) and adjacent edge weights, I can improve the fidelity of the proposed DWBPA synthetic model to $\varphi = 0.93$.

Putting it all together, by distributing and correlating the node fitness and edge weights though power-law preferential attachment, the increases in realism compared to the Barabasi-Albert model are +46%, +33%, +12% in terms of network fidelity relative to the Facebook, Twitter, respectively Google Plus empirical models.

5.6. Socio-psychological interpretation

The results from the previous section indicate that my BPA algorithms provide a much more accurate social network topological model, in comparison with the state-of-the-art models. However, I believe that the weighted and betweenness-driven preferential attachment approach transcends the mere topological view of the problem.

To get to the bottom of this, from a social standpoint, I need to address two aspects regarding the social network evolution. First, every person has weighted ties, with stronger ties being more influential than the weaker ones. Two natural questions to ask are how do the tie strengths get assigned to newly emerging ties, and how do tie strengths evolve over time? Second, everyone needs social recognition as we strive for status in our lives. The naturally emerging question is with whom, conscientiously or subconsciously, do we prefer to interact as we continuously want to improve our social status?

The answers to these questions come from social networks analysis and social psychology studies, where people are perceived as social creatures who strive for social recognition, validation, approval and fame [221, 5, 59, 154, 184]. Thus people tend to connect to two types of individuals: Popular people in their communities (i.e. typically they have high degree centrality), and influential persons who can bring people across communities together (i.e. high betweenness). While the first type of interconnection is related to the popularity of individuals within delimited communities, it appears to be a side-effect of the more important (second) type of interconnectivity.

State of the art has previously identified that social networks have apparent (degree) assortative mixing, while, on the other hand, technological and biological networks all appear to be disassortative in nature [184, 138]. The study in [138] explains that this is because most networks have a tendency to evolve, unless otherwise constrained, towards their maximum entropy state – which is usually disassortative. A similar debate was introduced by Borondo et al. based on the concepts of meritocracy versus topocracy [44]. The authors discuss the critical point at which value in society changes from being based on personal merit, to being based on social position, status, and acquaintances. My perspective on this issue concerns the balance between friends with less influence and ones with more influence than us; this translates into betweenness assortativity. Indeed, connecting

to persons with high betweenness and increasing our tie strength with them (through, say, a stable social relationship), we ourselves become, in turn, more influential bridges in social lives. This propagation of influence determines other persons, with lower betweenness, to interact with us and direct more tie strength towards us. I argue that there is a critical point at which we switch from being predominantly initiators of favorable social ties to becoming predominantly receivers of ties that favor both us and the initiator.

Towards this end, I introduce and further discuss the concept of *social evolution cycle*, which is based on betweenness preferential attachment and betweenness assortativity. Moreover, I consider this to be a true fingerprint of all social networks. This overarching concept differs from the state of the art because it revolves around betweenness assortativity rather than degree assortativity [184, 52, 220, 138, 297]. As a consequence, my perspective can be interpreted as a model in which social agents become more influential over time by increasing their own betweenness; indeed as they bridge the newly found communities with their own communities, they act as initial social hubs in communication. The exhibition of one social agent's desire to increase its betweenness is two-fold: it attracts new ties (i.e. increase in degree), and it creates stronger ties (i.e. edge weight); this process continues for the next generation of aspirants to climb on the social ladder.

These principles are illustrated in Figures 5.8 and 5.9. Figure 5.8 portrays the two options a social agent has in its goal to improve the social status. I consider that initiating contact with more influential agents leads to an increase in influence, which is followed by a natural increase in tie strengths – and not the other way around. This process is detailed and explained in Figure 5.9.

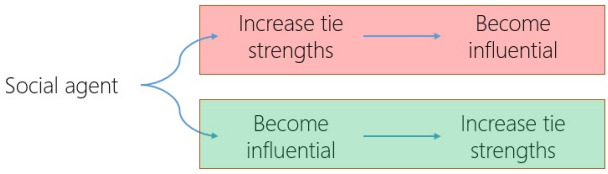


Figure 5.8.: One of the two envisioned ways for a social agent to increase its status. The first choice (depicted in red) relies on forcing tie strengths to increase first, then followed by an increase in influence. The second choice (depicted in green) relies on increasing one's influence, which will in turn trigger an increase in tie strengths. I consider the second choice as the plausible social process.

5.7. Conclusion

In this chapter, I have shown that betweenness preferential attachment is a fundamental concept in understanding the emergence of social networks; indeed, betweenness it is the main driving force behind how people interact, create new social bonds, and evolve their social status. My comprehensive social network analysis, based on fidelity assessments, has found that in real-world social networks weights and betweenness are both power-law distributed and also interdependent. Finally, I have shown that betweenness implies the formation of strong ties around the influential nodes.

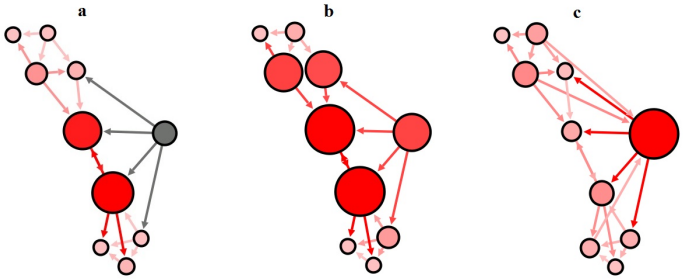


Figure 5.9.: An intuitive explanation of the social evolution cycle. All nodes are colored and sized proportional to their betweenness centrality (influence). **a.** A non-influential actor (gray) initiates social contact with other actors equal or more influential than himself. **b.** This action leads to a natural increase in influence (betweenness). **c.** Other nodes with less influence start connecting to the initial node. At this point, the initial node has become a predominant receiver of ties.

Based on my observations, betweenness centrality can be considered as a fitness function for explaining evolvable social ties. My results represent an important step towards modeling and understanding the fundamental mechanism for the formation of social ties. I believe my work paves the way towards a better understanding of the mechanisms that lie behind betweenness centrality and social network dynamics.

6. Tolerance-based interaction: A new model targeting opinion formation and diffusion in social networks

One of the main motivations behind social network analysis is the quest for understanding opinion formation and diffusion. Previous models have limitations, as they typically assume opinion interaction mechanisms based on thresholds which are either fixed or evolve according to a random process that is external to the social agent. Indeed, my empirical analysis on large real-world datasets such as Twitter, Meme Tracker, and Yelp, uncovers previously unaccounted for dynamic phenomena at population-level, namely the existence of distinct opinion formation phases and social balancing. I also reveal that a phase transition from an erratic behavior to social balancing can be triggered by network topology and by the ratio of opinion sources. Consequently, in order to build a model that properly accounts for these phenomena, I propose a new (individual-level) opinion interaction model based on tolerance. As opposed to the existing opinion interaction models, the tolerance model assumes that individual's inner willingness to accept new opinions evolves over time according to basic human traits. Finally, by employing discrete event simulation on diverse social network topologies, I validate my opinion interaction model and show that, although the network size and opinion source ratio are important, the phase transition to social balancing is mainly fostered by the democratic structure of the small-world topology.

"Quiet people have the loudest minds."

 Stephen Hawking

6.1. Motivation

Social networks analysis is crucial to better understand our society, as it can help us observe and evaluate various social behaviors at population level. In particular, understanding the social opinion dynamics and personal opinion fluctuation [128, 106, 101, 267, 117, 62] play a major part in fields like social psychology, philosophy, politics, marketing, finances and even warfare [82, 106, 217, 92]. Indeed, the dynamics of opinion fluctuation in a community can reflect the distribution of socially influential people across that community [142, 128, 193]; this is because the social influence is the ability of individuals (agents) to influence others' opinion in either one-on-one or group settings [180, 276, 82, 182]. Without social influence, the society would have an erratic behavior which would be hard to predict.

Existing studies on opinion formation and evolution [4, 292, 106, 101, 267, 62, 128, 112, 233] rely on the contagion principle of opinion propagation. However, such studies offer limited predictability and realism because they are generally based on opinion interaction models which use either fixed thresholds [75, 133, 162, 54, 74], or thresholds evolving according to simple probabilistic processes that are not driven by the internal state of the social agents [88, 76, 161] . To mitigate these limitations, I reveal some dynamical features of opinion spreading that previous models fail to identify. The consistent and recurring real-world observations are then explained by introducing a new social interaction model which takes into account the evolution of individual's inner state. I finally validate the proposed model by analyzing empirical data from Twitter, MemeTracker and Yelp, and by using my opinion dynamics simulation framework - SocialSim [253] - which includes multiple complex topological models, as well as customizable opinion interaction and influence models. Consequently, my main contributions are threefold:

1. **Observations** made on real-world datasets: I identify four distinct phases in opinion formation; this aspect is *not* captured by the existing models [248, 162, 4, 55, 112, 88] although previous research [124] notices that there are some stages of opinion evolution. The succession of opinion formation phases is critical to the *social balancing* phenomenon (i.e. the general opinion becomes stable despite constant local oscillations). I also identify a *phase transition* from an unstable opinion to social balancing which is related to the dynamics of opinion formation phases.

2. **Modeling** the opinion dynamics: I propose a new graph and threshold based interaction model with stubborn agents [3] which is able to reproduce the phenomena that I observe in real-world datasets. Inspired by social psychology, my new model assumes that individual's willingness to accept new opinions (i.e. tolerance) changes over time according to its inner state.

3. **Simulating** opinion dynamics: I provide new results that validate the newly proposed tolerance model via my discrete-event simulator SocialSim [253]. The analysis I provide reveals the deep connection between social balancing and the relevant parameters of social networks such as network size, topology, and opinion source ratio (i.e. stubborn agents distribution) [4]; this correlates well with my empirical observations on large social networks.

Taken together, these new contributions show that opinion dynamics in social networks exhibit specific patterns that are influenced by the network size and ratio of stubborn agents (which I consider to be opinion sources), but are mostly dependent on the underlying network topology. Consequently, my findings can be used to improve the understanding and predictability of social dynamics.

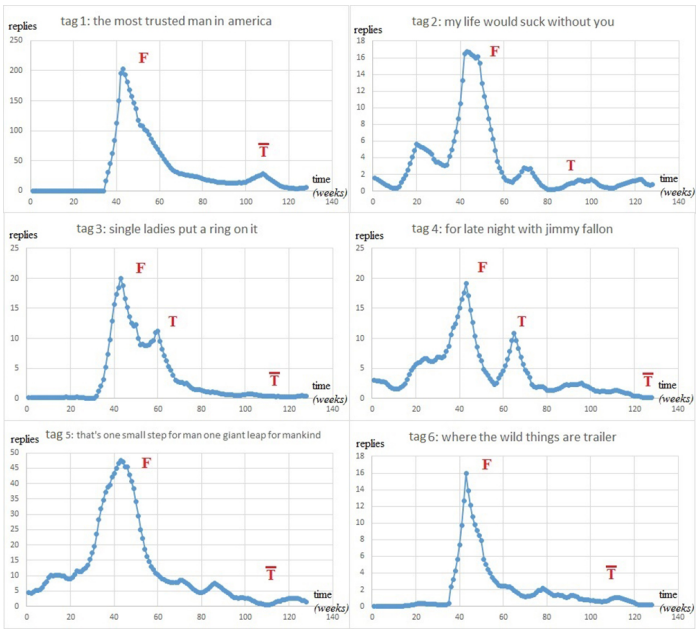
6.2. Results

6.2.1. Opinion formation phases and social balancing

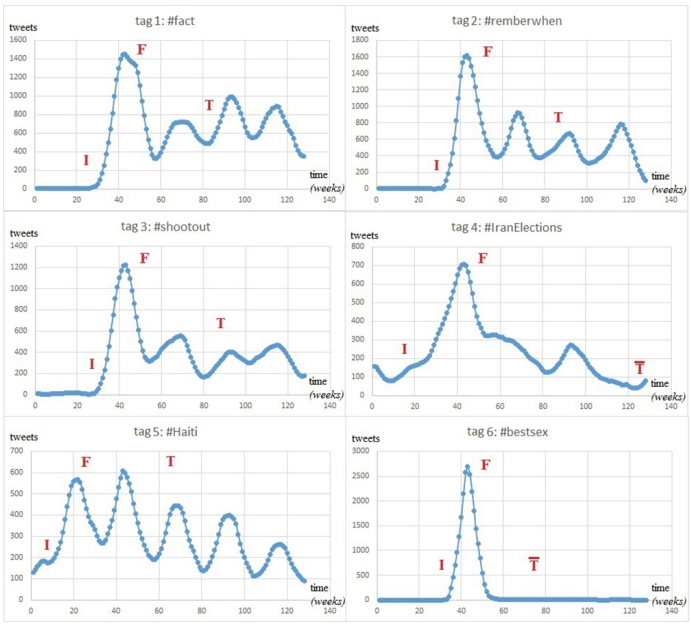
By analyzing data on opinion evolution using Twitter and MemeTracker hashtags, as well as user reviews and votes for local businesses from Yelp, I identify unique temporal patterns in all these datasets. Explaining and then modeling these patterns can improve our understanding of opinion formation and diffusion in social networks.

Figure 6.1 displays the popularity of six hashtags on MemeTracker and Twitter, expressed as posts/time evolution (posts are replies and tweets). Based on the observed fluctuations, I identify the following phases in opinion formation: an initiation phase (I) when new opinions are injected into the social network and the number of replies starts to increase rapidly; a fusion phase (F) when the opinion dynamics reaches a maximum and different opinions start to collide; a tolerance phase (T) which represents a fluctuating yet convergent behavior; and, occasionally, an intolerance phase ($\overline{T}$) when the fluctuations of opinion decrease and converge towards zero. Based on network topology and/or ratio of opinion sources, the diffusion process may reach the fourth phase of intolerance. Opinion sources, or stubborn agents [3, 4], are agents within the social network (i.e. Twitter or Yelp users) who try to instill a certain opinion by influencing their peers; they are represented by those people within the network who hold strong opinions that do not change over time. The concentration of opinion sources is expressed as their ratio relative to the entire population.

Additionally, the analysis of Twitter results in Figure 6.1b shows that tags 1-3 all exhibit a clear F phase (first spike), then they enter a balanced oscillation (T phase). This evidence supports the empirical observation of a phenomenon that I call *social balancing*, i.e., oscillations at microscopic scale of individuals opinion become stable and predictable at the macroscopic scale of the society. As such, social balancing is defined as the succession of $I - F - T$ phases, whereas social imbalance occurs if either the society does not reach T or, after reaching T, it decays into a $\overline{T}$ phase. For example, tag 4 (#Iran) in Figure 6.1b has a shorter, more abrupt oscillation. In this case, I consider that the number of opinion sources is not high enough (i.e. above a critical threshold) for social balancing to happen. Tag 5 (#Haiti) has a longer F phase because of the (probably) very high concentration of opinion sources. Indeed, the 2010 Haiti earthquake was breaking news so there were many outbreaks of opinion, scattered across the globe, resembling a random network topology of sources of opinion; nonetheless, for tag 5 the society reaches social balance. Tag 6 is an example of social imbalance with a decisive crystallization of just one opinion, as there is no T phase.



(a)



(b)

Figure 6.1.: Opinion dynamics for six popular hashtags on: **a.** MemeTracker. Tags 1, 5, and 6 all exhibit the fusion phase (F) (opinion spike), then they slowly converge towards intolerance. Tags 2 and 4 have an initial spike before the F phase and more oscillations after F. The tolerance phase is depicted in tag 2 as the oscillation exists, but it is balanced. Tag 3 exhibits a second spike after the F phase, then enters the intolerance phase; as such, social balancing does not occur in tag 3. **b.** Twitter. Tags 1, 2, 3 and 5 exhibit the fusion phase F (first opinion spike), then they oscillate during the tolerance phase keeping social balance. Tags 4 and 6 show an example of convergence towards the intolerance phase, as social balancing does not occur.

It can be debated whether hashtag dynamics could be equated to opinions dynamics. People use them driven by different forces, and many are not associated to changes of opinions or opinions formation. Indeed, in the Twitter and MemeTracker the actual state of the society s cannot be deduced, as the tweets are not processed (e.g. using sentiment analysis), so I interpret the number of replies as a measure of opinion change: more replies means more opinion injected in the society. When previously unopinated people reply or retweet, they do so because they have reached a clear opinion on a particular subject and they feel the need to express it by broadcasting the related story. The number of replies (OY axis) at a given moment (OX axis) corresponds to users expressing (injecting) opinion in the society. As such, the Twitter and MemeTracker datasets are supporting the observations related to the opinion change ω and not to those related to the opinion state s.

In case of the Yelp dataset, the state s is the current average number of stars awarded by users; it represent the opinion towards a business. The variation of s between two time moments $t-1$ and t is the opinion change ω. This information can be found in Figures 6.4-6.6 for the experiments' opinion change representation. As Yelp users are considered nodes in a social network, the opinion dynamics on local businesses will be affected by the underlying social network links. The social agents that influence opinion are linked and interact at least through the Yelp platform. Nonetheless, as Yelp is well-known for hosting social events for reviewers, I believe that these social ties are even stronger.

Using the Yelp context, I explain how the opinion formation phases (I-initiation, F-fusion, T-tolerance and $\overline{T}$-intolerance) are detected. For each business, I have automatically detected all spikes in the number of total votes (interpreted as opinion sources which never change their state, or stubborn agents SA) and have corroborated these with the point at which the state (average stars) has a variation of less than 1 star between maximum and minimum stars awarded. The reason behind considering the variation interval is that 1 star is the psychological threshold represented by the unit of measurement. Using an algorithmic explanation, I describe the pseudocode for detecting three points of interest - A (start of convergence of state), B (spike in SA concentration just before the convergence of state), C (spike in opinion change just after the spike in SA).

Algorithm 6.1 Detecting B: start of convergence in stars on OX-axis.

findt_B so that:
  maximum $(s(k))$-minimum$(s(k))< 1$ for all $t_B \leq k < t_{max}$
assign$B(t_B,\ s(t_B))$

Algorithm 6.2 Detecting A: spike in SA just before convergence of stars.

findspike[ ] := list of all local maximums in the number of total votes
findspike$[t_A]$ so that:
  $t_A < t_B$ (last spike before t_B)
assign$A(t_A,\ SA(t_A))$

Algorithm 6.3 Detecting C: spike in opinion change just after spike in SA.

findspike[ ] := list of all local maximums in the opinion change
findspike[t_C] so that:
  $t_B < t_C$ (first spike after t_B)
assign$C(t_C, \omega(t_C))$

By automatically performing this methodology on all 2331 businesses, I find that the average distance between the closest spike on the time axis OX (point A in the example from Figure 6.2) which occurred before the convergence of stars (i.e. point B in Figure 6.2, where the variation of awarded stars becomes lower than 1 star) is $d_{conv} = 4.131$ time units. Distance d_{conv}\$ is relatively small with respect to the observation interval of 100 time units or days, suggesting the fact that spikes in SA trigger a (shortly delayed) convergence of stars.

Further, I show that the spike in SA (point A) also triggers a maximum spike in the opinion change (point C). By running this methodology on all businesses, I obtain an average distance between the spike in SA and maximum spike in opinion change of $d_{fusion} = 4.828$ time units. These statistical results support the fact that spikes in SA trigger a maximum spike in opinion change.

Moreover, when I corroborate the average delays between the spike in SA and spikes in stars and opinion change, namely 4.131 and 4.828 time units, I can conclude that the convergence of opinion and the fusion phase are distanced, on average, by only $d_{corr} = 0.697$ time units. Backed up by this data, I can admit that the convergence of opinion (point B) and the triggering of the fusion phase (point C) are closely correlated.

Inspired by a similar approach on Twitter data [155], I have conducted a statistical analysis on all three datasets. Using all datasets from Twitter (1000 hashtags), MemeTracker (1000 keywords) and Yelp (2331 businesses) I have automatically detected the following characteristic phases in the opinion dynamics:

1. **Fusion** (2^{nd} phase) is the spike centered around the previously detected point $C(t_C, \omega(t_C))$ with t_C being the time projection and $\omega(t_C)$ the corresponding opinion change of point C. For convenience, we will refer to the local spike in opinion change $\omega(t_C)$ as fs (fusion spike).

2. **Initiation** (1^{st} phase): starting from time $k = 0$ (on OX-axis), find $0 \leq k < t_C$ so that $\omega(k) < 0.5 \cdot fs$ AND $\omega(k+1) > 0.5 \cdot fs$. In other words, time k represents the first point at which the opinion change ω exceeds 50% of the fusion spike fs. We have used this threshold value because it represents the half amplitude of the fusion phase, which it precedes.

3. **Intolerance** (4^{th} phase): starting from time $k = t_{max}$ (the highest registered time on the OY-axis), find $t_C < k < t_{max}$ so that $\omega(k) < 0.1 \cdot fs$ AND $\omega(k-1) > 0.1 \cdot fs$. In other words, time k represents the first point, from end to beginning of time, at which ω exceeds 10% of the fusion spike. We consider that a social network reaches intolerance if tolerance $\theta < 0.1$, so we use the 10% threshold for opinion change. Any higher than 10%, and opinion change is still in the tolerance phase, any lower, and opinion change is likely to converge towards 0.

4. **Tolerance** (3^{rd} phase): starting from time $k = t_C + 1$ (start of social balance), find $t_C < k < t_{max}$ so that $\omega(k) > 0.1 \cdot fs$ AND $\omega(k+1) < 0.1 \cdot fs$ (end of social balance). In other words, time k represents the point at which ω decreases below the 10% threshold which we consider a transition into the intolerance phase.

Figure 6.2.: Representative example for the evolution of reviews count and reviews votes for a popular businesses on Yelp. The ratio of review votes with respect to the review count, represented with the green line, is interpreted as stubborn agent SA (or opinion source) concentration. The average user defined popularity of the respective business over the same period of time represents the state of the social network. Also, the variation of the stars (blue) is represented with orange in the lower panel and it is interpreted as the participants opinion change ω. Point A depicts the SA concentration which triggers the delayed convergence in opinion (point B), and spike in opinion change (point C). In this example we have A(OX=28), B(OX=33), C(OX=32), $d_1 = 5$, $d_2 = 4$.

The averages of opinion change obtained for each considered dataset and for each phase are the following (their representation is given in Figure 6.3). Within square brackets are the minimum, maximum and standard deviation for each statistical average:

- **Twitter**: Initiation starts at OX=0 and ends and OX=33 [0, 39, 9.06], and has an average amplitude OY=21% [0%, 49%, 5.08]. Fusion happens at OX=42 and has an amplitude of 100% (i.e. it represents the maximum spike). Tolerance starts on average at OX=48 [43, end of time series, 4.07], and has an average amplitude OY=44% [13%, 83%, 4.01]. Intolerance starts on average at OX=68 [44, end of time series, 26.54], and has an average amplitude OY=5% [0%, 21%, 4.06].

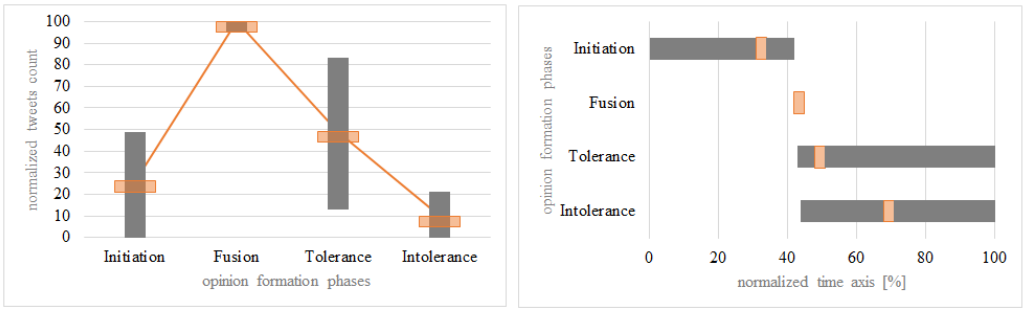
- **MemeTracker**: Initiation starts at OX=0 and ends and OX=37 [0, 40, 6.24], and has an average amplitude OY=13% [0%, 49%, 10.59]. Fusion happens at OX=42 and has an amplitude of 100% (i.e. it represents the maximum spike). Tolerance starts on average at OX=50 [43, end of time series, 4.88], and has an average amplitude OY=56% [45%, 97%, 3.90]. Intolerance starts on average at OX=62 [44, end of time series, 17.95], and has an average amplitude OY=5% [2%, 20%, 3.74].

- **Yelp** (all measurements are translated to the left on the time axis so that $t = 0$ coincides with the spike in SA, namely point A): Initiation starts at OX=0 and ends and OX=2 [0, 6, 2.1], and has an average amplitude OY=0.34 [0.1, 1.35, 0.14] stars. Fusion happens at OX=6 [3, 23] and has an amplitude of OY=2.25 [0.93, 4.9] stars. Tolerance starts on average at OX=33 [15, 73], and has an average amplitude OY=0.475 [0.275, 1.36] stars. Intolerance starts on average at OX=77 [47, end of time series], and has an average amplitude OY=0.175 [0.095, 0.46] stars.

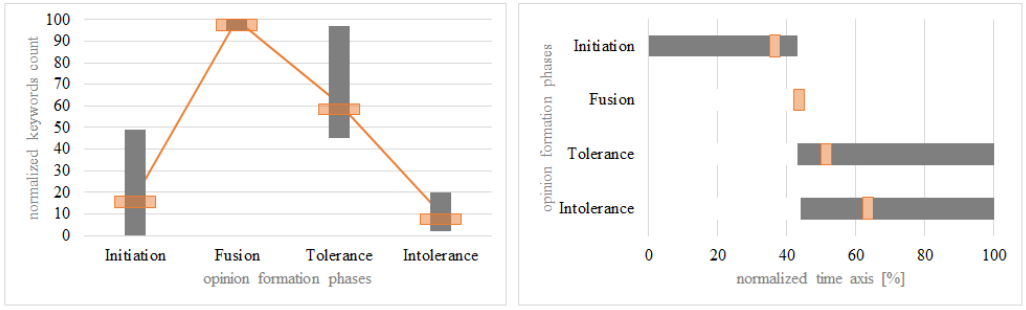
6.2.2. Phase transition

Apart from the quantitative measure of posts/time, I also consider the qualitative information from Yelp submitted by votes to local businesses (Figure 6.4a-c). With data from Yelp, I show the effects of a phase transition from social instability to social balancing which can occur when a critical concentration of opinion sources is reached in a social network. Figures 6.4-6.6 highlight the fact that the opinion (i.e. the stars given by users to a particular business) stabilizes only after reaching a critical ratio of opinion sources (i.e. votes representing strong opinions). This can be viewed in Figure 6.4 at time point $OX = 35$, in Figure 6.5 at time point $OX = 32$, and in Figure 6.6 at time point $OX = 28$, where the total number of reviews and votes rises dramatically (see the vertical red line). I interpret this phenomenon as a rise beyond a σ threshold for the concentration of opinion sources, which determines the *social balancing*, i.e. the average opinion stabilizes despite of opinion oscillations at local level. As such, in Figure 6.4, I can observe a stabilization of the average score given by users at time point $OX = 35$. The same type of stabilization occurs in Figure 6.5 at time point $OX = 32$. In Figure 6.6, I identify a stabilization point at $OX = 28$.

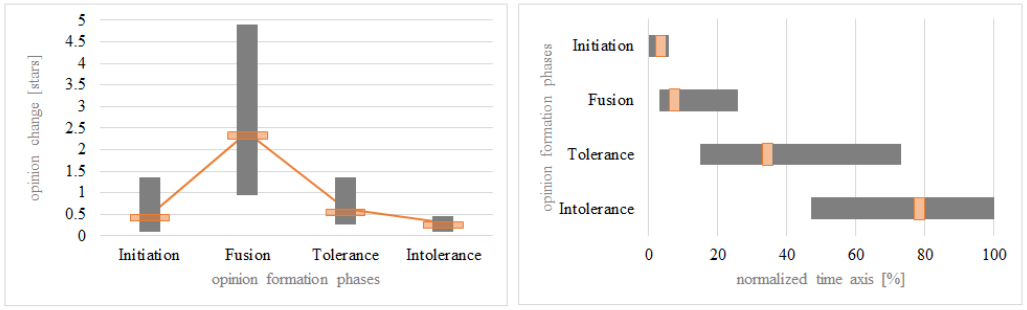
Corroborating all these empirical observations, I can state that Twitter and MemeTracker illustrate a *responsive* type of behavior, i.e. an immediate evolution towards the F phase, so a high opinion change is quickly reached for a relatively small ratio σ of opinion sources. This behavior, in turn, correlates well with another study which shows that Twitter online networks have a strong random and small-world component [78, 255].



(**a**) Twitter results: phases amplitude (left), phases time duration (right).



(**b**) Memetracker results: phases amplitude (left), phases time duration (right).



(**c**) Yelp results: phases amplitude (left), phases time duration (right)

Figure 6.3.: The four opinion formation phases represented in terms of: normalized amplitude (number of tweets / maximum number of tweets or opinion change in Yelp / maximum opinion change in stars), with each bar-plot depicting the minimum, maximum and average variation of opinion change; and time duration (on OX time-axis), with each horizontal bar depicting the minimum, maximum durations of the phase (gray), and the time at which it occurs on average (orange). All datasets indicate the same shape of opinion dynamics and the same succession of phases: I-initiation, F-fusion, T-tolerance and $\overline{T}$-intolerance..

Figure 6.4.: Evolution of reviews count and reviews votes for three popular businesses on Yelp over the period of 2010-2012. Accompanying each review trend, is the the average user defined popularity of the respective business over the same period of time. The critical opinion source concentration at OX=35 correlates with a stabilization of the state of the society given as the evolution of average stars awarded.

Figure 6.5.: Evolution of reviews count and reviews votes for three popular businesses on Yelp over the period of 2010-2012. Accompanying each review trend, is the the average user defined popularity of the respective business over the same period of time. The critical opinion source concentration at OX=32 correlates with a stabilization of the state of the society given as the evolution of average stars awarded.

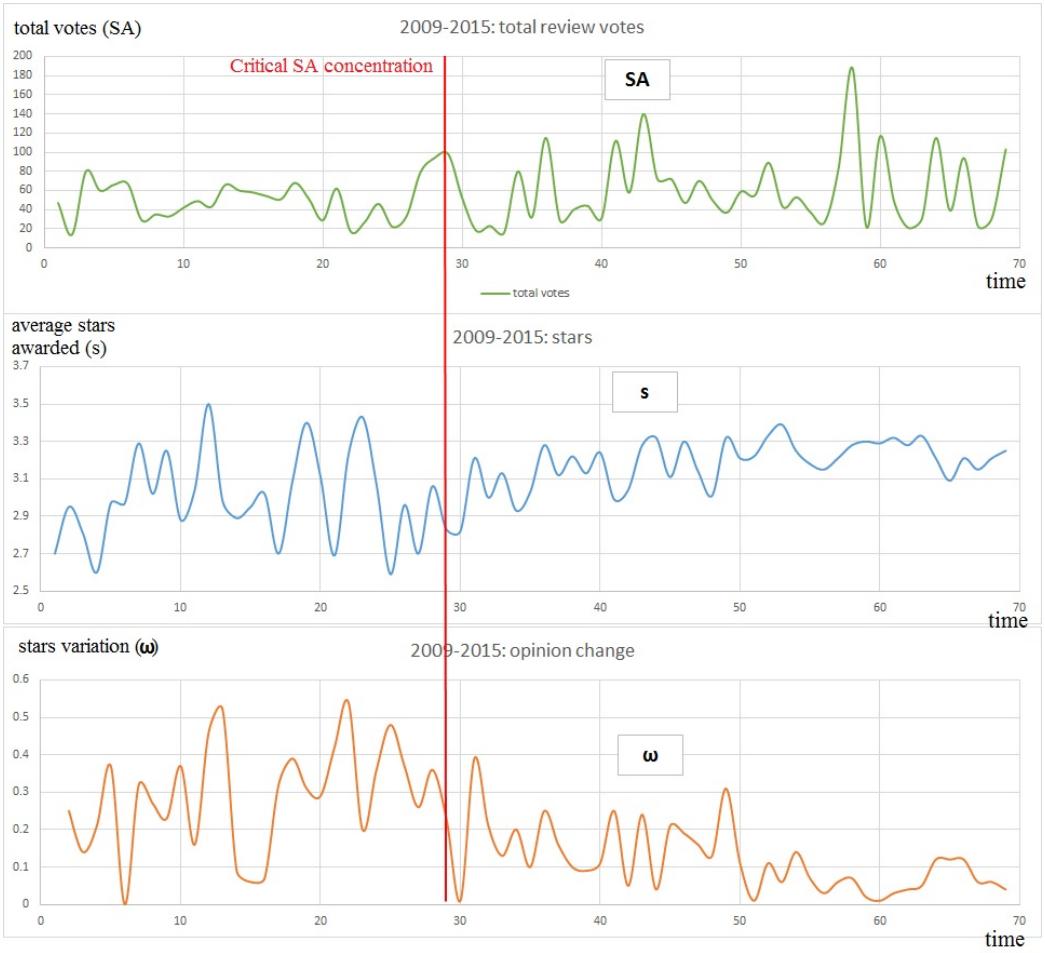


Figure 6.6.: Evolution of reviews count and reviews votes for three popular businesses on Yelp over the period of 2010-2012. Accompanying each review trend, is the the average user defined popularity of the respective business over the same period of time. The critical opinion source concentration at OX=28 correlates with a stabilization of the state of the society given as the evolution of average stars awarded.

In contrast, the Yelp dataset can be associated with a *saturated* type of behavior, as the ratio σ (relative to the maximum number of votes) needed to trigger the phase transition towards social balancing is high in all three cases. Balancing does not occur until a high concentration of opinion sources (I interpret them as similar to opinion-influencing "stubborn agents" [4] or "blocked nodes" [233]) are inserted into the social network.

6.2.3. New tolerance-based opinion model

This section analyzes the characteristics of a new opinion model that can reproduce the reported real-world phenomena, i.e. the four opinion formation phases and phase transition towards social balancing.

In terms of network *structure*, my analysis includes the basic topologies such as mesh, random [86], small-world [281], and scale-free networks [25]. Also, based on the last decade of research on the topic of realistic social network topology generation which either adds the small-world property to scale-free models [123, 96, 166], or adds a power-law degree distribution to the small-worlds [135, 57, 274, 296], I also consider the Watts-Strogatz with degree distribution (WSDD) [57].

In terms of *opinion dynamics* I rely on a predictive opinion interaction model that can be classified as graph and threshold based [112]. Generally, previous models use fixed thresholds [133, 37, 162, 74, 167] or thresholds extracted from real-world examples [99, 234]. However, there are a few models which use dynamic thresholds, but their evolution is not driven by the internal states of the social agents. On the other hand, my empirical references (i.e Twitter, MemeTracker and Yelp) indicate that opinion does not cease to oscillate and consensus is a rare case in real world. Therefore, I choose to extend Acemoğlu's opinion interaction model [2] based on stubborn agents, because it assumes that the society does not reach consensus. Based on recent research on stubborn agents which use a discrete [292] or continuous [4] representation of opinion, I integrate the following opinion models: one-to-one (simple contagion) versus one-to-many diffusion (complex contagion) [51], and discrete (0 or 1) versus continuous (0 to 1) opinion representation. By combining opinion representation and opinion diffusion, I obtain 4 distinct models; they are defined in Figure 6.7a and exemplified in Figures 6.7b and 6.7c. I build my tolerance-based opinion interaction model by using the SD (1) and SC (2) opinion representations as defined in Figure 6.7a.

Given a social network $G = \{V,\ E\}$ composed of agents $V = \{1,\ 2,\ ...,\ N\}$ and edges E, I define the neighborhood of agent $i\ \in\ V$ as $N_i = \{\ j\ |\ (i,\ j)\ \in\ E\}$. The disjoint sets of stubborn agents $V_0,\ V_1\ \in\ V$ never change their opinion, while all other (regular) agents $V\ \setminus \{V_0 \cup V_1\}$ update their opinion based on the opinion of one or all of their direct neighbors.

I use $x_i(t)$ to represent the real-time opinion of agent i at time t. Normal (regular) agents can start with a predefined random opinion value $x_i(0)\ \in\ [0, 1]$. The process of changing the opinion of regular agents is triggered according to a Poisson distribution and consists of either adopting the opinion of a randomly chosen direct neighbor, or an averaged opinion of all direct neighbors.

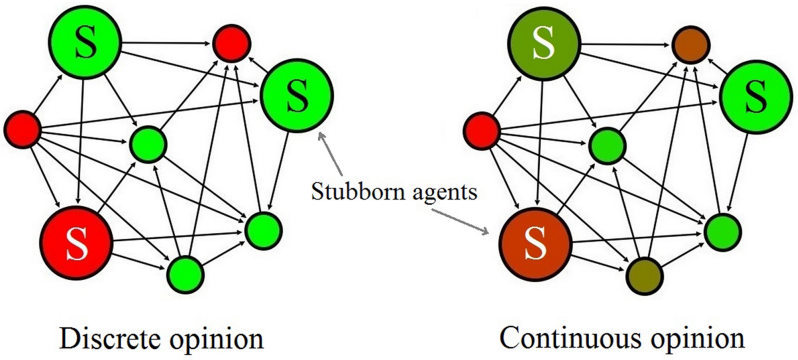
I represent with $s_i(t)$ the discrete opinion of an agent i at moment t having continuous opinion $x_i(t)$. In case of the discrete opinion representation SD (1) (Figure 6.7a), $x_i(t) = s_i(t)$; in case of the continuous opinion representation SC (2) (Figure 6.7a), $s_i(t)$ is given by equation 6.1.

$$s_i(t) = \begin{cases} 0 & if\ \ 0 \leq x_i(t) < 0.5 \\ 1 & if\ \ 0.5 \leq x_i(t) \leq 1 \end{cases} \tag{6.1}$$

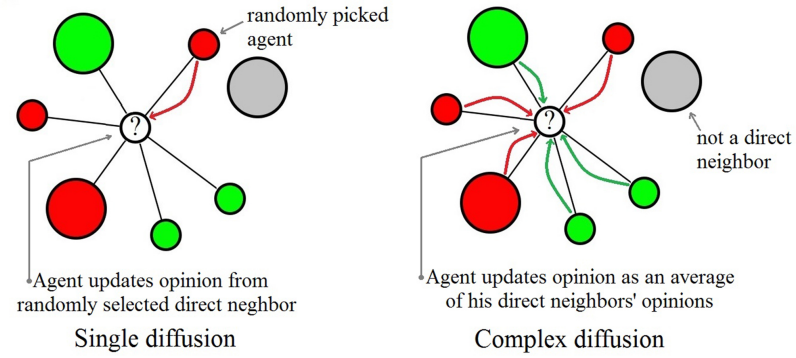
Table: interaction model taxonomy

		Opinion representation	
		Discrete	Continuous
Diffusion model	Simple	SD(1)	SC(2)
	Complex	CD(3)	CC(4)

(a)



(b)



(c)

Figure 6.7.: **a.** Interaction models taxonomy. **b.** Opinion representation types, where the larger nodes (labeled with S) represent stubborn agents. Discrete opinion (left): nodes have opinion 0 (red) or 1 (green) at any time (SD). Continuous opinion (right): nodes have any opinion between 0 and 1, highlighted by the color gradient transitioning from red to green (SC). **c.** Two opinion diffusion models for discrete representation: single diffusion (SD), respectivelty complex diffusion (CD).

Furthermore, $s(t)$ denotes the average state of the population at a certain time t by averaging the opinion of all individual agents $i \in V$.

$$s(t) = \frac{1}{|V|} \sum_{i \in V} s_i(t) \tag{6.2}$$

The previous social interaction models do not assign nodes (i.e. individuals or social agents) the basic properties of humans, i.e. humans evolve, learn, react, and adapt in time. The reason for the simplicity behind the existing models is twofold: first, the state-of-the-art models are only suited for theoretical contexts so bringing additional complexity to the interaction model would significantly increase the difficulty of mathematical analysis; second, involving measures of human personality (e.g. quantifying an individuals trust, credibility, or emotional state) is a complicated endeavor, in general; this was not the main goal of previous work.

Individual tolerance: interpretation and formalism

In order to improve the existing opinion interaction model based on a fixed threshold, I consider the evolution of personal traits by taking inspiration from social psychology. As a new contribution to the state-of-the-art, I introduce the concept of *tolerance* which reflects the individual's inner state and personal beliefs regarding surrounding opinions. For instance, egocentrism, as it is called in psychology, is highly correlated with individual's emotional state [84]. I choose to extend this model because the egocentrism-emotional state correlation is a trait that has been shown to influence and evolve with individual opinion [285].

Corroborating literature on attitude certainty [62], consensus [62], confirmation bias [208], social group influence [230], and ingroup emotion [192], I extrapolate the mechanism that leads to the formation of opinion into a measurable parameter. As such, I define *tolerance* θ as a parameter that reflects the willingness of an agent to accept new opinions. Similar to real life, individuals with higher tolerance will accept the opinion of others easier; thus, this parameter can be defined as a real number $0 \leq \theta \leq 1$. An agent with a tolerance value of 1 is called fully tolerant, whereas an agent with a tolerance of 0 is called fully intolerant (i.e. stubborn agent). Tolerance values which are greater than 0.5 describe a tolerance-inclined agent, while values smaller than 0.5 describe an intolerance-inclined agent.

Similar to the threshold-based continuous opinion fluctuation model described by Acemoglu et al. [4], tolerance can be used as a trust factor for an agent relationship; however, as opposed to the trust factor, tolerance changes its value over time:

$$x_i(t) = \begin{cases} 0 & if\ i \in V_0 \\ 1 & if\ i \in V_1 \qquad for\ t > 0 \\ \theta_i(t)\, x_j(t) + (1 - \theta_i(t))\, x_i(t-1) & if\ j \in N_i \end{cases} \tag{6.3}$$

where the new opinion $x_i(t)$ is a weighted sum of the agent's prior opinion $x_i(t-1)$ and the current opinion $x_j(t)$ of one randomly selected direct neighbor. The weights for the two opinions are given by the current tolerance $\theta_i(t)$ of the agent, thus, the extent of how much it can be influenced depends on its internal state.

As can be inferred from equation 6.3, the greater the tolerance of an agent, the easier it can accept

external opinion from others. At the beginning of the opinion formation process ($t = 0$), all agents are considered as having a high tolerance ($\theta_i(0) = 1$), but, as the society evolves, agents become intolerant, therefore segregated in clusters which tend to have a more stable opinion. I further define the tolerance θ of the entire population as a normalized average of all individual tolerances:

$$\theta(t) = \frac{1}{|V|} \sum_{i \in V} \theta_i(t) \tag{6.4}$$

I also introduce the concept of *opinion change* ω as the ratio of agents which have changed their current state (discrete time step t) since the last observation (time $t - 1$):

$$\omega(t) = \frac{1}{|V|} \sum_{i \in V} |s_i(t) - s_i(t-1)| \tag{6.5}$$

If an agent changes its state from one opinion to another, then the absolute difference$|s_i(t) - s_i(t-1)|$ will be 1; conversely, it will be 0 if the agent state does not change. This change, averaged over all agents at the interaction (discrete) moment t, defines the opinion change of the population $\omega(t)$. This metric is used to draw insights regarding the current tolerance level across the entire society.

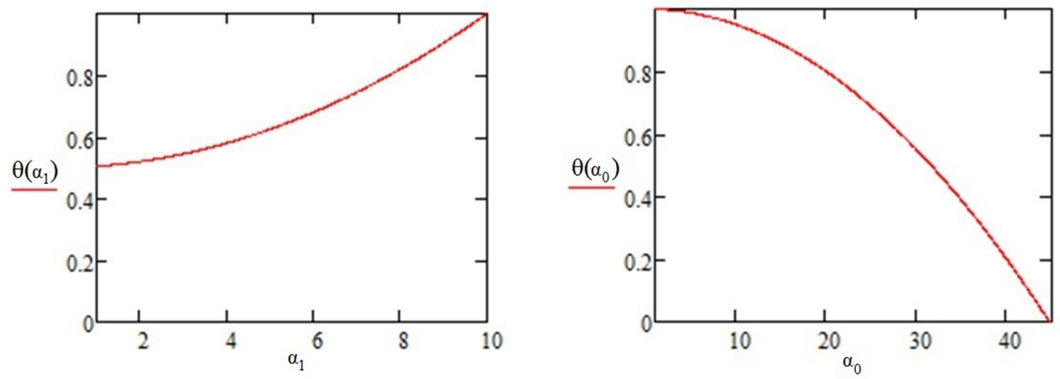
Progressive tolerance model

My model for tolerance evolution stems from the idea that the evolution towards both tolerance and intolerance varies exponentially [118, 282], e.g. a person under constant influence becomes convinced at an increased rate over time. If that person faces an opposing opinion, it will eventually start to progressively build confidence in that other opinion. Thus, my proposed progressive model represents the tolerance fluctuation as a non-linear function [118, 282], unlike other models in literature [4, 292]. For the first time, I integrate these socio-psychological characteristics in the dynamical opinion interaction model. As such, the new tolerance state is obtained as:

$$\theta_i(t) = \begin{cases} max\ (\theta_i(t-1) - \alpha_0\varepsilon_0,\ 0) & if\ \ s_i(t-1) = s_j(t) \\ min\ (\theta_i(t-1) + \alpha_1\varepsilon_1,\ 1) & otherwise \end{cases} \tag{6.6}$$

In equation 6.6, tolerance decreases by a factor of $\alpha_0\varepsilon_0$ if the state of the agent before interaction, $s_i(t-1)$, is the same as the state of the interacting neighbor (randomly chosen from all direct neighbors) $s_j(t)$. If the states are not identical, i.e. the agent comes in contact with an opposite opinion, then the tolerance will increase by a factor of $\alpha_1\varepsilon_1$. Moment t represents the time step where an opinion update is triggered; these moments are considered as being randomly distributed. The two scaling factors, α_0 and α_1, both initially set as 1, act as weights (i.e. counters) which are increased to account for every event in which the initiating agent keeps its old opinion (i.e. tolerance decreasing), or changes its old opinion (i.e. tolerance increasing). Therefore, I have:

$$\alpha_0 = \begin{cases} \alpha_0 + 1 & if\ \ s_i(t-1) = s_i(t) \\ 1 & otherwise \end{cases} \tag{6.7}$$



(a) (b)

Figure 6.8.: The tolerance function as defined by the progressive tolerance model. **a.** Tolerance scaling: shows how tolerance θ increases with the $\alpha_1\varepsilon_1$ scaling, as a result of continuous opinion change for an agent i. **b.** Intolerance scaling: shows how tolerance θ drops with the $\alpha_0\varepsilon_0$ scaling, from an initial tolerance $\theta_i(0) = 1$ to complete intolerance ($\theta_i(t) = 0$).

$$\alpha_1 = \begin{cases} 1 & if\ \ s_i(t-1) = s_i(t) \\ \alpha_1 + 1 & otherwise \end{cases} \tag{6.8}$$

Whenever an event occurs, the counter corresponding to the other type of event is reset. These factors are used to increase the magnitude of the two tolerance modification ratios ε_0 (intolerance modifier weight) and ε_1(tolerance modifier weight). The two ratios are chosen with the fixed values of ε_0 = 0.002 and ε_1 = 0.01. To determine these values, I have tried various $\varepsilon_0 : \varepsilon_1$ ratios as follows: if ε_0 is increased such that $\varepsilon_0 : \varepsilon_1 = 1 : 1$, most nodes will quickly become intolerant, as opinion will cease to diffuse; conversely, if ε_0 is decreased closer to a 1:10 ratio, then the society will become tolerance-inclined, with random opinion fluctuations. The used $\varepsilon_0 : \varepsilon_1$ ratio of 1:5 was chosen through consistent experimentation in order to provide a good balance between the deviations towards tolerance and intolerance, respectively.

As an illustration of the 1:5 ratio for $\varepsilon_0 : \varepsilon_1$, Figure 6.8 represents the non-linear tolerance function as implemented in equation 6.6. The displayed examples show that a total of 10 consecutive steps are required to maximize the tolerance if an agent starts with $\theta_i(0) = 0.5$, because the cumulative sum of $\theta_i(0) + \varepsilon_0 \sum_j \alpha_0$ reaches 1 after 10 iterations. Similarly, in Figure 6.8b, the sum $\theta_i(0) - \varepsilon_1 \sum_j \alpha_1$requires $t = 45$ iterations to reach intolerance ($\theta_i(t) = 0$), having started from $\theta_i(0) = 1$.

6.3. Model validation

My dynamical opinion model adds significant complexity to the opinion interaction model. Therefore, I use discrete event simulation (SocialSim [253]) over complex social network topologies, in order to validate my model's capability to reproduce real-world phenomena like the opinion formation phases and the phase transition towards social balancing.

6.3.1. Simulation on basic topologies

Regular networks

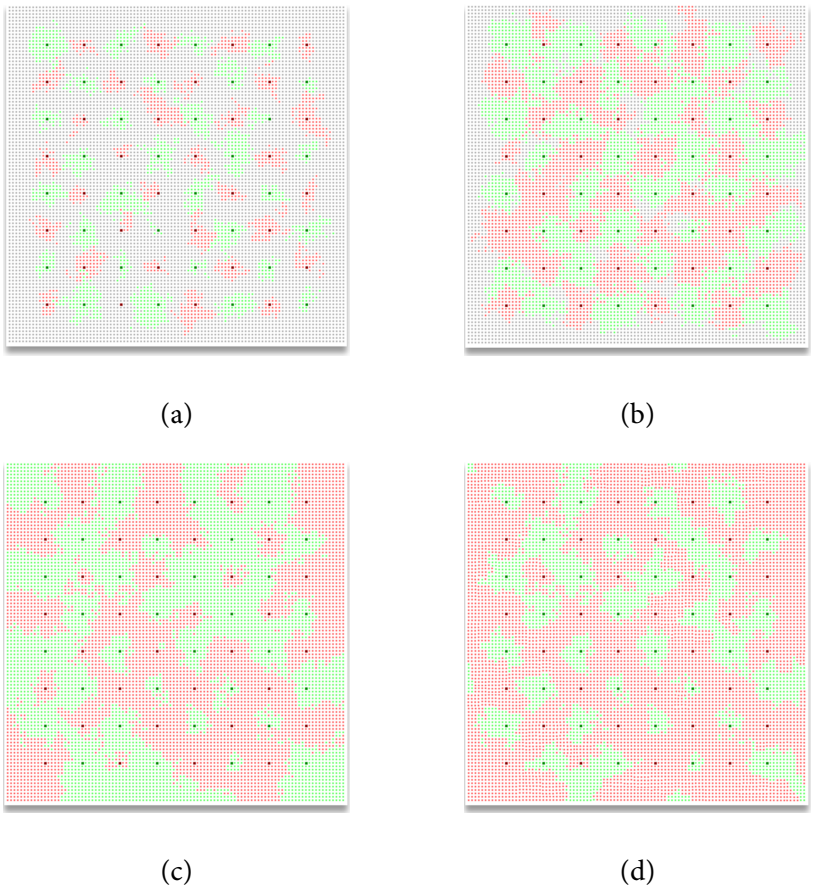
The first simulation setup is based on regular topologies, i.e. lattice and mesh. The results show that a homogeneous cluster of stubborn agents divides the overall society opinion (i.e. green (1) vs. red (0)) with a ratio that is directly proportional with their initial distribution. Figure 6.9 shows how a mesh network of 100,000 agents evolves under the influence of 64 stubborn agents – 32 of each opinion evenly distributed among the population. This way, I observe the same opinion formation phases as identified by my empirical observations: initiation I (Figure 6.9a), fusion F (Figure 6.9b), tolerance T (Figure 6.9c), and intolerance $\overline{T}$ (Figure 6.9d). The situation in Figure 6.9c may lead to one of two scenarios: a perpetual (proportional) balance of the two opinions, introduced by us as *social balancing* (the society remains in the T phase, and $\overline{T}$ is never reached), or a constant decrease in opinion dynamics which ultimately leads to a stop in opinion change (the society reaches the $\overline{T}$ phase), as depicted in Figure 6.9d.

Figure 6.10a illustrates a society which tends towards the tolerance phase T and social balance, by providing the evolution of the overall society state $s(t)$ (as defined in equation 6.2), tolerance $\theta(t)$ (see equation 6.4), and opinion change $\omega(t)$ (equation 6.5). For the society described in Figure 6.10a, the initiation phase I is revealed by the early increase of $\omega(t)$, as the number of individuals with opinion increases. The climax of $\omega(t)$ represents the fusion phase F. At this stage, there is a maximum number of bordering agents with distinct opinions (a situation that is also depicted in Figure 6.9b) and $s(t)$ evens out. In the tolerance phase T, the agents tend to stabilize their opinion, i.e. $\theta(t)$ stabilizes and $s(t)$ converges towards the ratio of stubborn agents (which was chosen as 1:1).

Another observation is that opinion fluctuation is determined by the stubborn agents density (see Figures 6.10b, c and d). Because of the regular topology, the fewer stubborn agents (regardless of their opinions) there exist in the society, the more the opinion fluctuates. This is explained by the fact that having few stubborn agents means few points of opinion control and stabilization in the local mesh structure; conversely, many stubborn agents make possible the control of more regular agents. Because of this, $s(t)$ may drastically get biased in someone's favor until the entire society stabilizes (Figure 6.10b). Also, due to the small influencing power of a few agents, the opinion will not necessarily stabilize with the same distribution ratio. As expected, the opinion distribution of a society with a high opinion source concentration will tend towards the ratio of the two stubborn agent populations (Figure 6.10c).

If the ratio of the two stubborn agent populations is not 1:1, then the opinion fluctuation will be around that ratio only during the initiation phase I. Afterwards, the overall opinion will get more biased towards the opinion of the larger stubborn agent population. In Figure 6.10d the ratio is 1:4 between green and red stubborn agents, therefore the fluctuation starts around 20% green opinions, but eventually stabilizes at 8%.

The scenarios presented above hold true for lattices. Consequently, these conclusions are more of theoretical interest, as real social networks are typically not organized as such regular topologies. Next, I consider more realistic network topologies.



(a) (b)

(c) (d)

Figure 6.9.: Green (1) vs. red (0) opinion evolution with homogeneous stubborn agent distribution in a 100,000 node social network. The network is initialized with 32 red and 32 green stubborn agents. Initially, the regular agents have no opinion and are colored with grey. I distinguish between the following phases of opinion formation: **a.** The initiation phase I where the society has no opinion, i.e. the stubborn agents exercise their influence to the surrounding neighborhood without being affected by any other opinion. **b.** The fusion phase F where the society is now mostly polarized (green or red) and different opinion clusters expand and collapse throughout the society. **c.** Tolerance phase T, where the cluster interaction stabilizes and new, larger, more stable clusters emerge. **d.** Intolerance phase $\overline{T}$, where the overall tolerance of agents has decreased to a point where opinion fluctuation ceases and the red opinion becomes dominant ($\theta(t) < 0.1$).

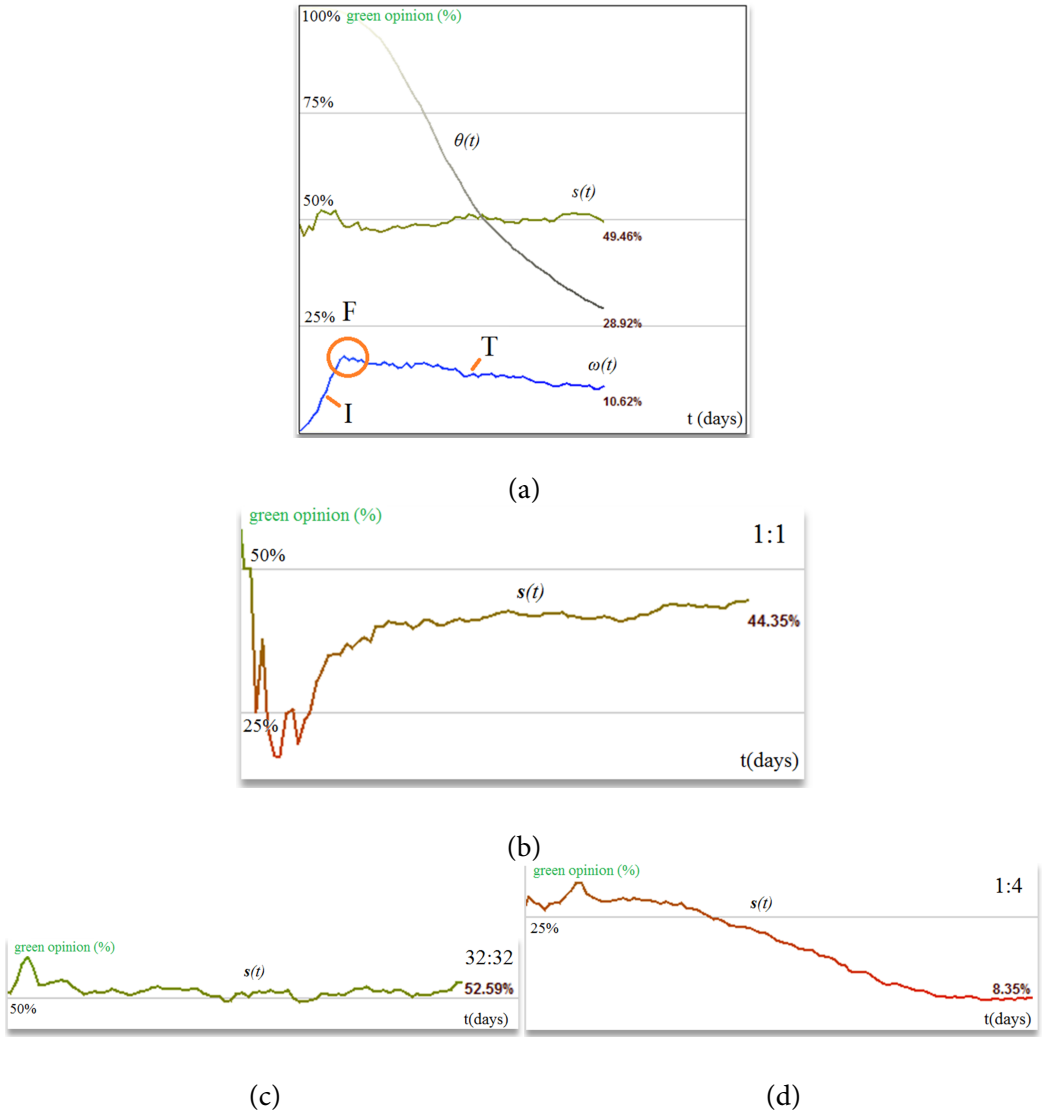


Figure 6.10.: Simulation of a 100,000 mesh network with SocialSim [253], displaying a representative example for the evolution of $s(t)$, $\theta(t)$, and $\omega(t)$, as well as the opinion evolution $s(t)$ with various stubborn agents distributions. **a.** Mesh topology, where the lowest panel displays the opinion change (ω) evolution over three simulation phases. **b.** Opinion evolution $s(t)$ with few and evenly distributed SA (1:1 ratio: 1 green, 1 red). **c.** Opinion evolution with many and evenly distributed stubborn agents (1:1 ratio: 32 green, 32 red), **d.** Opinion evolution with few and unevenly distributed stubborn agents (1:4 ratio: 1 green, 4 red).

Small-world networks

By constructing a Watts-Strogatz small-world network of 100,000 nodes, [281, 246, 276, 263, 57, 22] I show experimentally that a different type of behavior can emerge. For instance, Figures 6.11a and b present the society as having a mixed opinions distribution with no noticeable clusters. As opposed to the representation in Figure 6.9, this topology does not allow multiple agents to cluster around the stubborn agents and converge towards their opinion. Consequently, this model not only increases the dynamics of opinion fluctuation, but also keeps the society in *social balance*. The fourth and final phase of opinion evolution - the intolerance phase - does not occur, and opinion change $\omega(t)$ is maintained at a (high) constant level. Moreover, the state of the society $s(t)$ is stable.

The society depicted in Figure 6.11a is homogeneously mixed from an opinion standpoint. Clusters do not form because many agents have long range links to other distant agents whose opinion can be different from the local one. This leads to a perpetual fluctuation which remains in balance. The noticeable effect on a small-world network is that the opinion stabilizes very fast and always at the ratio of the two stubborn agent populations (i.e. 1:1 in my case). In a mesh network, having few stubborn agents leads to an imbalance of opinion, but in the case of small-world topologies, opinion across the entire population always stabilizes. Opinion change $\omega(t)$ is also much higher compared to the mesh (i.e. 42% versus 10% under the same conditions) due to the long range links.

Scale-free networks

I apply the same methodology by constructing a 100,000 node Barabasi-Albert scale-free network and highlight the unique behavior it enacts.[25, 217, 10, 276, 242, 57] As Figure 6.11c shows, the society does not reach a balance at the expected value ($32 : 32 \Rightarrow 50\%$); instead, it gets biased towards one opinion or another. The reason behind this behavior is related to the power-law degree distribution [276]. As such, scale-free networks behave more like a tree-structure with hubs rather than as a uniform graph. Indeed, as opinion flows from one agent to another, the higher impact of the hub nodes on the opinion formation at the society level becomes clear. If, for example, a green stubborn agent is placed as the root of a sub-tree filled with red stubborn agents, that sub-tree will never propagate red opinion as it cannot pass through the root and connect with other nodes. Experimentally, this is illustrated in Figure 6.11c. The green agents have been placed over nodes with higher degrees, and this can be seen in the evolution of the opinion. There is some initial fluctuation in the society and although the stubborn agent distribution is even, the fluctuation rapidly imbalances as the overall tolerance $\theta(t)$ plummets and all agents become sort of "indoctrinated" by the green opinion. The rapid drop in tolerance coincides with the drop in opinion change $\omega(t)$ and the stabilization of the state $s(t)$ at over 90%. Simulations were also run on the WSDD topology [57], which has a strong scale-free component, and yield similar results which lead to the same set of observations.

Additionally, I have implemented the algorithm for generating uncorrelated scale-free networks as defined by Catanzaro et al. [50]. Because of the random nature of this topology, the results obtained with SocialSim are much closer to what I obtain for random networks. Figure 6.11c represents the opinion formation phases obtained on correlated (assortative) Barabssi-Albert networks. Due to the the assortative connectivity, most nodes are influenced by the same hubs, with same opinion, thus tolerance decreases rapidly. This leads to an evolution towards the intolerance phase and a halt in opinion dynamics. The uncorrelated scale-free topology is worth mentioning in this thesis, since,

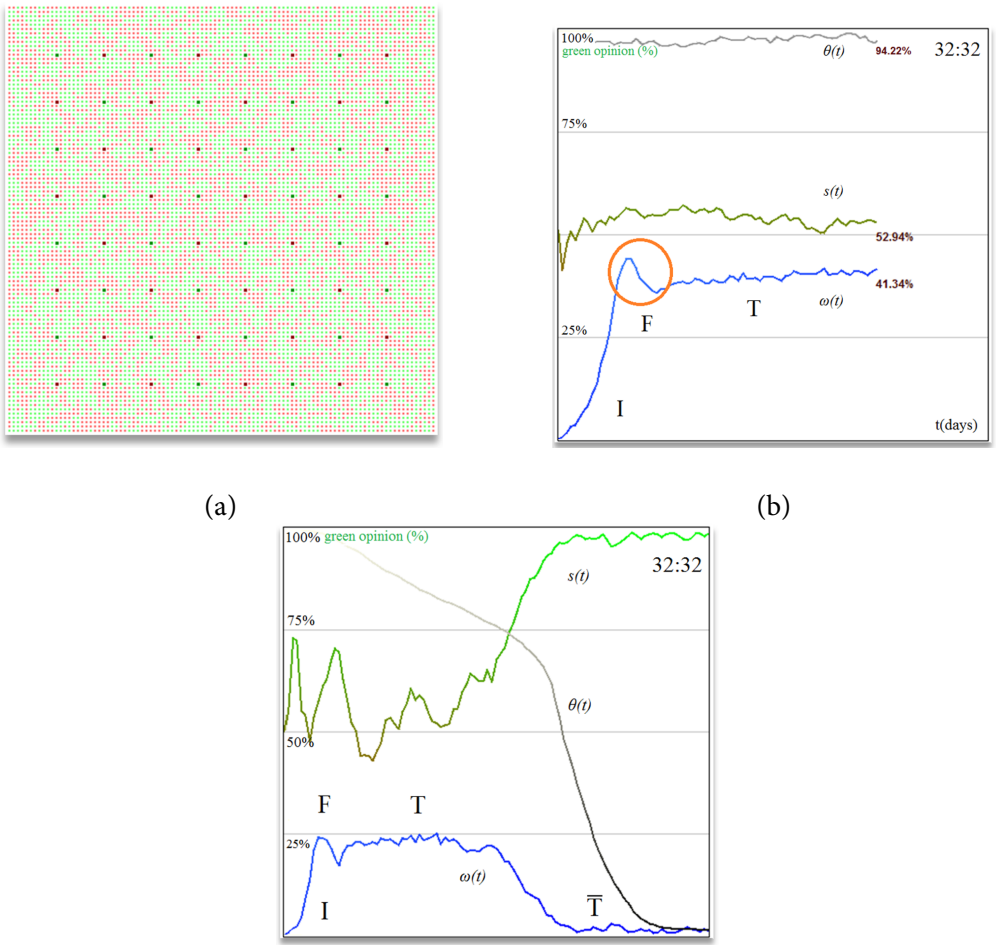
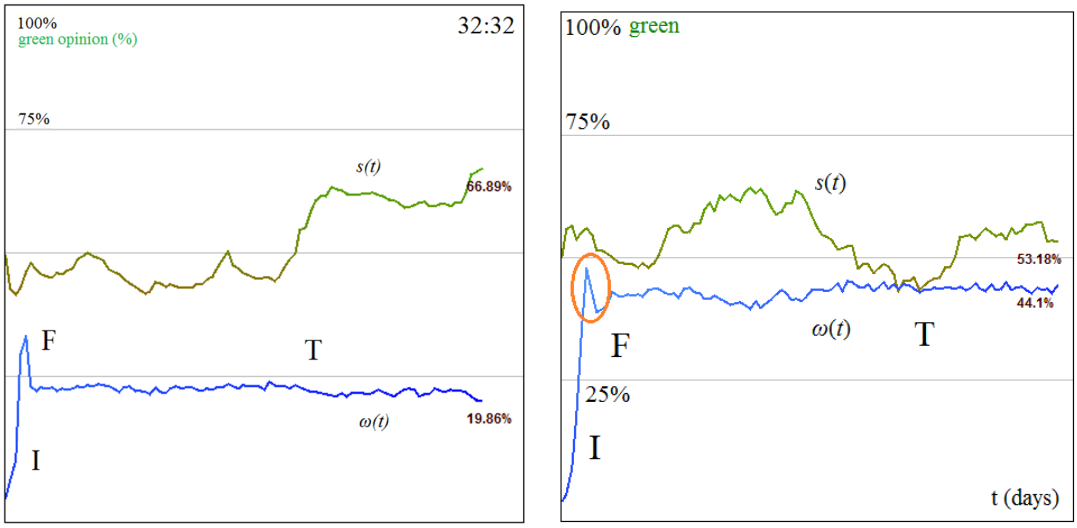


Figure 6.11.: Opinion evolution with homogeneous stubborn agent distribution (32:32) in small-world and scale-free networks. **a.** Tolerance phase where no visible clusters emerge for small-world networks. **b.** For small-world networks, social balancing is attained because tolerance remains extremely high ($\theta(t) > 90\%$), opinion change (ω) exhibits the three opinion evolution phases (initiation I, fusion F, and tolerance T), and never reaches intolerance. The state of the society $s(t)$ is stable. **c.** Social balancing is not achieved for scale-free networks: tolerance drops constantly and the society reaches the intolerance phase ($\overline{T}$). The state of the society $s(t)$ is unstable during the first three phases of opinion change, then stabilizes as tolerance (θ) and opinion change (ω) fall.



(a) (b)

Figure 6.12.: **a**. Representative simulation depicting opinion evolution in an uncorrelated random scale-free network with 32 red stubborn agents and 32 green stubborn agents: although opinion constantly oscillates, society becomes balanced and stabilizes in the tolerance phase. **b**. Representative simulation depicting opinion evolution in a random Erdos-Renyi network with 32 red stubborn agents and 32 green stubborn agents. Opinion change is maintained high and opinion presents high oscillations, but the overall state of the society becomes stable and predictable around 50% green opinion.

from what I obtain in Figures 6.12a,b, this topology behaves much like the random topology. The explanation is due to the the fact that nodes may be connected to any random hubs, so neighboring nodes will not adhere to the same community influenced by the exact same hubs. This diversity in connexions keeps tolerance high, so that opinion is kept in balance.

6.3.2. Phase transition in opinion dynamics

This section aims at analyzing the impact of topology, network size, interaction model, stubborn agent placement, ratio and concentration on the opinion change (ω), and on convergence towards intolerance (θ).

Simulations show that, in a society with a fixed stubborn agent distribution, the topology τ determines if:

- the society enters the intolerance phase I: $\theta \to 0$ (with $\theta < 0.1$), which also results in $\omega \to 0$;
- the society *balances* and never enters the intolerance phase I: $\theta \to \theta_{limit}$, where $\theta_{limit} > 0.1$ and maintains a high ω;
- the society continues to oscillate for $0.1 < \theta < 1$, but the tolerance level does not stabilize.

In case of the Yelp dataset, I notice that for a given topology τ, and a network of size N, when the concentration of stubborn agents is bigger than a critical ratio σ, the society never becomes intolerant. In such cases, the society becomes balanced, with slight oscillation in tolerance or opinion change. The goal is therefore to find the tuples (τ, N, σ) at which this phenomenon occurs.

To obtain my results I have used five topologies τ (mesh, random, small-world, scale-free and WSDD), network sizes N of 400 up to 100,000 nodes, my new tolerance interaction model, a ratio of 1:1 between green (1) and red (0) stubborn agents, and an increasing concentration of stubborn agents ranging from 1% to 36%.

Impact of topology

The tolerance and opinion change with respect to the number of stubborn agents, as depicted in Figurs 6.13a and b, highlight a clear difference between the five topologies, namely mesh, random, small-world, scale-free, and WSDD. There is a total of three clearly distinguishable behaviors: a *responsive* behavior (present in small-worlds and random graphs), a *linear* behavior (for mesh networks), and a *saturated* behavior (corresponding to scale-free and WSDD networks).

The tolerance increases *linearly* for the mesh, as the population of stubborn agents increases. Consequently, there is no critical σ for which a phase transition occurs due to the high regularity of the network, but there is a visible saturation point (when the blue graph begins to drop in Figure 6.13a). This happens because the society is physically filled with more stubborn agents than regular ones and because all stubborn agents have $\theta = 0$, the overall tolerance begins to drop.

The *responsive* behavior exhibited by the random network and small-world networks suggests that these two topologies behave similarly in the context of opinion source saturation. The two topologies are almost identical under the conditions defined here, as they behave almost as the opposite of mesh networks: once the critical point σ is reached, their tolerance rises to the maximum value. Then, as the stubborn agents population increases, the tolerance and opinion change values decrease proportionally. The random and small-world topologies are equivalent with the mesh topology as the society becomes saturated with stubborn agents (i.e. see Figures 6.13a and b in terms of tolerance θ and opinion change ω, respectively).

Finally, the *saturated* behavior groups together the scale-free and WSDD topologies, both of which have the feature of degree-driven preferential attachment. The two topologies show smaller responsiveness to social balancing. As depicted in Figures 6.13a and b, the critical point of stubborn agents concentration for scale-free is by far the greatest one (i.e. $\sigma = 16\%$) and the maximum tolerance θ reached is the smallest among the simulations aiming at the impact of topology (20%). The WSDD topology shows a better response, at a much lower critical stubborn agents concentration point ($\sigma = 4\%$) and reaches social balance at $\theta = 30\%$.

Impact of network size

When analyzing the opinion change at society level, the same observations and classification are valid for all other network sizes. The larger the size N is, the more accurate the delimitation shown in Figures 6.13a and b becomes.

The impact of size offers a comparison of different tolerance stabilization on the same topology. The results in Figures 6.13c, d, e, and f show how well the social balancing effect scales with increasing sizes of the network.

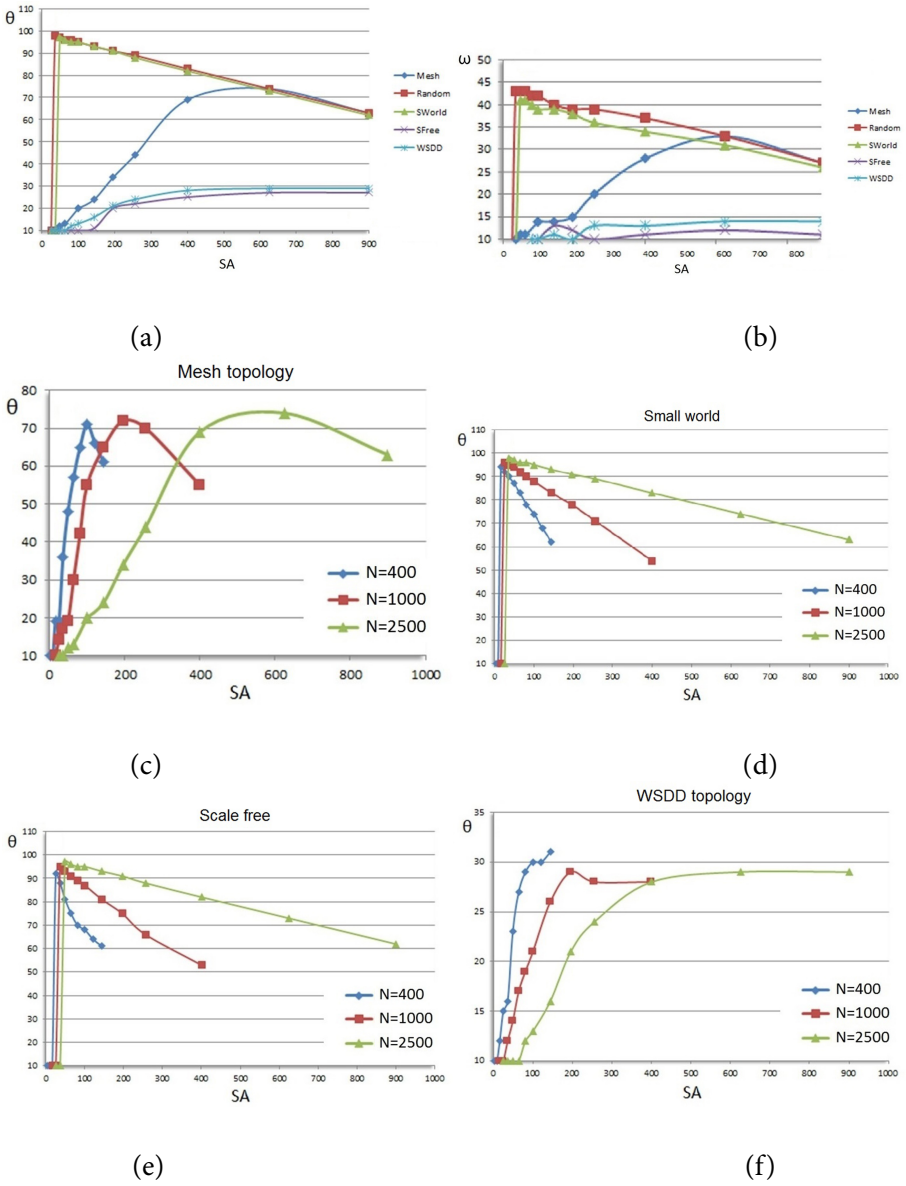


Figure 6.13.: Tolerance (θ) and opinion change (ω) evolution with the increasing concentration of evenly distributed SA and increasing network sizes. **a, b.** θ and ω over the five topologies when the size of the network is fixed at $N = 2500$, and the concentration of stubborn agents ranges from 4% to 36%. **c**, **d**, **e**, **f.** Tolerance θ stabilization values at which social balancing occurs over increasing network sizes (N=400 to 2500 nodes).

The behavior of meshes, presented in Figure 6.13c, shows a linearly proportional increase of the critical stubborn agents concentration σ (around 20-25%) in accordance with the network size N. A similar evolution is visible in Figure 6.13f, on networks with preferential attachment, where the required σ is also proportionally bigger on larger networks. In Figures 6.13d and 6.13e, the random and small-world networks exhibit similar behavioral patterns: they achieve the critical point σ with maximal opinion change, and then evolve towards intolerance at a pace that is corroborated with N (i.e. a slower drop in tolerance for larger networks occurs).

All simulations presented in this section confirm my main observations (Twitter, MemeTracker, Yelp) on opinion formation phases and phase transition towards social balancing. Figure 6.13 contains averages stemming from multiple experiments run in SocialSim, then processed separately in Microsoft Excel. The points on the OX axis are fixed SA concentrations which are used throughout these experiments, and the values on the OY axis are averages obtained from multiple runs (i.e. 10). An individual graph from one sub-figure is based on 8 (different SA concentrations) x 10 experiments = 80 simulations. One subfigure is the result of 3x80 = 240 simulations, therefore Figure 6.13 is based on 4x240 = 960 simulations.

6.3.3. Validation hypotheses

Extracting social state and opinion change from empirical data

It can be debated whether hashtag dynamics could be equated to opinions dynamics. People use them driven by different forces, and many are not associated to changes of opinions or opinions formation. Indeed, in the Twitter and MemeTracker the actual state of the society s cannot be deduced, as the tweets are not processed (e.g. using sentiment analysis), so I interpret the number of replies as a measure of opinion change: more replies means more opinion injected in the society. When previously unopinated people reply or retweet, they do so because they have reached a clear opinion on a particular subject and they feel the need to express it by broadcasting the related story. The number of replies (OY axis) at a given moment (OX axis) corresponds to users expressing (injecting) opinion in the society. As such, the Twitter and MemeTracker datasets are supporting the observations related to the opinion change ω and not to those related to the opinion state s.

In case of the Yelp dataset, the state s is the current average number of stars awarded by users; it represent the opinion towards a business. The variation of s between two time moments $t-1$ and t is the opinion change ω. This information can be found in Figures 6.4-6.6 for the experiments' opinion change representation. As Yelp users are considered nodes in a social network, the opinion dynamics on local businesses will be affected by the underlying social network links. The social agents that influence opinion are linked and interact at least through the Yelp platform. Nonetheless, as Yelp is well-known for hosting social events for reviewers, I believe that these social ties are even stronger.

For both of the Twitter and Yelp results I have taken into consideration the entire datasets. The Twitter data consists of 6 million hashtags out of which the top 1000 were processed. Similarly, the Yelp dataset consists of 1.5M reviews for 60K businesses of which I processed the top 1000. The same results (phases and phase transition) can be observed on all the data that was processed by the authors. Of course, Figures 6.1, 6.4-6.6 are limited to just 12 hashtags, and 3 businesses respectively, due to space and readability constraints. However, the same results are observed over all empirical data available in Yelp. The vertical lines in Figure 6.4 correspond to the critical σ threshold (stubborn agents SA concentration) being reached in the simulation, i.e. we reach social balance. Balance is

achieved in the empirical data when the state of the society (i.e. average stars awarded) presents an overall variation of less than 1 star out of 5 ($^1/_5$= 20% variation) between any maximum and minimum values in time. This balance is correlated with a spike in the number of opinion injected into the society. This spike is the required σ threshold of stubborn agents to reach balance. Therefore, the formal description of the algorithm that finds vertical lines in Figures 6.4-6.6 is:

if spike occurs in number of opinion sources **and** variation in the state of the society s is < 1 star, **then** the corresponding moment is marked with a vertical line.

Using the simple contagion principle

Since there are interaction models based on simple as well as complex contagion, I have implemented a complex contagion model in SocialSim and performed extensive simulations to compare with my obtained simple contagion results. I find that, when using complex contagion, the dynamics of the society is accelerated and the I, F phases occur very fast, the T phase is omitted, and the society enters the intolerance phase. This is due to the fact that averaging the opinion of neighbors does not allow a node to be in contact with the likely divergent opinions of his neighbors, one by one, and thus tolerance cannot increase; as a result, it will decrease after each interaction. Conceptually, I have defined the tolerance model to keep nodes tolerant through individual interactions which present diversity in opinion, like we would have in real life. Even if humans usually evolve towards the average opinion of their social group, they do so through individual interactions.

Using a more formal explanation, for tolerance to remain high and opinion to exhibit dynamics, the expected difference between a node's state and a neighbor's state should be large throughout the simulation (e.g. over time, a node may come in contact with opinions sequences like 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1). Such an interaction pattern will maintain a high tolerance. In the case of complex contagion, the average state of the neighbors will always be similar (e.g. a node will come in contact with almost the same opinion like $s(t)$, $s(t+1)$,... $s(t+k)$ with $s(t) \simeq s(t+1) \simeq ... \simeq s(t+k)$ and this lowers tolerance quickly). Figure 6.14 presents a simulation result that illustrates the above considerations.

Comparing against a null-model

A comparison with a null/random model can help to really understand the effect of the so-called social balancing, and such a study will contribute to increasing the robustness of my model. I have addressed this problem through implementation of the random interacting agents in my simulation tool, SocialSim; then Ie replicated experiments; and finally drew a short conclusion. Randomness has been added in two ways:

- **Fully-random** interaction model: all agents have random tolerance values, random initial opinion, interact with random neighbors, who posses random opinion, and tolerance is updated randomly after each interaction. When comparing the simulation results with a random behavior interaction model I obtain the same output regardless of topology and SA (stubborn agent) concentration. Figure 6.15a-b depicts a small-world with 50% of the population taking part in the opinion process. The state of the society remains balanced at 50% and there are no visible opinion formation phases.

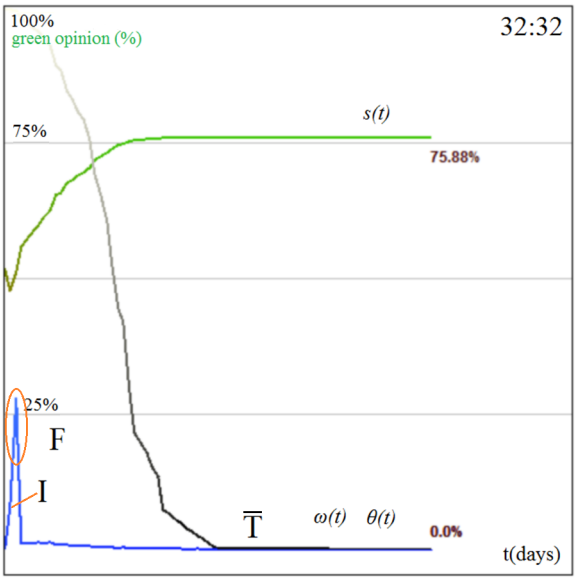
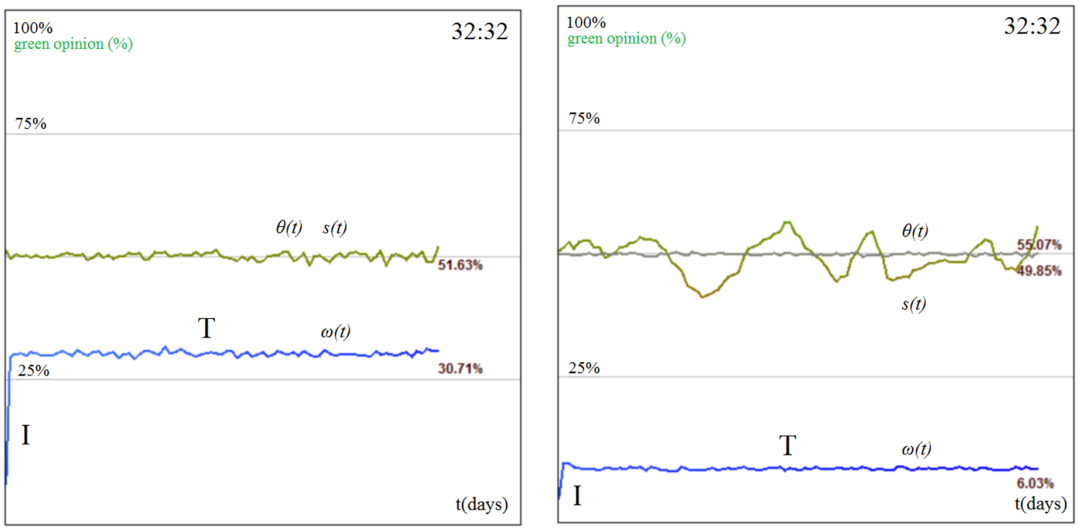


Figure 6.14.: Simulation results of the tolerance model tested using the complex contagion interaction principle. I use a 10,000 small-world network with a balanced number of stubborn agents (32 green : 32 red). The state s stabilizes quickly, and opinion change ω and tolerance θ converge towards zero. Consistently throughout simulations, the opinion formation phases are short in duration (generating a distinctive spike, as indicated with the orange oval in the figure) and the society always tends towards intolerance.



(a) (b)

Figure 6.15.: **a**. Representative simulation depicting opinion evolution in an uncorrelated random scale-free network with 32 red stubborn agents and 32 green stubborn agents: although opinion constantly oscillates, society becomes balanced and stabilizes in the tolerance phase. **b**. Representative simulation depicting opinion evolution in a random Erdos-Renyi network with 32 red stubborn agents and 32 green stubborn agents. Opinion change is maintained high and opinion presents high oscillations, but the overall state of the society becomes stable and predictable around 50% green opinion.

- **Random-tolerance** interaction model: it is similar with my proposed opinion interaction model but here each agent receives a static tolerance initialized with a random value in the [0,1] interval at startup. The same results are obtained as in the fully-random interaction scenario, with the state of the society stabilizing around the expected value (50%, see Figure 6.15b). With the random-tolerance interaction model, the state of the society oscillates much more in comparison with the fully random model, but less when compared with my proposed tolerance interaction model. As for the fully-random model, the opinion formation phases are not clear.

I can only conclude that using a random/null model for validation shows that tolerance actually plays an important role in the statistical results obtained in my proposed model. Tens of additional simulations have been run and I obtain the same results results as in Figure 6.15a for the fully random model, and as in Figure 6.15b for the random tolerance model.

Considering non-particpant agents

One of the underlying assumptions on society composition is that my experiments focus on the evolution of the overall society state, that can reach a stationary tolerance or intolerance phase. Such experimental effort is motivated by what has been observed during empirical analysis on real social

media data. However, if we focus on an overall society (with given topologies, sizes, clusters etc.), I need to represent nodes or agents that remain neutral during a given discussion: the opinion dynamics plotted in the empirical data in Figure 6.1 represent only a fraction of the entire population that are taking a partisan role over a given subject. The model assumes that every node or agent will take a part in a given discussion, with a tunable tolerance parameter that will influence the opinion itself. Nevertheless, this result is hard be compare with empirical observations, because a large part of the population will not explicitly state its opinion, and I cannot know if these agents, after being exposed to the discussion from a (set of) source(s), will remain neutral (e.g., not interested) or if they will just privately form their own opinion without communicating that to the neighbors. So I set to answer the question *Which is the role of such a neutral agents?*

As a consequence, I have added another type of agent in the simulations: *NullAgents*. These are agents which do not hold any opinion, namely their opinion is $x(t) = 0.5$ and thus, their state is $NONE$ (therefore, I have used a third state, along with $YES(1)$ and $NO(0)$). Theoretically, I consider that NullAgents (NAs) should act like edge-disconnections in the graph. They are distributed at random, with a given concentration C. Then, I verify if there is a threshold C^* for which if $C < C^*$ the society behaves the same as before (i.e. graph is mainly connected, and these agents have limited impact), and if $C > C^*$ the society behaves differently (i.e. the underlying graph becomes mainly disconnected).

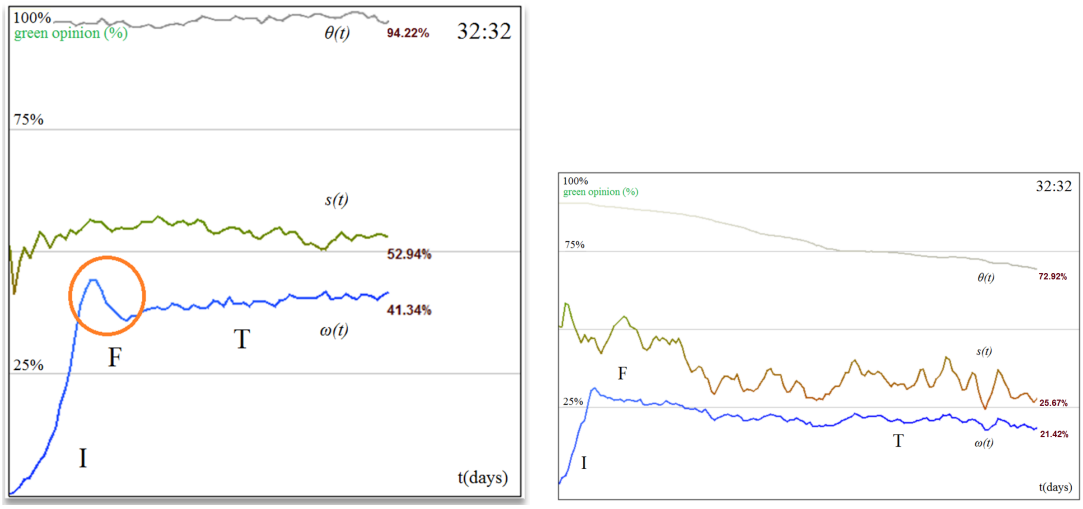
By adding NAs in SocialSim I was able to test them with all the synthetic topologies. The higher the population of randomly distributed NAs, the fuzzier the four phases become. Initiation (I) is less steeper, fusion (F) isn't that *spiky* anymore, tolerance T is achieved harder/later as the state oscillates more, but the society is still in balance and predictable; opinion change stabilizes with some delay. The phases tend to dissolve after a concentration of $C \sim 30\%$ population of NullAgents.

Additional simulations have been run with NAs and I obtained results similar to those presented in Figures 16a-f. All tests from Figure 16 were run on small-worlds with 10,000 nodes.

Proving the model holds

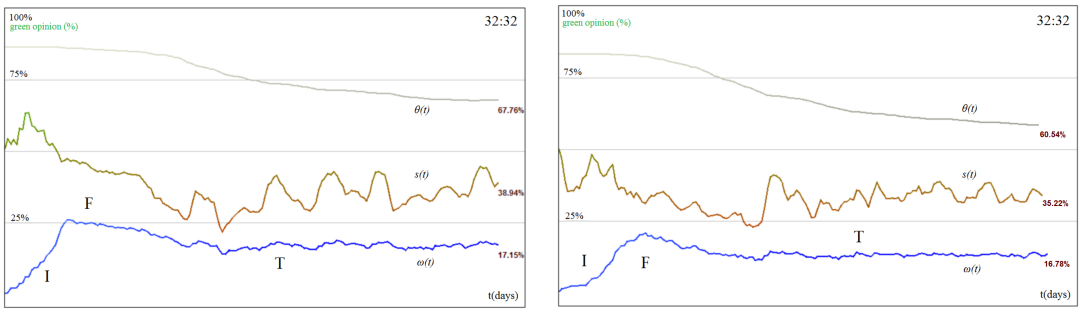
The validation and simulation conclusoons of this model are based on empirical observations, and this might seem a weak point of this contribution. A comparison of the results with an analytical solution of the model at infinite time could help to understand things much better. In fact, the model includes many complex parameters that can be tuned, and an analytical resolution can help understand the general bias of the model itself, with a mathematical proof of its assumed stationarity after a given point.

To discuss this aparent problem I come with the following discussion. My original interaction model is non-trivial, i.e. the evolution of nodes opinion and tolerance depend on dynamic parameters and it is inspired from the previously introduced model of Acemoglu et al. [4, 2]. The difference consists of the fact that the so-called trust factor from [4, 2] is no longer fixed, as my model describes the evolution of this parameter, which I identify as corresponding to an individual's tolerance towards other opinion according to basic human traits. In reference [4], the authors try solve the equations that describe the stationarity of opinion evolution by using random walks from regular agents to stubborn agents which influence their state. Even if their model is simpler than the model proposed in my thesis, they have to come up with some simplifying assumptions in terms of network topology (only regular topologies are tractable) and number of agents (they solve equations on small networks



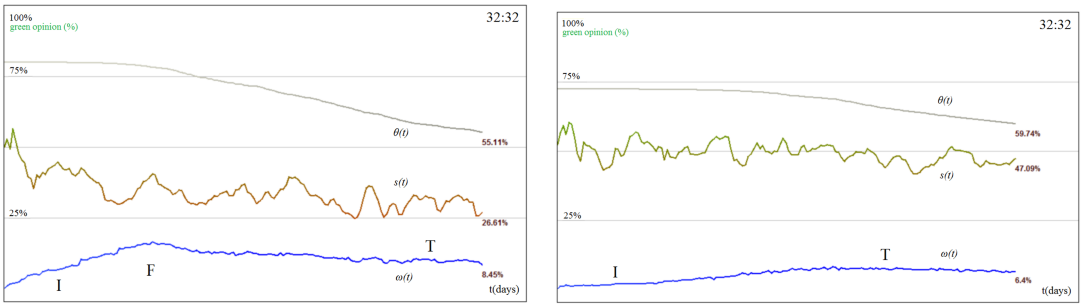
(a) (b)

Figure 6.16.: Simulation results for the tolerance-based opinion interaction on a small-world network with 10,000 nodes with 32:32 green-red SAs. **a**. There are no NullAgents in the population. **b**. The population consists of 20% randomly placed NullAgents.



(a) (b)

Figure 6.17.: Simulation results for the tolerance-based opinion interaction on a small-world network with 10,000 nodes with 32:32 green-red SAs. **a**. The population consists of 30% randomly placed NullAgents. **b**. The population consists of 40% randomly placed NullAgents.



(a) (b)

Figure 6.18.: Simulation results for the tolerance-based opinion interaction on a small-world network with 10,000 nodes with 32:32 green-red SAs. **a**. The population consists of 50% randomly placed NullAgents. **b**. The population consists of 80% randomly placed NullAgents.

and then generalize the results in a qualitative discussion). Because my work adds significant complexity, considering that node tolerance is not a fixed threshold, but a dynamic one which depends on the interactions with neighbors, solving stationarity equations require even more simplifying assumptions. That was the reason for using simulation in order to analyze the stationarity situation. Nonetheless, in all simulation scenarios, the obtained stationarity described in the study coincides with one of two exclusive cases:

- The society reaches **intolerance**, i.e. the overall tolerance converges towards 0. When this happens, my model implies that opinion dynamics comes to a halt. As such, no further modifications to the state of the society can be achieved. I obtain this behavior on **mesh** and **scale-free** topologies. Meshes imply only local connectivity to neighbors that converge towards a similar state, thus tolerance is bound to decrease to 0 (see Figure 6.10a). Scale-free networks imply connections to hub nodes, which means that all neighbors are influenced by the same local hubs, which in turn decreases tolerance to 0 (see Figure 6.11c). Such a situation, in the case of regular small networks, was already mathematically described by [4]. The authors measure the probability of being influenced by a SA using random walks.

  As such, in reference [4] it was proven that:

  1. The stationary expected belief of the society (regular agents) is a linear interpolation of the beliefs of the stubborn agents. Since I use a balanced population of SAs (32 green, 32 red), it is expected that opinion oscillates around the value of 0.5. This fact is validated in my small-world experiments which show social balancing.
  2. The belief of each regular agent keeps on fluctuating ergodically around a value which depends on the relative distance of the agent from the two stubborn agents. This is directly deductible on all used topologies, as is shown by the clusters forming around stubborn agents.

  Using Equation 6.3 it is clear that if the society reaches intolerance ($\theta = 0$), for the majority of

nodes), the equation for regular agents becomes:

$$x_i(t) = x_i(t+1)$$

so the state of the society becomes stable. The requirement for such a stationary state is thus the condition that tolerance converges towards 0. This fact is deductible if we look at Equations 6.6-6.8. Tolerance decreases as long as a node interacts with neighbors with identical state. In meshes and scale-free networks, nodes within vicinity of each other connect to the same nodes. This leads to a convergence of opinion, so Equation 6.6 will cause tolerance to decrease to 0.

- The society remains in **social balance**, i.e. the overall tolerance converges towards a non-zero constant in time ($\theta(t) > 0,\ for\, t \to \infty$), which causes the state and opinion change to also stabilize (for $t \to \infty$). I obtain this phenomenon on **random** and **small-world** topologies. Small-worlds have the unique feature of being both regular and random in a proportion p, given by the rewiring parameter of the Watts-Strogatz algorithm. Thus, nodes interact with equal probability (for $p = 0.5$) with neighbors with similar opinion, and with distant random nodes with different opinion. A proportional value $p = 0.5$ will keep tolerance at maximal value as can be seen in Figure 6.11b. My experiments use such a value of $p = 0.5$, as can be seen in Figure 6.11b ($\omega = 41\%$). Random networks have an even higher stabilization value for tolerance, as can be seen in Figure 6.12b ($\omega = 44\%$).

  Due to the random distribution of initial opinion and links (in random networks and small-worlds with $p = 0.5$), nodes will oscillate ergodically, and both Equations 6.7,6.8 will be activated with relatively equal probability. This keeps the tolerance variation of each node around a certain convergence value:

$$\theta_i(t) = \theta_i(t-1) \pm \alpha\varepsilon$$

  where both α_0 and α_1 are small integers and imply small variation in $\theta_i(t)$. In such a case, for a relatively stable tolerance, the stationarity can also be described as in [4] (where θ is assumed as fixed).

6.4. Discussion

The results for the proposed tolerance-based opinion interaction model show that, if individual traits are considered for modeling social agents, then I can realistically reproduce real-world dynamical features of opinion formation such as opinion formation phases, as well as their evolution towards social balancing. At the same time, I demonstrate that the dynamics of opinion formation is influenced by topology, network size and stubborn agent (opinion source) distribution across the entire population. Overall, the topology seems to have the strongest influence on opinion formation and spread; this can be summarized by the following different tendencies:

- **Responsive behavior**: Tolerance stabilization is attained right after reaching a relatively low critical ratio of stubborn agents. Inserting additional stubborn agents entail a drop in autonomy and opinion flow. Such a behavior is achieved by random and small-world topologies, and it can be motivated by the uniform degree distribution and the existence of both local and long-range links, which foster opinion diversity and *social balancing*; this can be representative for a decentralized and democratic society.

- **Linear behavior**: The critical threshold at which tolerance becomes stable for mesh topologies increases linearly with the stubborn agents concentration. The mesh topology corresponds to a limited, almost "autistic" social interaction behavior (where each agent only interacts with close proximity neighbors); therefore, the probability of opinion diversity only increases with the proportional addition of stubborn agents. For meshes, *social balancing* is attained only if a substantial number of stubborn agents is inserted.

- **Saturated behavior**: Tolerance converges slowly around a specific low value. This behavior is achieved in scale-free and WSDD networks. Due to the nature of these topologies, even though long-range links exist, nodes tend to be preferentially attached to the same hub nodes, meaning the same opinion sources. The amount of stubborn agents required to reach *social balance* is much higher and the resulting balance saturates quickly. It is thus a conservative, stratified and oligarchic type of society which reacts later and slower to new stimuli. Most individuals within this type of society remain intolerant and opinion change is treated as suspicious and non-credible.

Besides these original contributions, the results obtained with my model confirm prior studies which show how individuals converge towards the state of their ingroup [192, 269]. This is especially noticeable on networks with high modularity, like the WSDD network in which every member in a community converges towards the community's dominant opinion, yet every community converges towards a different state.

Inspired by works of Axelrod on the evolution of cultural traits [18, 229] this contribution has found much inspiration from interdisciplinary studies.

An important real-world aspect of my new tolerance model (which assumes that the level of acceptance of neighboring opinions evolves over time) is that the tolerance level of an agent $\theta_i(t)$ is proportional to the degree of the node. In other words, the more neighbors a node has, the more likely it is to receive different influences which can guarantee a higher tolerance level. This observation is backed up by a recent study which proves that individuals with a higher (in)degree are less

likely to be influenced, and the influence of friends is not significantly moderated by their friends' indegree and friendship reciprocity [101].

The results rendered with my tolerance model also fall in line with a research direction started by Gross et al. [111] where the authors show that there is a self-organization in all adaptive networks, including multi-agent opinion networks. My real-world observations and opinion simulation results show a similar topological self-organization based on stubborn agent topological properties.

Finally, the study of opinion dynamics through my proposed concept of *social balancing* shows key features that may be used in practical applications, like marketing or conflict resolution. Under the requirement of keeping the social state stable, while never reaching intolerance, I provide a classification of network topologies based on the social balancing property. Networks with the democratic small-world structure promote balancing; the phenomenon is also exhibited if there is a high concentration of stubborn agents to stabilize opinion in mesh networks. If there are significantly fewer stubborn agents in the network, balancing will only be achieved if one side is using a placement strategy to counter its rivals [103]. A small-world network will not offer an advantage to any of the opinions due the link layout and uniform degree distribution. On the other hand, the oligarchic scale-free topology shows a clear importance of strategically placed agents in hub nodes which intrinsically render the opposing nodes on lower levels of the tree virtually powerless. The balancing phenomenon does not occur in networks with scale-free properties. Clearly, the social balancing concept remains open for further debate, improvement, and real-world validation.

6.5. Methods

I rely on the following datasets, which contain opinion fluctuation data with time information:

The Yelp dataset: contains graded (1-3 stars) user reviews of American businesses, each with a timestamp. One can obtain insights on the popularity of a business at a given time. The usable information is the number of reviews at a given moment in time (interpreted as network size of individuals with an opinion), the average grade in time (the average opinion over time), and the number of votes to each review (ratio of agents with strong or "stubborn" opinions, because when an agent votes, his opinion is already made up). The dataset contains 366,715 users, 61,814 businesses and 1,569,264 reviews.

MemeTracker and Twitter hashtags with time information from the Stanford Large Network Dataset Collection (SNAP); which contain the history (repost rate in time) of diverse, popular hashtags. I can use this data to analyze the evolution of a particular opinion in time. MemeTracker phrases are the 1,000 highest total volume phrases among 343 million phrases collected within 2008-2009. Twitter hashtags are the 1,000 highest total volume hashtags among 6 million hashtags from Jun-Dec 2009.

6.5.1. Discrete simulation methodology

Like any discrete event simulation, I define the salient properties of the experimental setup which was used to obtain the simulation results with my Java-based opinion dynamics simulator, SocialSim [253].

Events are synchronized by the simulation clock; the period of this clock is called a simulation *day*. One day is a simulation period in which agents can interact with their neighbors. However, an agent does not interact daily, in fact each agent picks a random number of days to be inactive after each

active day. In the presented simulations, I have defined a random timeout interval between 1 day and 50 days. Only after this time has elapsed, will an agent interact again with one random neighbor. After that interaction, the agent will again choose to be inactive for a random period of days.

6.5.2. Simulators for social networks

There is a wide variety of tools for visualizing and manipulating graph data. I discuss the most popular tools, their advantages and their limitation in terms of simulation.

Gephi is one of the most popular data visualization tools [30]. It is open-source, plug-in based, has a wide variety of tools for social networks researchers, and it provides a rich framework that helps developers extend functionalities using Java. Its main functionalities are that of importing graph data in multiple formats, visualizing data using various intuitive layouts, measuring graph metrics, coloring nodes and communities based on custom criteria, filtering out nodes based on custom conditions, exporting data as images etc. In terms of performance it scales good with larger graphs but as nodes contain more data the visualization and loading times escalate. There is the ability to measure dynamic events (e.g. creation of nodes, addition of a new edge) but there is no implementation or support for diffusion models and real-time graphical feedback.

Cytoscape is another open source software platform for visualizing complex networks and integrating these with any type of attribute data [238]. It excels at visualization, being used for many researchers in biology and genetics to highlight protein interactions, cell-signaling pathways and other phenomena. GraphViz has been around for more than 20 years and is an open source graph visualization software. It has important applications in networking, bioinformatics, software engineering, database and web design, machine learning, and in visual interfaces for other technical domains[85]. Pajek is a more low level framework which is under constant development [31]. It offers high customization to developers and can be used as a simulator. Other existing state of the art tools are iGraph [68], Tulip [16], GUESS [7], Neo4J, yED, and Walrus.

In terms of simulating diffusion in complex networks there are powerful tools in other adjacent field like infectious diseases and computer security. EpiFast offers parallelized stochastic disease propagation in large contact networks [36]. EpiCure is a scalable high performance computing oriented modeling environment to study malware propagation over realistic mobile networks [53]. Both of these tools offer the ability to simulate millions of nodes, given the necessary hardware resources. Complementary with the presented results in this chapter, I have also highlighted the problem that there are no *complete* tools for studying opinion diffusion in social contexts, nor are there any tools capable of working with huge data sets of million of nodes. I hope that SocialSim will be able to better fulfill the needs of researchers worldwide.

7. Conclusions

The work presented in this thesis represents an original exploratory analysis meant to highlight the role of social networks in a multitude of sciences, and to discover and compare the results of the proposed models with observations taken from medicine, computer networks, voting etc. The exploratory studies rely on a set of observations using real-world data. These have helped me assemble a wide and valuable perspective over social networks analysis and modeling.

As such, I mention two studies regarding the collaboration of musicians (MuSe Net) [258] and fashion models (FMNet) [254], which represent innovative and applicative approaches which bring novelty to literature. In both studies I have used graph metrics and centralities analysis to showcase the importance of the emergent communities which develop and explain real-world particularities of the two artistic fields. Together with a theoretical study of the impact of the underlying topology on online social networks [261], all these studies have helped me understand the importance of graph metrics like average degree, path length, clustering coefficient, diameter, graph density and modularity, as well as the role of centralities like degree, eigenvector and betweenness.

Additionally, I mention two studies of complex networks applied in medical science, following the so-called path of network medicine. The results obtained in predicting central sleep apnea [186, 265] bring landmark novelty and improvement in the field of sleep medicine, The study undertaken of assessing the treatment response of patient with hypertension [247] also showcases a new, useful, perspective for medical doctors. These experiences outside the field of social networks analysis have helped me greatly to understand the role of different metrics to take into consideration when modeling empirical data using graph analysis.

The observations obtained in all these empirical studies have helped me pave the way for the essence of my thesis, namely understanding social structures and creating mathematical models which can reproduce the topology, dynamicity and interactivity within social networks. Consequently, two main directions of original improvement have been presented:

First, on a topological level, I proposed to create a more realistic topological model, called Genosian. It was compared to the existing synthetic models and validated using the statistical fidelity metric and online social network datasets. In accordance with current work, my proposed topology creates a synergy between the small-world and scale-free networks. The small-world property [279] is useful for creating small clusters of individuals, and the scale-free property [25] greatly improves the realism of the degree distribution. In order to achieve the needed realistic graph parameters (e.g. triadic closure, modularity, path length), genetic algorithms have been used. Using empirical datasets for validation, my topology achieves a 63% better realism compared to the best existing alternative synthetic topologies. Another contribution on the same topological level is the explanation of network growth in time [264]. I discuss degree preferential attachment and its limitations, as presented in literature, then come up with a more realistic alternative: the betweenness preferential attachment. This concept is implemented into a modified Barabasi-Albert algorithm [25], where betweenness is used instead of degree. The resulting networks have, on average, a realism of 80-90%, as compared

to the scale-free network which lies within the interval 60-70% in terms of similarity to empirical datasets. All the realism assessments were done using the proposed and validated fidelity metric [257, 256], presented in Appendix A.

Second, on the interaction level, I designed the tolerance based model which better explains how individuals adopt opinion and adapt in time [262]. The socio-psychological discussion backing up this proposal is supported by a rich literature review, and, as such, I propose for the first time a fully dynamic-threshold model in opinion dynamics. The model is inspired by observing four opinion formation phases on empirical data: Twitter replies, MemeTracker replies, and Yelp reviews. The statistical analysis, using the fidelity metric [256], supports the fact that the observed opinion formation phases are representative. This analysis led to the mathematical description of the tolerance model. The simulation results of this model are interpreted by the results obtained from the research conducted in Appendix B [261]. The classes of topologies analyzed in the study using network motifs helped me better understand the types of behaviors that are present in opinion diffusion based on the underlying topology.

It can be admitted that the work presented in this thesis provides many entry point for further research and experimentation, as the plethora of contributions is diverse. It also facilitates the study of new proposals in the field of social networks science, and it allows new proposals to be added and analyzed. I propose to remain on the same track of studying and improving models in SNA.

7.1. Publications and milestones

To this date, I have the following publications at international conferences or international journals, and am a member of the following project teams. I have included a list of the main articles relevant to my PhD thesis, which are all submitted, accepted and presented at international conferences relevant to computer science, or submitted and accepted at computer science journals.

7.1.1. Social Networks Analysis

Applied research projects

1. I am part of the 2-year research project NOVAMOOC led by Assoc. Prof. Gabriela Grosseck, from West University Timisoara. The project is named "Dezvoltarea şi implementarea inovativă a MOOCurilor în învăţământul superior" (Innovative implementation and development of MOOCs in higher level education) and has the identifier PN-II-RU-TE2014-4-2040. Its goal is to create the first massive open online courses (MOOC) for highschool teachers in Romania. This endeavor is a premiere in Romanian education and my role in the project team is to analyze the control groups of students used for motivational and technical tweaking of the platform. The analysis consists of skills acquired from social networks analysis, namely community detection and analysis, and network growth.

2. I am part of the 2-year research project MORPHEUS led by Assoc. Prof. Stefan Mihaicuta, from Victor Babes University of Pharmacy and Medicine in Timisoara. The project is a joint endeavor of the ACSA team (from the Department of Computer and Software Engineering) and the Department of Pneumology to improve the prediction and diagnosis accuracy of sleep apnea, at an EU level. In over 3 years of collaboration, we have over 10 medical congress

attendances with notable distinctions. I received gold sponsorship (<20 awarded in the world in 2014) for the ERS congress from Munchen in 2014.

International journals with impact factor

1. **Alexandru Topirceanu**, Alexandra Duma, Mihai Udrescu (2016). Uncovering the Fingerprint of Online Social Networks Using A Network Motif Based Approach. In Elsevier *Computer Communications* (vol. 73PB, pp. 164-172). **IF=1.695**.

2. **Alexandru Topirceanu**, Mihai Udrescu, Mircea Vlăduțiu, Radu Mărculescu (2016), Tolerance-based interaction: A new model targeting opinion formation and diffusion in social networks, In *PeerJ Computer Science*, 2, p.e42.

3. **Alexandru Topirceanu** and Mihai Udrescu (2016). Statistical Fidelity: A Tool to Quantify the Similarity Between Multi-variable Entities with Application in Complex Networks. In *International Journal of Computer Mathematics.* (accepted), **IF=0.825**.

Book chapters at international publishers

1. **Alexandru Topîrceanu**, Mihai Udrescu, Mircea Vlăduțiu. Genetically Optimized Realistic Social Network Topology Inspired by Facebook. In *Online Social Media Analysis and Visualization* (pp. 163-179). Springer International Publishing, 2014. **ISI indexed**.

International conferences

1. **Alexandru Topîrceanu**, Mihai Udrescu, Mircea Vlăduțiu. Network Fidelity: A Metric to Quantify the Similarity and Realism of Complex Networks. In *Cloud and Green Computing* (CGC+SCA), 2013 Third International Conference on (pp. 289-296). IEEE, **ISI indexed**.

2. Alexandra Duma, **Alexandru Topirceanu**. A network motif based approach for classifying online social networks. In *Applied Computational Intelligence and Informatics* (SACI), 2014 IEEE 9th International Symposium on (pp. 311-315). **ISI indexed**.

3. Gabriel Barina, **Alexandru Topirceanu**, Mihai Udrescu. MuSeNet: Natural patterns in the music artists industry. In *Applied Computational Intelligence and Informatics* (SACI), 2014 IEEE 9th International Symposium on (pp. 317-322). **ISI indexed**.

4. **Alexandru Topirceanu**, Gabriel Barina and Mihai Udrescu. MuSeNet: Collaboration in the Music Artists Industry. In *Network Intelligence Conference* (ENIC), 2014 European (pp. 89-94). IEEE. **ISI indexed**.

5. **Alexandru Topirceanu**, Mihai Udrescu. Measuring Realism of Social Network Models Using Network Motifs. In *Applied Computational Intelligence and Informatics* (SACI), 2015 IEEE 10th International Symposium on (pp. 443-447). **IEEE indexed.**

6. Mihai Udrescu and **Alexandru Topirceanu**, What Drives the Emergence of Social Networks? In *Control Systems and Computer Science* (CSCS), 2015 20th International Conference on (pp. 999). **IEEE indexed.**

7. **Alexandru Topîrceanu**, Dragoş Tiselice, Mihai Udrescu. The Fingerprint of Educational Platforms in Social Media: A Topological Study Using Online Ego-Networks. The 2nd International Conference on *Social Media in Academia: Research and Teaching* (2014). ***Pending ISI indexing.***

8. **Alexandru Topirceanu**, Mihai Udrescu (2015, September). FMNet: Physical Trait Patterns in the Fashion World. In *Network Intelligence Conference* (ENIC), 2015 2nd European, Received **best-paper award**. **IEEE*, pending ISI indexing.***

Additionally, since my passion for the thesis domain exceeded the envisioned goals, I have additional original contributions in social sciences, medical science, and computer engineering. Some of the papers listed in this section are still pending indexing, but all have been accepted.

7.1.2. Network Medicine

International journals with impact factor

1. L. Suciu, C. Cristescu, **A. Topirceanu**, L. Udrescu, M. Udrescu, V. Buda, M.C. Tomescu (2015). Evaluation of patients diagnosed with essential arterial hypertension through network analysis. *Irish Journal of Medical Science* (1971-), p. 1-9. **Impact factor=0.827**.

2. Ștefan Mihăicuță, Răzvan Avram, **Alexandru Topîrceanu**, Mihai Udrescu (2013). A Network Based Approach to Sleep Apnea Syndrome. *European Respiratory Journal*, 42(Suppl 57), P2046. **ISI indexed**,

3. Mihai Udrescu, **Alexandru Topirceanu**, Razvan Avram, Ștefan Mihăicuță (2014), AER Score: A Social-Network-Inspired Predictor for Sleep Apnea Syndrome, In *Chest Journal*, 145(3), 609A. **ISI indexed.**

International medical conferences

1. **Alexandru Topîrceanu**, Mihai Udrescu, Răzvan Avram, Stefan Mihăicuță. Data Analysis for Patients with Sleep Apnea Syndrome: A Complex Network Approach. The *6th International Workshop On Soft Computing Applications* (SOFA 2014), ***Pending ISI indexing***.

2. Alexandru Iovanovici, **Alexandru Topirceanu**, Mihai Udrescu, Lucian Prodan, Stefan Mihaicuta (2014). A high-availability architecture for continuous monitoring of sleep disorders. *Studies in health technology and Informatics*, 210, 729-733. **PubMed indexed.**

7.1.3. Communication Networks

1. **Alexandru Topîrceanu**, Alexandru Iovanovici, Cristian Coșariu, Mihai Udrescu, Mircea Vlăduţiu. Social Cities: Redistribution of Traffic Flow in Cities Using a Social Network Approach. The 6th *International Workshop On Soft Computing Applications* (SOFA 2014). ***Pending ISI indexing.***

2. Alexandru Iovanovici, **Alexandru Topîrceanu**, Mihai Udrescu, Mircea Vlăduţiu. Heuristic Optimization of Wireless Sensor Networks using Social Network Analysis. The 6th *International Workshop On Soft Computing Applications* (SOFA 2014). ***Pending ISI indexing.***

3. Alexandru Iovanovici, **Alexandru Topirceanu**, Mihai Udrescu, and Mircea Vladutiu. Design Space Exploration for Optimizing Wireless Sensor Networks using Social Network Analysis. In *System Theory, Control and Computing* (ICSTCC), 2014 18th International Conference, pp 815 - 820. ***Pending ISI indexing.***

4. **Alexandru Topirceanu**, Alexandru Iovanovici, Mihai Udrescu, Mircea Vladutiu. Social cities: Quality assessment of road infrastructures using a network motif approach. In *System Theory, Control and Computing* (ICSTCC), 2014 18th International Conference, pp 803 - 808.. ***Pending ISI indexing.***

7.1.4. Research milestones

My 3-year period of doctoral studies has been planned with the following milestones and results, as presented in the following table. In order to achieve, not only these goals, but a meaningful contribution to the field of social networks analysis, the scientific activity revolving around the field had to be kept under constant monitoring on an international level to ensure that the proposals made in this thesis are new, original, coherent, relevant, and useful. Important research centers for social networks can be found in the USA (at MIT, Notre Dame, Cornell, Stanford, Northeastern, Michigan), Israel, Sweden, Turkey, Italy and the UK. Their results are well acclaimed by the scientific community as they are published in renowned journals like: Nature, Science, Physics, PlosOne, PerrJ, Elsevier Social Networks, New England Journal of Medicine.

Nr	Description	Results	Due date
1	Analytic study into the state of the art to balance the proposals of the thesis with the current hot topics of interest for social networks analysis.		Continuous, 2013-2015
2	Propose and develop a complex topology to model social ties between individuals as realistic as possible. Also make it customizable for specific simulation scenarios.	The Genosian algorithm for creating realistic social network topologies.	March 2013
3	Propose one or more metrics to quantify: (1) the realism of a social network topology, (2) the similarity between two complex networks, (3) the sociability of a node in regard to the social features	(1) and (2): The statistical fidelity metric. (3): sociability S for complex networks	December 2013
4	Keep track of new emerging fields, as many fundamental sciences add social network analysis to their research methodologies (e.g. medical fields, sensor networks, genetics, gamification, gaming theory, marketing strategies etc).	Involved in the Morpheus (sleep medicine) and NOVAMOOC (educational analytics) projects.	Continuous, 2013-2015
5	Further development of SocialSim in regard to releasing it as a simulation and social scenario testing tool.	Simulator is available online	March 2014
6	Refine and validate the social interaction model proposed during the dissertation.	The tolerance based interaction model.	January 2015
7	Confronting the proposed diffusion model with real data gathered from diverse empirical sources (Twitter, Facebook, opinion polls, user-expressed opinion etc.).	Validation of the tolerance model using Twitter, MemeTracker chat and Yelp user reviews.	April 2015
8	Collaborate with other researchers from around the world in one or more joint contributions.	Ongoing collaboration with Radu Marculescu from Carnegie Mellon University (submission at PeerJ Computer Science)	September 2015
	Future milestones		
9	*Apply the structural and behavioral model on real-world market and political situations.*		*2016-2017*

7.2. Future research directions

Due to the passion for research, and backed up by the encouraging results during my years as a PhD student, I foresee two directions to contribute to SNA:

- Network growth model: nodes will also evolve in age (i.e. they will die after some time) and the links of the topology will be layered depending the type of interaction (i.e. multi-layered social networks [173]). The locality of the long range links will also be influenced by the interaction type, and will determine a better, more realistic mixture of agents which can communicate in the society.

- Improving the tolerance based interaction model: addition of new parameters to better model the human subjective nature of taking decisions. Tolerance is but a first step in improving agent based interaction. The egocentric model (i.e. opinion is changed only by personal beliefs of that person) also covers other concepts like trust or confidence. These may be added on edges and model the friendship strength of two agents. In addition to the egocentric nature of the model, there is also an exocentric model (i.e. opinion only depends on the other person's parameters). New parameters include an agent's credibility or authority. External parameters are public opinions of a person, shared by all agents initiating communication. All these parameters may be added in time, after a thorough validation.

- An even more customizable interaction model may be necessary for real-world prediction as it has been seen that people use different behaviors in different situations. As such, I might need to parametrize the classes of opinion as well, and then parametrize the interaction model accordingly.

7.3. Closing thoughts

For many fields of science like Psychology, Philosophy, Politics, Marketing, Finances and even Warfare, understanding social opinion dynamics is a major concern. One of the requirements of improving profits, for example, is the study of markets and consumers. Economy and marketing strive to better understand the needs of consumers, but also their strengths and weaknesses. Winning elections is one of the major goals in Politics. Because political parties are always interested in the overall public opinion rather than the opinion of individuals, social studies are used to understand the political influence of parties and the means to create a consensus among voters. Even in war there has always been the need to understand social opinion dynamics. Counter-intelligence to stop enemy propaganda and spies to influence the enemy's morale have been used for a long time in confrontations. Medicine uses social networks to model the dynamics of diseases, to help determine the outbreaks of infections and the stop the spreading of epidemics.

Regardless of the science in question, social studies are still in their infancy from a theoretical point of view. There is the need for a fundamental set of rules, as human behavior is mostly unpredictable. This makes the mentioned market study, political propaganda, infectious spreading, and voting somewhat unpredictable to better understand the social processes a better collaboration between natural sciences and applied sciences is needed, as both possess valuable knowledge. Using the current computational power, computers can help researchers analyze interleaved mathematical

and psychological models at a faster rate. Of course, validating results with empirical data is the final step in proving that a social process is understood. Recent mathematical research proposes new ways of modeling societies or clusters of individuals and present results of great theoretical value.

Having evolved from basic network topologies, like the mesh and ring, complex networks have emerged by studying empirical networks in our world. Ranging from natural networks, like food-chains, actor's relationships, protein chains and correspondence patterns, to synthetic networks, like the World Wide Web and airplane traffic, these networks have generated interest in engineering around the world. However, better understanding of social networks, fostered by social, economic and marketing research, has led to the proposal of newer and more advanced topologies which better resemble real networks.

On a parallel direction, there is the possibility of researching social studies and predictions based on results gathered from social platforms like Twitter and Facebook. Because of the popularity of these platforms the available data quantity is near infinite. Opinions can be quantified and opinion dynamics can be monitored through time. This is a mandatory step in order to demonstrate a theoretical proposal. However, the relevance of the acquired data is still somewhat limited due to the facts that social platforms are mainly popular among young people; they have a limited, heterogeneous spreading around the world and have relative impact due to the disproportionate popularity of a certain binary decision, or belief.

Modern social science is trying to create improved topologies that better resemble the real world. Many studies focus on refining the way individuals interconnect into forming a more realistic social layer (see chapter 4). Other focus on understanding how new individuals are added into layers or clusters of society, modeling realistic growth (see chapter 5). The studies presented in this thesis conclude with the fact that individuals present different behavioral patterns depending on the type of belief, and they interact with different persons depending on the problem circumstances. With this in mind, it cannot be denied that on the level of social opinion expression there is no single correct topology. The only accurate method for modeling social behavior is research and understanding the topological layers that come into action in various situations (see Appendix B). Based on observations drawn throughout the thesis, it is considered that most individuals will use small-world networks to interact with their family and friends. In other situations, the same individuals will often cluster in WSDD-like networks at parties of public debates, where most stay within a group of moderate size, and very few will stay in small or very large groups. People carrying infections or planning terror attacks will travel along scale-free networks using the airways. Finally, people expressing political or religious (i.e. strong beliefs that rarely change) belief are more likely to be organized as a mesh network. As polls are centralized, public opinion plays a major role in the behavior of individuals, thus the impact of long range connections with individuals outside the cluster tends to be minimized. As in a city with mesh-like streets separating neighborhoods, a population at a larger scale (e.g. region, state, country) will behave as if in a mesh topology for a certain set of strong beliefs. The complexity with which all these factors combine has been discussed in chapter 6 and appendix B.

As a conclusion, social research is proving useful in understanding the mechanisms which transform individual opinion into a wide-spread social opinion; how opinions affect individuals and how they evolve in time as seen on a macroscopic scale. Modeling social behavior can be both a means of defending and boosting democratic rights as well as a means to impose and manipulate a society or a social layer.

As one of the top researchers in big data, Alex Pentland explained *"The fact that we can now begin*

to actually look at the dynamics of social interactions and how they play out, and are not just limited to reasoning about averages like market indices is for me simply astonishing. To be able to see the details of variations in the market and the beginnings of political revolutions, to predict them, and even control them, is definitely a case of Promethean fire. Big Data can be used for good or bad, but either way it brings us to interesting times. We're going to reinvent what it means to have a human society" [219].

Appendix

A. Statistical fidelity: quantifying similarity between multi-variable entities

The computer-based analysis of complex networks relies on fundamental properties of natural and synthetic networks that surround us. These properties of networks are characterized by graph metrics, yet there is no unified computational method for comparing networks to each other. To address this issue, I introduce the new statistical fidelity metric, which can compare any types of complex networks, by using specific individual metrics. The composite fidelity metric offers an insight on the structural similarity of networks, as well as on the topological realism of synthetically-generated networks. I also provide an overview of the alternatives to measuring similarity, then I apply my metric in the context of social and technological networks. This way, I highlight the superior analytic power of my composite metric compared to cosine similarity, Pearson correlation, Mahalanobis distance, and fractal dimension. Therefore, network fidelity is able to better capture the combination of fundamental properties of complex networks.

"There is no law except the law that there is no law."

 John A. Wheeler

A.1. Motivation

Network science is receiving an increased interest from many fields of science, since many empirical observations of our surrounding world show the same properties, regardless of whether the underlying complex graphs are of natural or synthetic origin [89, 154]. There are topological network models which describe geographical proximity, friendship distribution, brain neural networks, protein interaction mechanisms, natural food chains, the distribution of means of transportation, citation networks, sexual interaction patterns, the World Wide Web, power distribution networks, relationship of words in a language, interaction between ingredients in a recipe, the world markets, political structures [10, 276, 95, 250, 131, 82, 214, 154]. As such, complex networks fall into four main categories [276]: technological [154, 274, 259], biological [11, 82, 178], social [82, 10, 57], and semantic networks [276, 250].

Within the network paradigm, the capacity to collect and analyze massive amounts of data is transforming fields like biology, economy and physics [154]. However, the emergence of data-driven computational science has been much slower, carefully directed by a few intrepid computer scientists, physicists, and social scientists [187, 281, 25, 205, 154]. Regardless of the representation of nodes, edges, edge directions, and edge weights, graph models of big data [95, 89] are often subjected to numerical comparison, sampling, and statistical analysis to extract relevant patterns. To that end, network scientists employ a wide range of state of the art comparison techniques, but there is no single computational methodology to express similarity/dissimilarity in an objective and synthetic manner.

In light of existing research on network metrics [246, 276, 154], this appendix presents a solution to the problem of quantifying the comparison between any two complex networks, based on a set of metrics. Additionally, by comparing to a reference network structure, I can order any other topologies - synthetic or empirical - by the degree of similarity to that chosen reference structure. To this end, I propose a similarity metric, namely the network fidelity φ (phi); it uses only the topological properties of the underlying graph, and has symmetric and scale free properties, which are further discussed in this section.

I also propose to showcase how φ may be applied by researchers in any network-related context. First, I rely on a social network context to illustrate the effectiveness of my fidelity metric in terms of realism assessment of widely accepted synthetic topologies. A second illustration for φ is in the context of measuring similarity between cities, modeled by their road infrastructure network. As such, I use a set of road networks which are represented through their distribution of network motifs, i.e. recurrent and statistically significant subgraphs or patterns in complex networks [189, 259].

This section also provides a review of the current statistical methods being used in complex networks analysis and sets out to demonstrate that no existing methodology offers the same analytical value as the fidelity φ. I compare the efficiency of φ against available statistical methods (cosine similarity, variance, covariance, Pearson correlation, Mahalanobis distance), as well as against the fractal dimension [242]. Taken together, φ offers a better (and synthetic) overview of the analyzed data sets; it is also being integrated by the authors into Gephi [30] - the leading tool in visualization and analysis of large networks.

The appendix is organized as follows. Section A.2 describes the perspectives and existing solutions to measure network similarity, then presents the empirical and synthetic data sets used in the study for validation of my metric; Section A.3 explains the mathematical theory behind the fidelity metric;

Section A.4 presents a detailed discussion that proves the superior power of my proposal compared to existing similarity measures, and explains how one can define a realism threshold for synthetic networks; Section A.5 draws the main conclusions for this proposed metric.

A.2. Analyzing complex network structures

In this section I present an overview on the state of the art concerning the comparison of complex networks, and discuss the corresponding statistical methods. Furthermore, I introduce an experimental setup that aims at exemplifying the usage of φ in both technological and social networks.

A.2.1. Graph metrics and motifs

Empirical studies done over a variety of natural and synthetic networks have resulted in the definition of several metrics used to describe and measure networks. Out of these, the fundamental metrics are: the average path length (L), the clustering coefficient (C) and the degree distribution ($P(k)$) [281, 25, 246]. A more in-depth analysis of complex networks is obtained by measuring the centralities, modularity, graph density and diameter [10, 276].

The average path length (L) of a network is the mean distance between two nodes, averaged over all pairs of nodes [276]. The clustering coefficient (C) is defined as the average fraction of pairs of neighbors of a node that are also neighbors of each other [276]. The degree of a node is defined as the total number of its (outgoing) edges. Thus, the average value of the degrees, measured over all nodes, is called the average degree of the network. The degree distribution over a network is characterized by a function $P(k)$ [276]. The diameter of a network is the maximal distance among all distances between any pair of nodes in the network [276]. The network density is defined as the ratio of edges in the network to the total number of possible edges [150]. Modularity is a measure that shows the strength of the division of a network into communities [205]. A high modularity means a strong presence of well-delimited communities, while a low modularity models an interleaved small-world society.

Another considered property of graphs are network motifs, which were introduced by Milo et. al. [189]. Each motif is a subgraph defined by a particular pattern of interaction between graph nodes, and can reflect a framework in which particular functions are achieved efficiently. Motifs have recently gathered much attention as a useful concept to uncover structural design principles of complex networks [178]. Although network motifs may provide a deep insight into the functional abilities of a network, their detection is computationally challenging even by current standards.

Considerable progress has been achieved in the areas of biology and genetics, where motifs are associated with functional roles of transcription regulation networks which control the expression of genes [11]. Another related study proves that motifs can be efficiently used to describe urban topologies and quantify their traffic flow properties [259].

A.2.2. Network specific comparison methods

Comparing data that characterizes real-world systems and phenomena aims at a deeper understanding of the interaction patterns between these systems [281, 25, 246, 95, 131]. Using network modeling to reveal shared topological properties of different networks helps understand these patterns even

further [281, 141]. However, current network comparison methodologies suffer from limitation in terms of analytical efficiency [165].

In current network analysis, the comparison of two or more networks is done by performing individual metric comparisons [116]. While such an approach is useful when trying to capture one specific feature of the network, it fails to create a general overview of the similarity between the analyzed networks [66, 165]. Similar work aimed at comparing the importance of graph metrics concludes that each metric captures only specific attributes of the network [35, 49]. Consequently, Bigdeli et al. consider that further study on the effect of each graph metric is needed in order to be able to reliably find synthetic topologies that fit the characteristics of empirical networks [35].

The network dimension is a key feature in understanding not only network topology, but also dynamical processes on networks, such as diffusion, percolation and other critical phenomena [73]. The fractal dimension d_B [242] is proposed based on the belief that complex networks are not invariant or self-similar under a length-scale transformation. Fractal dimension has been measured on multiple varied real-world networks like the WWW, biological networks, actor networks.

A similar topological approach, oriented towards the similarity of graph nodes, proposes a measure based on the concept that two nodes are similar if their immediate neighbors in the network are themselves similar [156]. The approach relies on simulator-generated networks and proves that the topological perspective offers much better insight into similarity than generic statistical metrics.

From a topological perspective, existing studies are available on both classifying complex networks [141, 261] and structural pattern detection [215]. However, rather than being a measure of similarity, these methods serve as meta-analysis techniques..

A.2.3. Statistical methods for similarity

The statistical methods with which I can assess the topological similarity of networks are the cosine similarity [249], variance, covariance, Pearson correlation coefficient (PCC) [245], the Mahalanobis distance [174]. Other methods used in network analysis which are adopted from statistics include the T-test and the ANOVA test (analysis of variance). There is no single statistical approach that is used in current research on complex networks, because there is no unified metric that provides normalized values which are specifically tailored for comparing networks. Yet, the most intuitive and used metrics are the Euclidean distance, Pearson correlation [244] and cosine similarity [249].

The cosine similarity S_C is a measure of similarity between two vectors that expresses the cosine of the angle between them, not from the perspective of magnitude, but from that of orientation. The technique is commonly used to measure cohesion within clusters in the field of data mining [249]. The cosine can be defined as:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{A.1}$$

The resulting similarity between vectors A and B ranges from -1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence. This metric can be applied in my context by building a vector with elements consisting of each measured graph metric of interest. It would translate into a multi-dimensional vector with each dimension given by a different metric (e.g. degree, average path length, modularity)

Variance measures how far a set of given numbers is spread out, proportionally indicating the difference between the numbers. A variance of zero indicates that all the values are identical. The variance can be defined as:

$$var(X) = E[(X - \mu)^2] \tag{A.2}$$

where $E(X)$ is the probability that X occurs and μ is the reference value for X.

Covariance is the measurement of how much two variables change together. The sign of the covariance shows the tendency in the linear relationship between the variables. The covariance between two variables x and y can be defined as:

$$cov(x,\, y) = E[(x - \mu_x)\,(y - \mu_y)] \tag{A.3}$$

The Pearson correlation coefficient (PCC) is a measure of the linear dependence between two variables X and Y, giving a value between +1 and −1, where +1 is total positive correlation, 0 is no correlation, and −1 is negative correlation. The PCC is commonly being used for network comparison, like in the work of Barabási et al. where the authors use it to correlate co-expression networks in the context of gene networks [270]. The PCC for two populations is defined using the covariance as:

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X\, \sigma_Y} \tag{A.4}$$

where σ_X is the standard deviation of X.

The Mahalanobis distance, also known as the D^2-statistic, is a method of multivariate analysis that has gained wide popularity in problems arising in biological, psychological and economic research [22]. It is a descriptive statistic that provides a relative measure of a data point's distance from a common point. Given a vector $x = (x_1, x_2, .., x_n)$, formed by the actual measured metric values, which we want to compare with a reference (mean-value) vector $\mu = (\mu_1, \mu_2, ..., \mu_n)$ the Mahalanobis distance can be defined as:

$$D(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)} \tag{A.5}$$

where S is the covariance matrix.

Fractal dimension (d_B) is a mathematical function representing the topological complexity of a structure, which is also applicable on complex network topologies [242]. As self similarity - the infinite scaling property of fractals - is capable of synthetically capturing the common network structure and offers a way to unravel the universal characteristics of many complex networks [175, 169], I consider it a strong competitor for φ. Fractal dimension is defined as a ratio comparing how detail in a pattern changes with the scale at which it is measured; the most common algorithm for determining fractal dimension in graphs is the box-covering method [243]. This method consists of covering an entire network with the minimum number of boxes N_B of linear size B. If the number of boxes scales with the linear size B following a power-law, then d_B is the fractal dimension of the graph according to the following Equation:

$$N_B(l_B) \sim l_B^{-d_B}$$

A.2.4. Experimental setup

I present a set of empirical data that will be used to demonstrate the efficiency of using φ. While the metric remains open to be used in any complex network context of choice, my demonstration focuses on the fields of social and technological networks. To this end, I introduce validation datasets consisting of empirical online social networks, and a set of urban road networks.

Social network datasets

The empirical online networking data sets are made available by the Stanford Large Network Dataset Collection [157]. As social platforms are unquestionably part of our lives, I use Facebook, the largest virtual society to date with over 1 billion users [93], as the main source for collecting data, but I also use Twitter, Google Plus, Wikipedia and Gnutella.

The data obtained for the online networking sites is represented under the form of *.edges* files which contain the set of edges between nodes. Each edge is defined by a pair of source node ID and target node ID. All data is anonymous (i.e. numeric IDs), as it represents user information from the particular site. I parse the data and create *.gdf* files which are used as input for Gephi. The upper half of Table A.1 details the empirical data to be used as a baseline for demonstrating the efficiency of φ. The presented networks originate from Facebook – FB1 (333 nodes), FB2 (747 nodes), Google Plus – GP1 (347 nodes), GP2 (521 nodes), Twitter – TW1 (231 nodes), Wikipedia – Wiki (7115 nodes) and the Gnutella file sharing service – Gnu (8114 nodes).

Table A.1.: The basic metrics for the seven representative online social networks and the five synthetic topological models. The numerical values for average degree (AD), average path length (L), average clustering coefficient (C), modularity (Mod), diameter (Dmt), density (Dns) are measured using Gephi.

		AD	*L*	*C*	*Mod*	*Dmt*	*Dns*
Empirical	FB1	15.13	3.75	0.557	0.45	11	0.046
	FB2	80.39	2.56	0.639	0.53	7	0.108
	GP1	12.15	3.90	0.404	0.44	12	0.035
	GP2	38.09	2.34	0.447	0.20	7	0.073
	TW1	12.39	2.68	0.239	0.28	7	0.054
	Wiki	14.57	3.34	0.081	0.42	10	0.002
	Gnu	3.21	7.05	0.005	0.45	20	1E-04
Synthetic	S-World	3.99	5.61	0.321	0.73	11	0.005
	S-Free	3.12	4.60	0.015	0.62	10	0.003
	Cellular	11.39	3.79	0.599	0.91	7	0.02
	Geographic	6.63	3.34	0.065	0.52	8	0.013
	WSDD	21.58	4.59	0.738	0.9	9	0.041

In terms of synthetic data, I use a range of network models. My study uses the *small-world* model of Watts and Strogatz [281], the *scale-free* model of Barabasi and Albert, based on preferential attachment [25], the *Watts-Strogatz model with degree distribution* (WSDD) [57], c*ellular networks* [263] inspired from the observation of covert networks like terrorist organizations, and the *static-geographic model* [154] used when taking spatial distances into consideration. The motivation for choosing these five networks (see Figure A.1) lies within the topological diversity of each: the first two are fundamental models for network science, while the latter three combine properties of the the previous.

A representative topology is generated for each of the five synthetic networks, as I present the resulting graph metrics in the lower half of Table A.1. The algorithms for generating all the analyzed networks, as originally described by their respective authors, are implemented as Gephi plug-ins by this paper's authors. Trying to shed light over the results from Table A.1 and Figure A.1 in order to find a synthetic network that best fits an empirical social network is a cumbersome job, because each of these structures fails to capture the nature of the empirical data for most of the metrics. Relying on just a graphical comparison between Figure A.1c (empirical friendship network) and the rest of Figure A.1 can be considered subjective and non-quantifiable. A numerical comparison is much more desirable, but a metric to correctly measure the similarity between the empirical data set and the five synthetic networks does not exist.

Road network datasets

The road information data is obtained from the online repository OpenStreetMap. This data is parsed into a *gexf* file format using a customly implemented python plugin. Thus, I render all intersections as a node list, and all city streets as edges between nodes (i.e. intersections). The edge data is further parsed to extract only a plain edge list text file. This serves as an input for FANMOD (FAst Network MOtif Detection tool) [284], a fast and lightweight tool based on the RAND-ESU algorithm [283]. FANMOD takes as input an edge list in text format, in order to generate a detailed motif distribution statistic.

The explained process is repeated for each selected city. My study relies on six diverse cities for analysis: Augsburg (272K inhabitants, 6097 nodes, 7929 edges), Bratislava (415K inhabitants, 5968 nodes, 7494 edges), Budapest (1.7M inhabitants, 12038 nodes, 17309 edges), Cluj-Napoca (324K inhabitants, 2321 nodes, 3051 edges), Constanta (283K inhabitants, 2794 nodes, 3994 edges), and Timisoara (319K inhabitants, 4070 nodes, 5542 edges). The geographical information originates from Wikipedia, and the graph modeling and analysis is done in Gephi [30]. After obtaining the motif distributions my main goal is to quantify the similarity between the studied cities.

The acronyms of the cities used further in my discussion (subsection A.4.2) and in Tables A.2 and A.10 correspond to: Augsburg (Agb), Bratislava (Brt), Budapest (Bud), Cluj-Napoca (Clj), Constanta (Cns), and Timisoara (Tsr). For motif size $k = 4$, I obtain six relevant motifs that characterize each city, thus $D_4^{city} = \{m_1^{city}, m_2^{city}, m_3^{city}, m_4^{city}, m_5^{city}, m_6^{city}\}$. Each motif occurrence m_i^{city} is a percentage representing the number of motifs of size $k = 4$ that exist in the topology of $city$ (see Table A.2). All motifs are named by the IDs resulting from their adjacency matrix, as found in literature [11, 261].

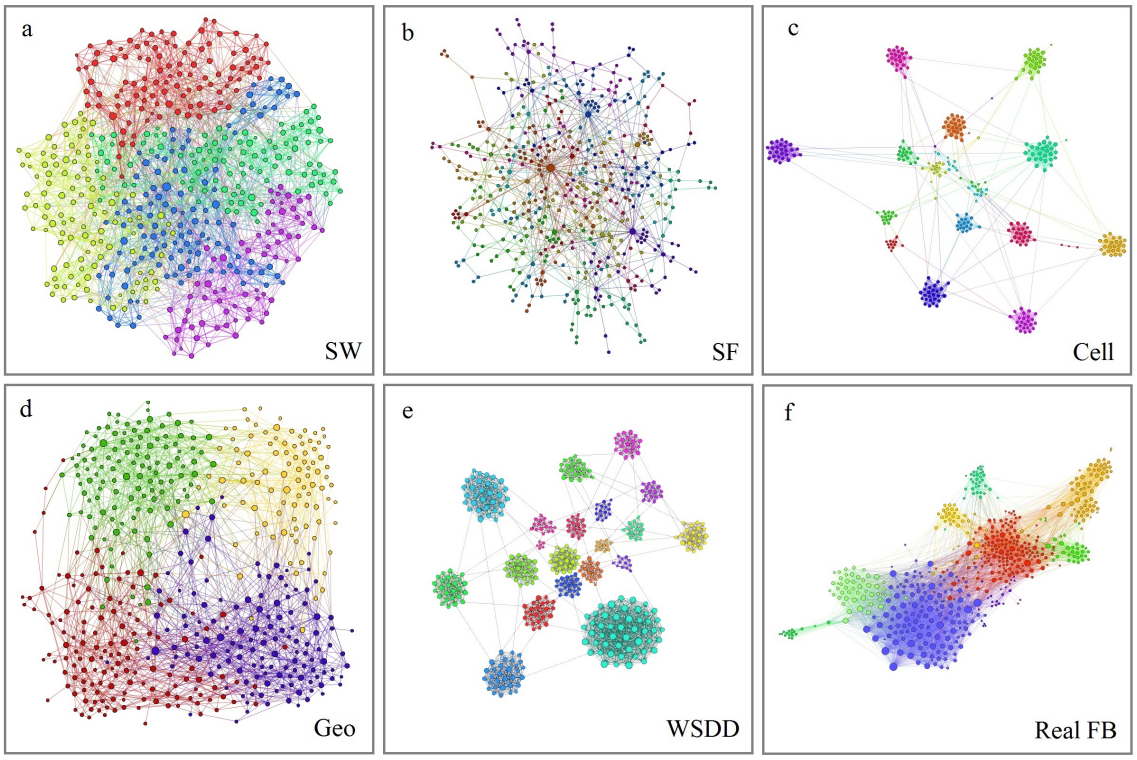


Figure A.1.: State of the art social networks. All topologies are synthetically generated using Gephi. **a.** A small-world network with 500 nodes. **b.** A scale-free network with 500 nodes. **c.** A cellular network with 500 nodes. **d.** A static-geographic network with 500 nodes. **e.** A WSDD network with 437 nodes. **f.** A real Facebook network with 590 nodes. By running a community detection algorithm, all nodes are colored according to their belonging community.

Table A.2.: Occurrences of undirected motifs of size 4 in the road networks of each city. The most relevant six motifs are given, which account for approximately 90% of the total network structure.

		Motif IDs						Total [%]
		28	74	140	392	536	2116	
Cities	Agb	6.02	19.57	21.44	18.66	19.1	5.8	90.6
	Brt	5.62	18.78	23.24	19.03	17.43	5.77	89.9
	Bud	7.22	18.42	16.37	18.44	24.51	7.25	92.2
	Clj	6.23	18.01	18.67	19.93	22.3	6.64	91.8
	Cns	8.33	17.41	11.67	17.13	30.72	8.36	93.6
	Tsr	7.16	18.04	14.92	18.62	26.56	7.39	92.7

A.3. Theory and calculation

Let there be a reference complex network topology $M = (V, E)$, where V is the vertex (node) set and E is the edge (link) set, and another network M_i of any origin, be it natural or synthetic. I want to numerically express the similarity between M_i and the reference M. A maximum fidelity of 1 represents complete similarity, while a minimum fidelity of 0 represents complete dissimilarity between the two compared networks. I do this by measuring and comparing their common individual graph metrics. The fidelity is not dependent on the choice of metrics of interest, however it is customizable to allow a weighted comparison.

Depending on the context of the problem, any numerical value that is representative for a network can be used. In the results section I consider all six mentioned graph metrics as relevant, however, in this section I propose a more abstract exemplification. As an illustrative example, let there be four network metrics, noted as *a, b, c* and *d*, with their respective average measurements m_a, m_b, m_c and $m_d \geq 0$. These represent the averages over an arbitrary number of repeated measurements over M. Now let there be network M_1 which can be described through the same metrics, having m_a^1, m_b^1, m_c^1 and $m_d^1 \geq 0$ as respective measurements. The goal is to quantify the similarity between M and M_1 using the provided measurements, and be able to determine how close the network is to the reference M. The step-by-step mathematical formulas are accompanied by a set of examples, as shown in Table A.3.

Table A.3.: Example values for M and M_1 used to demonstrate the presented formulas.

		a	*b*	*c*	*d*
(1)	M	0.3	0.7	3	50
(2)	M_1	0.4	0.5	7	82
(3)	d_i^1	0.1	0.2	4	32
(4)	r_i^1	0.75	0.71	0.43	0.61

Lines (1) and (2) in Table A.3 define the values of the four metrics for the two networks being compared. The first approach is to determine the absolute distance d between each pair of measurements:

$$d_i^1 = \left| m_i - m_i^1 \right| \tag{A.6}$$

where $i = \{a, b, c, d\}$. Thus, d_i^1 is the distance of metric indexed i of network M_1 towards the metric with the same index of the reference M (e.g. it can express the distance between the modularities of the two networks being compared). Line (3) in Table A.3 shows that while metrics *a* and *b* can be compared using just this simple method, metrics *c* and *d* are of a different magnitude. Thus, a normalization of the values is imposed to obtain a ratio r as follows:

$$r_i^1 = \frac{min\left(m_i,\, m_i^1\right)}{max\left(m_i,\, m_i^1\right)} \tag{A.7}$$

where $i = \{a, b, c, d\}$. The normalization imposes that all values measured on different scales are brought to a common scale for proper comparison. Line (4) in Table A.3 demonstrates this, as all four metrics are brought within the unit interval [0,1] by dividing the smaller value by the larger value of the two metrics. The remaining problem, however, is of intuitive nature, as I can further try to express the dissimilarity (distance) or the similarity between the networks. A distance, as defined by $1 - r_i^1$ would suggest that the closer the metrics are, the closer φ should be to 0. On the other hand, a similarity, as defined by r_i^1, imposes that the closer two metrics are, the closer φ is to 1 (i.e. 100% match). Both formulas can be discussed, but as I want to quantify similarity, I will further use Equation A.7 in defining φ.

After obtaining the normalized comparison values I can combine them in three manners to compute φ: as a product (φ_G: geometric φ), as a sum (φ_A: arithmetic φ) or as a combination of both (φ_H: harmonic φ). By using the example with network M_1, I obtain:

$$\varphi_G = \sqrt[n]{\prod_i r_i^1} \tag{A.8}$$

$$\varphi_A = \frac{\sum_i r_i^1}{n} \tag{A.9}$$

$$\varphi_H = \frac{\prod_i r_i^1}{Avg\left(\frac{\prod_i r_i^1}{r_i^1}\right)} \tag{A.10}$$

where Avg is the arithmetic mean operator computed over the example set of four metrics, with $i = \{a, b, c, d\}$.

While all three variants of φ can be used, individually or in parallel, I propose the usage of the arithmetic function (φ_A) due to its simplicity. As there are no negative metric values, the arithmetic fidelity is reliable (i.e. not summing up negative numbers to positive ones), and if there are no measurements of 0, then the geometric and harmonic fidelities are also reliable (i.e. multiplication with 0 is avoided). Table A.4 exemplifies the calculation of the three fidelities. I also introduce another network M_2 with the measured values given in Table A.4, on which r_i^2 is computed. The difference between the two fidelities is expressed as percentage and is obtained as a ratio of the distances: $(1 - \varphi_A^1)/(1 - \varphi_A^2)$.

Having the quantified expressions for the similarities $(M,\ M_1)$ and $(M,\ M_2)$ from Table A.4, one can say that, for example, the first network topology M_1 is 3.3% more similar to the reference than the second network topology M_2 in terms of φ_A, 13.3% in terms of φ_G, and 22.3% in terms of φ_H.

Table A.4.: Example values for M, M_1 and M_2 used to demonstrate the calculation of φ_A, φ_G and φ_H.

	a	*b*	*c*	*d*	φ_A	φ_G	φ_H
M	0.3	0.7	3	50			
M_1	0.4	0.5	7	82			
r_i^1	**0.75**	**0.71**	**0.43**	**0.61**	0.62	0.61	0.59
M_2	0.5	0.2	3	90			
r_i^2	**0.6**	**0.29**	**1**	**0.56**	0.61	0.56	0.52

There is one last observation regarding the ratio r defined in Equation A.7, which is used in the composition of φ. The distance function d defined in Equation A.6 is asymmetric, and this undesired property can be observed in Figure A.2. While the distance from the reference value should grow symmetrical on both sides (towards the origin and towards infinity), the ratio r falls linear towards the origin (left) and logarithmic towards infinity (right), meaning the distance grows faster towards the origin. The example in Figure A.2 displays a metric comparison for two networks: one with a measured value $x_1 = 1$ and the second one with a measured value of $x_2 = 5$. As the reference is $x = 3$, both measurements x_1 and x_2 are equally distant, thus their r ratio should be equal. However, due to the asymmetry of the function, $r(3, 1) = 1/3 = 0.33$ and $r(3, 5) = 3/5 = 0.6$. Equation A.7 needs to be improved so that both values of r result in 0.6, offering a logarithmic decrease on both sides of the reference value.

Consequently, we introduce the notation ar for the asymmetric r function (A.7), the one depicted in Figure A.2, and sr for the symmetric function. Also, another issue is the scenario in which the reference value is $x = 0$. If we look at Equation A.7, this scenario would lead to a ratio $r(0, *) = 0$, regardless of the measurements being compared to x. To make the fidelity usable with reference values of $x = 0$, we add another branch to Equation A.7, which enables the comparison with real-world measurements that are approximated to 0. By improving on Equation A.7, sr is defined as:

$$sr_i^1 = \begin{cases} \frac{min\,(m_i, 2m_i - m_i^1)}{max\,(m_i, 2m_i - m_i^1)} & if\ m_i^1 < m_i,\ m_i > 0 \\ \frac{min\,(m_i, m_i^1)}{max\,(m_i, m_i^1)} & if\ m_i^1 \geq m_i,\ m_i > 0 \\ \frac{1}{m_i^1 + 1} & if\ m_i = 0 \end{cases} \tag{A.11}$$

The first branch of the function sr ensures that the result for all measurements which are smaller than the average have the same value as the symmetric measurements which are greater than the average. Figure A.3 displays this property of the symmetric function. The third branch shows a logarithmic convergence from $sr_i^1(0, 0) = 1$ to $sr_i^1(0, \infty) = 0$.

Using the symmetric ratio (A.11), Equations A.8-A.10 are redefined using sr instead of r. That is, the arithmetic network fidelity metric becomes:

$$\varphi_i^1 = \frac{1}{n} \sum_i sr_i^1 \tag{A.12}$$

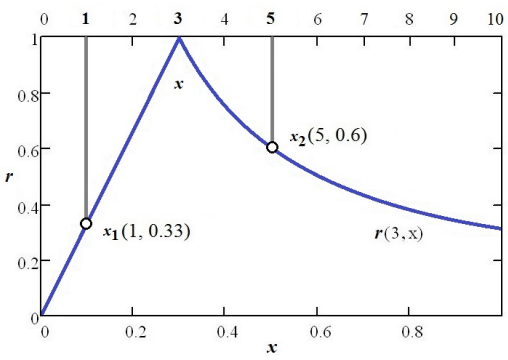


Figure A.2.: The asymmetric r function results in different values for two equally distant values, with regard to the average $x = 3$.

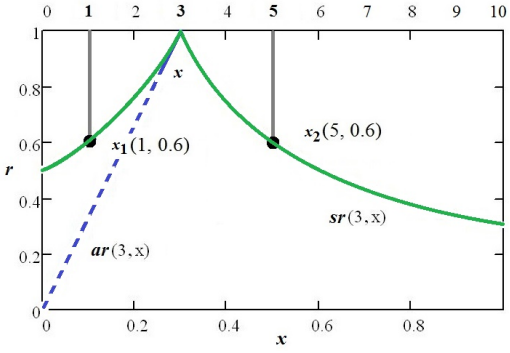


Figure A.3.: The symmetric r function renders the same values for two equally distant values, with regard to the average $x = 3$.

Table A.5.: Example values for M, M_1 and M_2 used to demonstrate the calculation of φ_A, φ_G and φ_H using the symmetric ratio r defined in Equation 13.

	a	*b*	*c*	*d*	φ_A	φ_G	φ_H
r_i^1	0.75	**0.78**	0.43	0.61	0.6425	0.626	0.608
r_i^2	0.6	**0.58**	1	0.56	0.685	0.66	0.648

which means that the fidelity of network M_1 towards network M, using the four supplied metrics $i = \{a, b, c, d\}$ and $n = 4$, is φ_i^1- the average mean of the symmetric ratios for each metric.

If I repeat the similarity measurements done in Table A.4 over the example network M, M_1 and M_2 I obtain a new conclusion which is quantified in Table A.5.

By applying the symmetric ratio formula, the two highlighted values in column b change in such a way that, compared to the results from Table A.4, network M_2 is now closer to the reference M. I can conclude that M_2 is 13.5% closer (more similar) to M than M_1 in terms of φ_A.

By analyzing the expression of sr in Equation A.11 I conclude that:

- If $m_i^1 < m_i$ then the expression $min\,(m_i,\ 2m_i - m_i^1)$ can be simplified to m_i, and the expression $maxn\,(m_i,\ 2m_i - m_i^1)$ can be simplified to $2m_i - m_i^1$.

- If $m_i^1 \geq m_i$ then the expression $min\,(m_i,\ m_i^1)$ becomes m_i and expression $max\,(m_i,\ m_i^1)$ becomes m_i^1. Using both observations, Equation A.11 can be simplified as:

$$sr_i^1 = \begin{cases} \frac{m_i}{2m_i - m_i^1} & if\ m_i^1 < m_i,\ m_i > 0 \\ \frac{m_i}{m_i^1} & if\ m_i^1 \geq m_i,\ m_i > 0 \\ \frac{1}{m_i^1 + 1} & if\ m_i = 0 \end{cases} \tag{A.13}$$

DEFINITION 1 (Network fidelity φ): Given a reference network M, and any network M_j being compared to M, the arithmetic fidelity φ_A^j which expresses the similarity of M_j towards M is defined as:

$$\varphi_A^j = \begin{cases} \frac{1}{n} \sum_{i=1}^{n} \frac{m_i}{2m_i - m_i^j} & if\ m_i^j < m_i,\ m_i = 0 \\ \frac{1}{n} \sum_{i=1}^{n} \frac{m_i}{m_i^j} & if\ m_i^j \geq m_i,\ m_i = 0 \\ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i^j + 1} & if\ m_i = 0 \end{cases} \tag{A.14}$$

where j is the index of network M_j, i is the index of the metric which describes the two networks being compared, and n is the total number of metrics used in the comparison. The geometric fidelity φ_G^j is defined in the same manner, but it uses the geometric mean operator instead of the arithmetic mean, and the harmonic fidelity φ_H^j uses the harmonic mean operator respectively.

When doing multivariate analysis it often happens that I want a weighted approach, namely, one or more metrics should be more important than others in specific contexts of interest. As Equation A.14 defines an unweighted, or equally weighted fidelity, I define the weighted fidelity metric to address this issue.

DEFINITION 2 (Weighted network fidelity ϕ): Given a reference network M, a network M_j being compared to M, and a vector $w = \{w_1, w_2, ..., w_n\}$ of weights for each common metric, the weighted arithmetic fidelity ϕ_i^j which expresses the similarity of M_j towards M is defined as:

$$\phi_A^j = \begin{cases} \sum_{i=1}^{n} \left(w_i \times \frac{m_i}{2m_i - m_i^j} \right) & if\ m_i^j < m_i,\ m_i = 0 \\ \sum_{i=1}^{n} \left(w_i \times \frac{m_i}{m_i^j} \right) & if\ m_i^j \geq m_i,\ m_i = 0 \\ \sum_{i=1}^{n} \left(w_i \times \frac{1}{m_i^j + 1} \right) & if\ m_i = 0 \end{cases} \tag{A.15}$$

$$\sum_i w_i = 1 \quad for\ i = 1, 2, \ldots, n \tag{A.16}$$

where j is the index of network M_j, i is the index of the metric which describes the two networks being compared, n is the total number of metrics used in the comparison, and w_i is the weight of metric with index i. The sum of all weights must be equal to 1. The weighted geometric and harmonic fidelities are defined identically, using their respective operators.

Choosing the actual weights is the responsibility of anyone who uses the ϕ-metric. If, for example, all metrics are considered to have an equal importance, then each weight $w_i = 1/n$, and thus I obtain the formula from Equation A.14.

A.4. Results

This section aims to validate the efficacy of φ by placing it in an applicative context. Concisely, I first use the unweighted φ to express network *realism* in a social network scenario (A.4.1), then I use the unweighted φ to express network *similarity* in a technological network setup (A.4.2).

A.4.1. Realism assessment in social networks

I apply φ in a real social network scenario, namely comparing and ordering state of the art social networks in terms of fidelity towards empirical topological models. To this end, I use the empirical data sets presented in Table A.1 as references M for measuring φ, and compare against the selected state of the art synthetic topologies. In this particular application of φ, Equation A.14 is instantiated as follows: each empirical data set is used as reference M, and its metric measurements represent m_i; each of the five synthetic networks, indexed j, is compared against M based on its topological measurements m_i^j; there are $n = 6$ metrics used for comparison, as described in Table A.1.

By using the empirical FB1, GP1 and TW1 networks as references, Table A.6 presents the resulting φ for the evaluated synthetic topologies: the scale-free (SF) model, the small-world (SW) model, the cellular (Cell) model, the static-geographic (Geo) model and the Watts-Strogatz model with degree distribution (WSDD). Table A.6 also provides the values for the statistical methods presented in Section A.2.3. By ordering the synthetic models in descending order of their φ_A values I obtain the ranking of similarity towards the empirical reference networks. To demonstrate how φ can be used as a measure of realism I have added another empirical dataset in each of the Tables A.6.1-3. Through the usage of φ_A I can make a distinction between synthetic and real-world networks. Specifically in Table A.6.1, the GP1 network is 34% more realistic than the best synthetic model (Cellular). Moreover, through comparison with the empirical network GP1, which has an inherent similarity to FB1 that is network independent (i.e. both are empirical friendship networks), only the Mahalanobis distance and the fidelity φ determine GP1 to be the most similar friendship replica of FB1. The same observation holds for Table A.6.2, where TW1 is shown as the most similar to the reference GP1. In the last scenario, only φ manages to find GP2 as the most similar to the reference TW1, while the Mahalanobis distance finds it very dissimilar.

All similarity measures are applied using six unweighted graph metrics: average degree, average path length, average clustering coefficient, modularity, diameter and density. Table A.7 highlights the best matching network in terms of single-metric comparison for each of six used metrics.

By using only φ, it can be concluded that the cellular network (Cell) is the best network fitting a real friendship network with φ_A = 0.765, while the scale-free model provides the least similarity with φ_A = 0.673. We can also conclude that the WSDD model is a good candidate for modeling friendships because its φ_A is close to the cellular fidelity, while the small-world provides less realism because it has a lower φ. Comparing the best and the worst networks using φ, it can be said that the cellular network is 39% better, or closer to the reference network FB1 than the scale-free network. Analyzing Table A.7 it is clear that it is not possible to make accurate distinctions between the networks if the comparison is made using each metric individually (i.e. using the distance Equation A.6 for all six mentioned metrics). Even though Table A.7 highlights the Cellular network as the most similar (in

Table A.6.: The fidelity metric (φ_A), cosine similarity (cos), variance (var), covariance (cov), Pearson correlation (PCC) and Mahalanobis distance (Mah) applied over state of the art networks, using each of the three empirical friendship networks as references. Unique values marked with star (*) on each column correspond to the best network as measured by the respective metric. An additional empirical social network is added to serve as a reference for comparison in each table.

A.6.1. Using the FB1 Facebook friendship network as a reference.

FB1	cos	var	cov	PCC	Mah	φ_A
SW	0.82	25.25	15.22	0.67	52.09	0.682
SF	0.80	24.44	**13.35***	0.64	59.30	0.673
Cell	**0.99***	25.83	23.67	**0.99***	**5.8***	**0.765***
Geo	0.96	**23.30***	17.56	0.93	29.67	0.717
WSDD	0.97	45.90	42.26	0.96	17.38	0.754
GP1	0.99	31.08	30.28	0.98	3.654	0.825

A.6.2. Using the GP1 Google Plus friendship network as a reference.

GP1	cos	var	cov	PCC	Mah	φ_A
SW	0.89	21.91	16.41	0.79	28.36	0.697
SF	0.87	21.0	**14.60***	0.78	33.52	0.668
Cell	0.97	22.54	20.57	0.95	**0.34***	**0.745***
Geo	**0.99***	**19.87***	16.71	**0.98***	12.60	0.718
WSDD	0.93	42.98	35.18	0.88	36.69	0.683
TW1	0.96	24.61	22.66	0.94	0.64	0.756

A.6.3. Using the TW1 Twitter friendship network as a reference.

TW1	cos	var	cov	PCC	Mah	φ_A
SW	0.75	17.94	10.01	0.56	32.33	0.559
SF	0.73	16.74	**8.65***	0.54	36.61	0.554
Cell	**0.99***	18.75	18.53	**0.99***	**1.12***	0.658
Geo	0.93	**15.63***	12.86	0.88	13.74	**0.673***
WSDD	0.99	40.34	34.0	0.98	36.27	0.562
GP2	0.94	108.16	57.59	0.92	269.9	0.712

Table A.7.: The best individual matches to the FB1 friendship network according to each separate metric.

	AD	*L*	*C*	*Mod*	*Dmt*	*Dns*
FB1	Cell	Cell	Cell	Geo	SW	WSDD

terms of *AD*, *L*, *C*) one cannot express this similarity numerically, nor can any conclusions be taken on behalf of the other networks.

The conclusions that can be drawn with regard to the competing statistical methods are:

- Variance and covariance are not suitable for quantifying similarity since their numerical values express a correlation between the magnitudes of the different metrics rather than correlation between the networks being compared. Additionally, neither of them are normalized (see Table A.6), making the numerical interpretation harder. Particularly, covariance behaves worst on all data sets.

- Cosine similarity and PCC offer identical results on these data sets, together with the Mahalanobis distance. They are both normalized values within the unit interval, but both offer different results for networks that are symmetrically distant to the reference metrics. This is one of the key features and contributions of φ, namely that it provides equal values for symmetrical networks. This aspect was detailed in Figure A.3 and Equations A.7, A.11.

- The Mahalanobis distance and φ are the only symmetric measures, but on the other hand, unlike the cosine similarity and PCC, it is an absolute, rather than a normalized distance (e.g. Euclidean distance in particular cases), thus making fair comparison harder.

The Mahalanobis distance and φ result as the best options for quantifying similarity, both being symmetric, but there is one more aspect which makes φ stand out uniquely. Table A.8 shows a synthetic example in which four networks M_{1-4} are compared with a reference M using two metrics a and b.

The first observation is that for M_1 and M_2, whose metrics vary by the same magnitude, but on a different scale (i.e. 50% variation from reference), φ gives the same results, while the Mahalanobis distance scales up by a ratio equal to the square of the variation (i.e. $\times 10^2$). Therefore, the Mahalanobis distance is not scale-free (i.e. depends on the network size) and is impossible to normalize.

Table A.8.: The fidelity metrics φ_A and φ_H, and the Mahalanobis distance measured on four networks compared to the reference M. The cells marked with star (*) are marking the counter-intuitive results. The results display the fact that φ does not depend on the network size (i.e. it is scale-free).

	a	*b*	Mah	φ_A	φ_H
M	20	200			
M_1	10	200	**70.71***	0.833	0.8
M_2	20	100	**7071***	0.833	0.8
M_3	20	199	0.707	0.997	0.997
M_4	20	$+\infty$	$+\infty$*	0.5	$\mathbf{0}_+$ *

The second observation is highlighted by the results for networks M_3 and M_4. The first one varies by a small margin, while the second one has a theoretically infinite variation for metric b. While the

Mahalanobis distance rises to $+\infty$ (i.e. complete dissimilarity), φ_A lowers to only 0.5 and not 0. The feature of φ is that it keeps a proportion of the similarity for each metric that is compared. In this example, having two metrics with equal weights, both have a proportion of 0.5 of φ. That is, the (huge) variation of one metric will not affect the impact of the others in the final value. The final observation is an argument of why I propose φ_A as the most robust of the three variants. The geometric and harmonic fidelities rely on multiplication rather than addition, a problem that is highlighted in Table A.8. The φ_H multiplies the distance ratios of the two metrics and because one of them is 0, the resulting φ will be 0.

The conclusions that can be drawn from Tables A.6 and A.7 are that single metric comparisons, variance and covariance are not well suited for measuring the similarity between social networks, while the cosine similarity, Pearson correlation, Mahalanobis distance and φ offer correct insights on similarity. However, φ is the only metric that is normalized - making scale-free comparisons possible - and symmetric - making symmetrically distant networks equally similar towards the reference.

As a consequence, I can further define realism as φ over a certain threshold θ (i.e. $0 < \theta \leq \varphi \leq 1$) in a specific context. Figure A.4 represents the numerical data measured on the presented social networks. In addition to the conclusions already discussed for each empirical network, the figure highlights the fact that, overall, Facebook friendships (FB1) are better replicated by all networks (i.e. higher φ), while the Gnutella file-sharing network (Gnu) is the least accurately replicated (lowest φ_A). Thus we can conclude that collaboration networks with the characteristics of the Gnu network are less accurately reproduced with a state of the art synthetic complex network topological model.

Fractal dimension of complex networks

Many researchers have studied the self-similarity property and dimension of complex networks [169] in search of a deeper understanding of the underlying mechanism based on the common graph metrics. As self-similarity refers to the infinite scaling property of fractals (scale-free-ness), it is considered capable of quantitatively capturing the common network structure of complex networks [175, 169]. Table A.9 shows the d_B values measured on all networks: empirical and synthetic. I have used the box counting algorithm described in [242] to compute the numerical values of the fractal dimension.

The results from Tables A.6 and A.9 offer a comparison between the five synthetic networks and each empirical dataset. As can be seen, the fractal dimension does not offer clear and consistent results regarding similarity between networks. For example, while the scale-free network is the least similar to Facebook friendships, it is the most similar with regard to Google Plus friendships, a conclusion which is conflicting with the fact that both of them are similar friendship networks.

A.4.2. Similarity assessment in technological networks

Based on the datasets introduced in Section A.2.4, I present the results of the similarity measurement in the context of technological networks.

For each city I obtain a distribution D_k^{city} , which is a vector of normalized percentages corresponding to the occurrences of each individual motif of size k. To correlate any two vectors of same

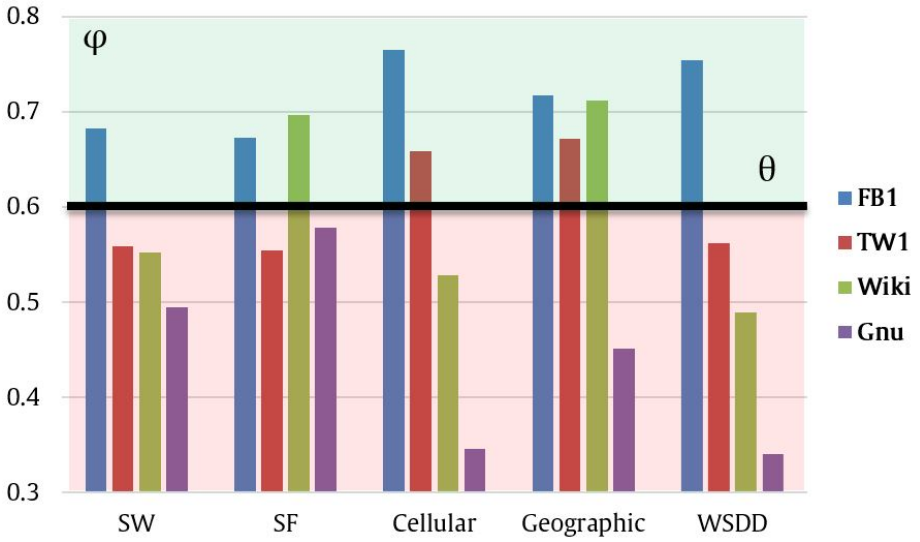


Figure A.4.: Graphical representation of the network fidelity φ_A measured for each of the five state of the art networks: small-world (SW), scale-free (SF), cellular, static-geographic and WSDD. φ is measured against the four empirical reference networks: friendships on Facebook (FB1) and Twitter (TW1), respectively collaborations on Wikipedia (Wiki) and Gnutella (Gnu). The threshold θ, at 60% similarity, divides the networks into realistically accurate ones (green upper-half) and realistically inaccurate ones (red lower-half). The value for θ chosen here is purely illustrative for this example.

Table A.9.: The fractal dimension of each synthetic network, replicated over five columns, each corresponding to an empirical reference network. Each column highlights the closest d_B value compared to the d_B values of each reference network (in column header) using a star (*) symbol.

		Empirical references				
		FB1	FB2	GP1	GP2	TW1
	d_B	2.60	3.21	4.12	5.95	3.83
Synthetic	SW	**2.65***	2.65	2.65	2.65	2.65
	SF	4.46	4.46	**4.46***	**4.46***	4.46
	Cell	2.51	2.51	2.51	2.51	2.51
	Geo	3.51	**3.51***	3.51	3.51	**3.51***
	WSDD	2.03	2.03	2.03	2.03	2.03

Table A.10.: The fidelity metric (φ_A), cosine similarity (cos), variance (var), covariance (cov), Pearson correlation (PCC) and Mahalanobis distance (Mah) applied on the motif distributions of each road network. Values marked with star (*) on each column correspond to the best network as measured by the respective metric.

A.10.1. Using Augsburg (Agb) as a reference.

Agb	cos	var	cov	PCC	Mah	φ_A
Brt	**.998**⋆	44.61	44.04	**.988**⋆	**2.83**⋆	**.804**⋆
Bud	.981	**41.14**⋆	36.16	.88	24.45	.802
Clj	.993	42.02	40.14	.955	9.27	.651
Cns	.928	51.12	**30.41**⋆	.603	101.8	.597
Tsr	.968	44.36	35.69	.806	52.59	.687

A.10.2. Using Cluj-Napoca (Clj) as a reference.

Clj	cos	var	cov	PCC	Mah	φ_A
Agb	.993	42.02	40.14	.955	**4.74**⋆	.816
Brt	.985	43.68	39.78	.912	19.24	.815
Bud	**.996**⋆	**40.17**⋆	39.01	**.971**⋆	5.68	**.869**⋆
Cns	.962	50.14	**38.87**⋆	.788	55.3	.709
Tsr	.989	43.38	40.44	.933	14.43	.854

size, for different cities, D_k^{city1} and D_k^{city2}, I make use of my proposed fidelity metric. I sequentially use each city as a similarity reference towards all other cities.

I offer an illustrative instantiation of Equation A.14 for my road networks case as follows: cities Agb and Clj are consecutively used as references M, and their specific motif measurements m_i^{city} represent m_i (in Equation A.14); each of the other five city networks, indexed j, are compared against M based on their topological measurements m_i^j; there are $n = 6$ motifs used for comparison, with measurements m_1^j to m_6^j.

Table A.10 presents the resulting fidelities for the chosen topologies, and also provides the values for the statistical methods presented in Section A.2.3. The table contains the results for comparing the empirical data in terms of motif size $k = 4$ (i.e. subgraph size). By ordering the urban topologies in descending order of their φ_A values I obtain the ranking of similarity to each reference network. Looking at the results in Table A.10.1, the φ-metric promotes the cities Brt, Bud and Tsr as the top 3 most similar to Agb based on their roads' motif distribution. The cosine, Pearson and Mahalanobis metrics give the same results in terms of the most similar city, but order the remaining cities in different manners. Variance and covariance fail at even finding the most similar city, namely Brt. Similar observations are available for Table A.10.2 in which Bud is found as the most similar city to Clj, by the cosine similarity, variance, Pearson and φ. Covariance and Mahalanobis fail to find the best match. Furthermore, the other metrics order the cities by similarity to Clj in different manners. As in the social network context, I show that φ is the superior metric for network comparison.

An important observation form Table A.10.1 is that even though Bud has approximately twice the number of nodes as Brt, and the reference Agb, φ easily finds the two networks similar to Agb (i.e.

$\varphi_A^{Brt} = .804$ and $\varphi_A^{Bud} = 0.802$). This is due to the scale-free property of φ. On the other hand, all other metrics - without exception - find a discrepancy between Brt and Bud.

By applying φ in social and technological networks I highlight that its usage is not limited to a specific network type, or to a given set of graph metrics. One can choose any type and number of metrics that are relevant to a particular study; anything from number of nodes to slope of degree distribution may be used for comparison.

A.5. Discussion

In this section I have introduced the fidelity metric φ which represents a framework for comparing complex network topologies based on common graph metrics. As such, I show the capability of φ to asses network structure *realism*, with an application in the context of comparing synthetic social networks to empirical social network topologies, as well as to asses network *similarity*, with an application in the context of technological, urban road networks. In a similar manner, any type of complex networks can be used for topological comparison. Moreover, I have presented the superior analytical power of my metric as compared to individual metric comparisons, such as the fractal dimension, and the available statistical methods like the Pearson correlation and Mahalanobis distance. Pearson correlation and the Mahalanobis distance are proven to be less efficient when compared to φ. The fidelity can be measured using any (and any number of) graph metrics, including the ones not mentioned in this work (e.g. networks size, average betweenness, slope of the power law degree distribution, criticality etc.), in an unweighted context, as well as in a weighted context.

I show that the proposed fidelity metric stands out as the most accurate and intuitive framework for assessing the similarity of complex networks, This work was published in the International Jounal of Computer Mathemtics [256]. My future work is oriented towards applying this metric on other natural and synthetic models and obtaining innovative insights.

B. Uncovering the fingerprint of online social networks using a network motif based approach

Complex networks facilitate the understanding of natural and man-made processes and are classified based on the concepts they model: biological, technological, social or semantic. The relevant subgraphs in these networks, called network motifs, are demonstrated to show core aspects of network functionality and can be used to analyze complex networks based on their topological fingerprint. I propose a novel approach of classifying social networks based on their topological aspects using motifs. As such, I define the classifiers for regular, random, small-world and scale-free topologies, and then apply this classification on empirical networks. I then show how my study brings a new perspective on differentiating between online social networks like Facebook, Twitter and Google Plus based on the distribution of network motifs over the fundamental topology classes. Characteristic patterns of motifs are obtained for each of the analyzed online networks, and are used to better explain the functional properties behind how people interact online, and to define classifiers capable of mapping any online network to a set of topological-communicational properties.

"If your experiment needs statistics, you ought to have done a better experiment."

 Ernest Rutherford

B.1. Motivation

Complex networks cover an active area of scientific research inspired largely by the empirical study of real-world networks such as communication networks, economical networks and social networks. They are classified into four major types, based on the context which they model: biological networks (e.g., metabolic networks, transcription regulatory networks, protein-protein interaction networks, protein structure networks, neural networks, ecological networks, natural food chains) [10, 276, 82], social networks (e.g. friendship networks, citation networks, voter networks, world markets, political structures) [246, 276, 214], technological networks (e.g., computer networks, electrical circuits, road networks) [10], and semantic networks (e.g. word-net [188], recipe networks [250]). Without exception, all these networks can be represented as graphs, which include a wide variety of subgraphs. One fundamental property of networks are the so called network motifs, which were introduced by Milo et. al. [189]. They represent recurrent and statistically significant subgraphs or patterns in these complex networks. The fact that motifs repeat themselves in specific networks, or even among various networks, is highly correlated with the concepts of evolutionary theory. Each of these subgraphs, defined by a particular pattern of interactions between graph nodes, may reflect a framework in which particular functions are achieved efficiently. Motifs are considered to have a notable importance today because they may reflect underlying functional properties [178]. In light of their ability to uncover structural design principles of complex networks, motifs have been slowly adopted from Systems Biology into the broader perspective of Network Science. Although they foster a deep insight into the functional abilities of a network, their detection is computationally challenging even by current standards.

Particular research has been done in the areas of biology and genetics where motifs are associated with functional roles of transcription regulation networks which control the expression of genes [11]. Experimental studies show how motifs serve as basic building blocks of transcription networks. Another example is the understanding of how some cellular components are conserved across species but others evolve rapidly [286]. A notable study brings forward this new motif-inspired paradigm to uncover drug development strategies that help in the identification of drug target candidates [69]. A similar scientific track to my proposal is presented by Wang et. al. in a study focused on detecting important nodes, not through the classic centrality metrics approach, but through specific motif patterns[275].

While conceptually (and functionally), complex networks can represent biological, technological, social or conceptual relationships between entities, I propose a motif-based analysis of networks from the topological perspective. As such, the fundamental topological families are: regular networks, random networks, small-world networks and scale-free networks [276]. Regular [56] and random networks [86] represent the basics of complex networks. The effort to mathematically express accurate and realistic models of natural phenomena (e.g. social influence, collaboration, internet communication) has been triggered by the observation of the three fundamental properties of complex networks: average path length, clustering coefficient and degree distribution [246, 276]. The well-known models of small-world [281] and scale-free [25] networks both present these network properties. Since their introduction to literature, a considerable amount of new networks have been added, yet all fall into one of the two categories: small-world or scale-free. To recreate natural processes with a higher fidelity, there are proposals which add the small-world property to scale-free models [123, 96, 166], or ones that add power-law degree distribution to the small-worlds [135, 57, 274, 296]. Thus the

motivation of this section is to provide an analytical perspective over existing state of the art complex topologies using an novel approach - classification using the network structure, namely through network motifs.

In the second part of this section, I apply this novel perspective to differentiate between online social networks. I use empirical data to demonstrate how real social networks can be classified with different levels of appurtenance to the four topological models. Even though similar in nature, it is shown in this appendix that Facebook networks, Twitter networks and Google Plus networks have very distinct topological features, as revealed by the motif-based analysis. This points out to the different features the three social platforms have in the real world.

I set out to measure the motif distributions of sizes 3 and 4 on a comprehensive database of undirected online social networks. For this, I obtain encouraging results regarding the particular patterns each of the three mentioned online platforms reveals. Their fingerprint is highly visible in terms of distribution of triadic closures, which is correlated with the clustering of nodes and short paths in the graph. The mark of triads is important as it has been shown to drive the scaling and emergence of social networks in general [145]. Also, using my approach to reveal triadic closure formation is correlated with the predictability of evolving contacts in human proximity networks [236], an important aspect of modern communication frameworks. The classifiers I obtain for each of the three online social network classes are mapped onto the four topological families, and also provide a new methodology of identifying key functional properties for new network data.

B.1.1. Research goals

In light of the general concept-driven approach to complex and social networks analysis, I propose a new perspective of looking at networks from their topological point of view. This perspective is conceptualized in Figure B.1 using the four main complex network classes: regular, random, small-world and scale-free, and is provided by in-depth network motif analysis. Thus, I bring forth the following main contributions:

- Large-scale computational generation and motif distribution analysis for the synthetic topology classes. I obtain a distinct motif pattern for each such class.
- Comprehensive motif analysis of online social networks (Facebook, Twitter, Google Plus) from which I obtain three quantifiable characteristic motif fingerprints.
- Mapping and similarity assessment of empirical networks onto topology classes, and defining a general methodology for such an approach.
- Correlation and discussion of the individual motifs that occur in each fingerprint, and an outlining of the functional properties behind the three online social platforms.

B.2. A new perspective over the related work

Comparing complex networks is aimed at a deeper understanding of the interaction patterns between these systems [281, 25, 246], and extracting their common properties helps improve the models even further [281, 13, 141]. However, the predominant method of graph metric comparison suffers from

limited information [165]. Some notable means of comparison are the distance ratio measure [49], used to compare individual mental models, a comparison from the data analysis perspective [165] and the study of the self-similarity of complex networks [242]. The network dimension is a key feature in understanding not only network topology, but also dynamical processes on networks, such as diffusion, percolation and other critical phenomena [73]. The fractal dimension d_B is proposed based on the belief that social networks are not invariant or self-similar under a length-scale transformation. Fractal dimension has been measured on multiple varied real world networks like the WWW, biological networks, actor networks, and I will use it as an alternative to the standard metric comparison.

From a topological perspective there are studies done both in the direction of classifying social network models [141] and of structural pattern detection [215]. These methods however serve a higher level of meta-analysis rather than as measures of similarity.

The work done in the field of network motifs, since their introduction [189], has seen the definition of several super-families of evolved and designed networks by the same authors [190]. They present families of complex networks grouped together by the similar significance profiles (SP) of motifs in the networks compared to the normal occurrence in random networks. These families include:

- Direct transcription interactions (in bacteria and yeast)
- Signal-transduction interactions (cell signaling, neural networks)
- Web hyperlinks and social networks
- Word-adjacency networks networks (in English, Spanish, Japanese)

With great preponderance, all studies revolve around the classification of networks - empirical or synthetic - from the conceptual point of view, into one of the mentioned four main categories. However, many of the functional properties inherent to the different classes of complex networks stem from their underlying topological features. I thus propose an alternative perspective, in which I consider any emergent complex network a mixture of the four fundamental topology classes: random, regular, small-world and scale-free. In the context of social networks, for example, it is a well know fact in literature that characteristics of each topology are present. Collaborations, sexual interactions, friendships, and citation networks are good examples of scale-free networks [200, 276]; voter networks, influence networks, food chains, and human communities are examples of small-worlds [276, 82], and they feature properties of regular organization and/or random long range links as well. My main motivation for reclassifying networks based on topology is driven by the fact that each network model can be characterized by a certain mixture of topological properties. I find out that this mixture of properties creates specific patterns over which apparently diverse networks can overlap. By applying this methodology on online social networks I bring an original contribution of how we can do social networks analysis.

The core analytical instrument with which I define the classification based on topology classes is network motifs. More specifically, given a distribution of motifs D_N over a network N one *can* classify the network into one super-family which encompasses a particular concept (e.g. social, technological), but one *cannot* associate the distribution D_N with the fundamental complex network topologies. Figure B.1 depicts the two types of classifications for complex networks. The solution to this main outline is discussed in the next section.

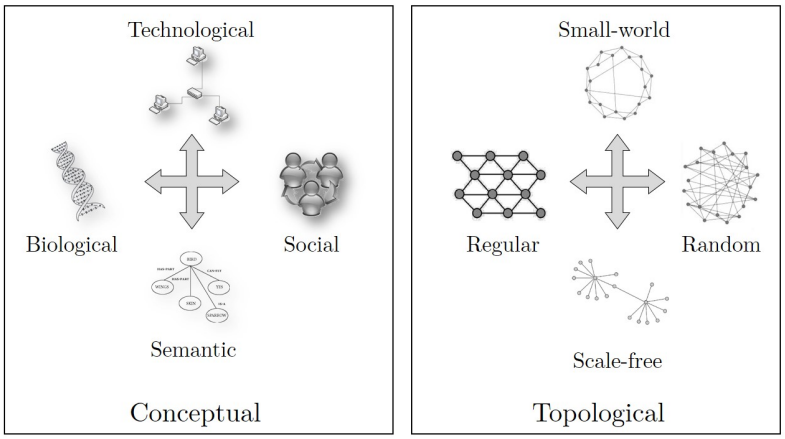


Figure B.1.: The two classifications of complex networks: the conceptual perspective versus the topological perspective.

B.3. Methodology

I propose a two step approach into classifying online social networks. First, I measure the distributions of motifs of sizes 3 (i.e. subgraphs with 3 nodes) and 4 on synthetically generated networks. I have implemented the algorithms for generating regular mesh networks, Erdős-Rényi random networks [86], Watts-Strogatz small-world networks [281], and Barabási-Albert scale-free networks [25] in Gephi [30]. Gephi is a world-leading open-source large data visualization tool built on the Netbeans framework using Java. After generating a relevant amount of such networks, ranging from 100 to 5000 nodes, with parameter values characteristic for each class, I use FANMOD to run the motif detection [284]. FANMOD is a light-weight tool for fast motif detection designed using one of the fastest detection algorithms available, RAND-ESU [283]. As depicted in Figure B.2, the first step is to find the distributions D_{reg}, D_{rnd}, D_{sw}, D_{sf} for the four corresponding topology classes.

All the generated and used networks are undirected and unweighted, since edges model mutual social ties with no additional information regarding tie strength, reciprocity etc. Many studies (from the originating fields of Medicine) rely on the analysis of motifs of size 3 in directed contexts. The upper size limit is commonly imposed due to the computational complexity of detecting larger motif structures. However, since we deal with an undirected context, the processing time is greatly reduced. For example, there are 13 different combinations of motifs of size 3 in a digraph, as depicted in Figure B.3, but only 2, respectively 6 undirected motifs of sizes 3 and 4. The codes of each motif depicted in Figure B.3 are standardized in literature, and represent the serialized binary value of the adjacency matrix (row by row) converted to a decimal value. For example, code 14 originates from the matrix 000 001 110 converted to base 10. In this section, I measure the distributions of motifs depicted in Figures B.3b and B.3c, and will refer to them using the corresponding codes.

The second step is to run the same process of detecting motifs and determining the distributions on the three chosen online social networks. I have chosen Facebook, Twitter and Google Plus as they are the most popular sites in this field [93, 172]. The empirical data is gathered from the Stanford large network dataset collection [157], and from a comprehensive private repository populated with

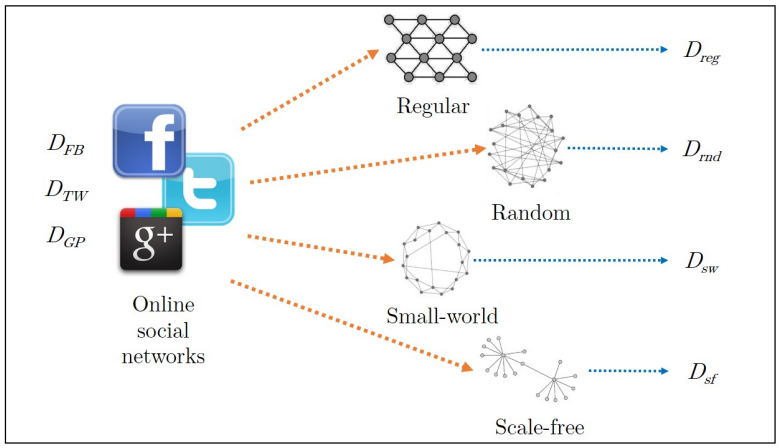


Figure B.2.: The process of classifying the three online social networks (Facebook, Twitter, Google Plus) using the four topological classes. Each motif distribution of the social networks (D_{FB}, D_{TW}, D_{GP}) is expressed as a combination of the four theoretical distributions (D_{reg}, D_{rnd}, D_{sw}, D_{sf}).

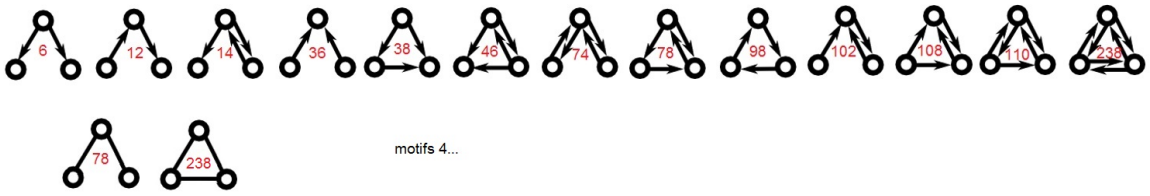


Figure B.3.: Motifs representation. **a.** All existing motifs of size 3 in a directed graph. b. The two types of motifs of size 3 in an undirected graph. **c.** All existing motifs of size 4 in an undirected graph. The code of each motif corresponds to the decimal value of its serialized adjacency matrix.

Facebook friendship graphs of students aged 19-25. The averaged results of running FANMOD on these networks yields the characteristic distributions D_{FB}, D_{TW}, D_{GP}.

To correlate the distribution vectors of the empirical datasets with each vector of the reference distributions I use the existing fidelity metric φ [257] (See Appendix A). The metric is tailored to express of similarity between any two generic vectors, in a weighted or unweighted context. In this work I use the unweighted arithmetic fidelity metric.

By measuring all similarities one can express each empirical distribution using one or more distributions of the four topological classes as:

$$D_j = \alpha_j^{reg} \times D_{reg} + \alpha_j^{rnd} \times D_{rnd} + \alpha_j^{sw} \times D_{sw} + \alpha_j^{sf} \times D_{sf} \tag{B.1}$$

where j is the index of any of the three social network distributions (i.e. FB, TW, GP; e.g. $j = FB \rightarrow D_{FB}$, α_{FB}^{reg}, α_{FB}^{rnd}, α_{FB}^{sw}, α_{FB}^{sf} etc.), or any empirical complex network in general. The coefficients α are obtained from the normalized similarities with each topology respective class. For example, α_{FB}^{reg} is the normalized similarity of the Facebook motif distribution (vector) towards the distribution found in regular networks.

The motif sizes used in this study are fixed to 3 and 4, that is subgraphs with 3/4 nodes are quantified, not larger ones. While there are approaches in literature studying network functionality using motifs of larger sizes (up to 6), I rely only on the size 3 and 4 motifs since there are few such distinct patterns, are much more numerous to be found in graphs, and thus substantially more relevant [11].

B.4. Dataset analysis

The presented motif-driven methodology requires the synthetic generation of networks pertaining to each of the four topology classes (within the characteristic parameter values), and the acquisition of friendship networks for each of the three online platforms. In this section I briefly present the parameters and settings used for generating the data, as well as the graph metrics obtained for each network class.

Even though friendship graphs vary in size significantly, from as few as 100 nodes to as many as 5000 nodes, it is a known statistic that the predominant majority of such networks revolve around the size of 300 nodes [113]. I thus generate data accordingly, and concentrate on synthetic networks within that range. Moreover, I rely on public data gathered from the Stanford large network dataset collection [157] which offers networks of hundreds up to millions of nodes. Taking into consideration the fact that graph size significantly impacts motif distributions, in order to enable a comparison at the same scale, all chosen synthetic networks are within the range of real-world ego-networks. The following datasets are used in this appendix:

- Regular: I have generated standard 2D mesh networks of sizes 200, 300 and 500.
- Random: I have generated random networks of the same sizes using the Erdős-Rényi algorithm [86], and the wiring probabilities $p_1 = 0.05$ and $p_2 = 0.1$.
- Small-world: multiple networks have been generated using the Watts-Strogatz algorithm [281], with sizes 300 and 500 nodes, wiring distance $k_1 = 2$ and $k_2 = 5$, and rewiring probability $p_1 = 0.05$ and $p_2 = 0.1$.

- Scale-free: multiple networks have been generated using the Barabási-Albert preferential attachment algorithm [25], with sizes 200, 300 and 500 nodes.
- Facebook: over 50 different friendship ego-networks have been used for metric measurements and motif analysis. Ten ego-networks are obtained from the Stanford large network dataset collection [157, 158] and have a total of 4039 nodes and 88234 edges, when combined. Furthermore, I also rely on personally gathered data using the *netvizz* Facebook application [228] with which I have obtained 50 ego-networks of sizes 150-5000 nodes.
- Twitter: using the same online repository [157], 973 Twitter circles are provided. The combined network consists of 81306 nodes and 1.7M edges. For this study, I rely on 50 chosen ego-networks, with sizes within the mentioned ranges of 200-500 nodes.
- Google Plus: I use 50 ego-networks from the same study of Leskovec et al. [158]. The combined friendship network consists of 107614 nodes and 13.7M edges. The chosen networks are all within 200-500 nodes.

Measuring the representative graph metrics over the acquired data gives conclusive results for average degree (*AD*), average path length (*L*), average clustering coefficient (*C*), modularity (*Mod*), network diameter (*Dmt*), and network density (*Dns*). Table B.1 shows the distribution of averaged topological properties on each network class.

Table B.1.: Specific values for average degree (*AD*), average path length (*L*), average clustering coefficient (*C*), modularity (*Mod*), diameter (*Dmt*), and density (*Dns*) averaged for each data set.

	AD	*L*	*C*	*Mod*	*Dmt*	*Dns*
Regular	6.63	3.34	0.065	0.05	8	0.013
Random	7.55	2.40	0.049	0.27	4	0.050
Small-world	3.99	5.61	0.321	0.73	11	0.005
Scale-free	3.12	4.60	0.015	0.62	10	0.003
Facebook	19.82	2.48	0.266	0.47	8.5	0.050
Twitter	12.39	2.68	0.239	0.28	7	0.054
Google Plus	12.15	3.90	0.404	0.44	12	0.035

B.5. Results and interpretation

Following the methodology description in Section 3, the first result is the motif distribution on the four topology classes. The distributions D_{reg}, D_{rnd}, D_{sw} and D_{sf} are depicted in Figure B.4 and, numerically, in Table B.2. Important to note is that, for each class of networks in part, I have obtained the same motif distributions regardless of network size or other specific parameters (presented in Section 4). For example, all small-worlds exhibit the same distribution D_{sw} independent of the generated network size (100-5000 nodes) and of the rewiring probability p (0.05-0.1).

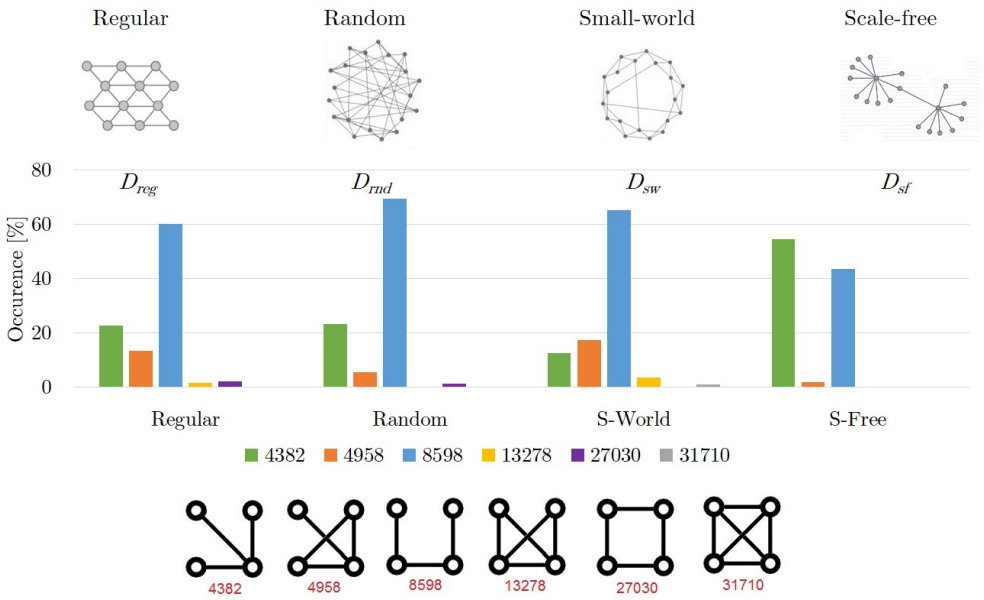


Figure B.4.: The resulting motif distributions on the regular (D_{reg}), random (D_{rnd}), small-world (D_{sw}) and scale-free (D_{sf}) topologies. The occurrence of each motif is expressed in percentage in the central histogram for each network class in part. As can be seen, only distinct motifs (not all) characterize each network class. All 6 motifs of size 4 are depicted at the bottom of the figure.

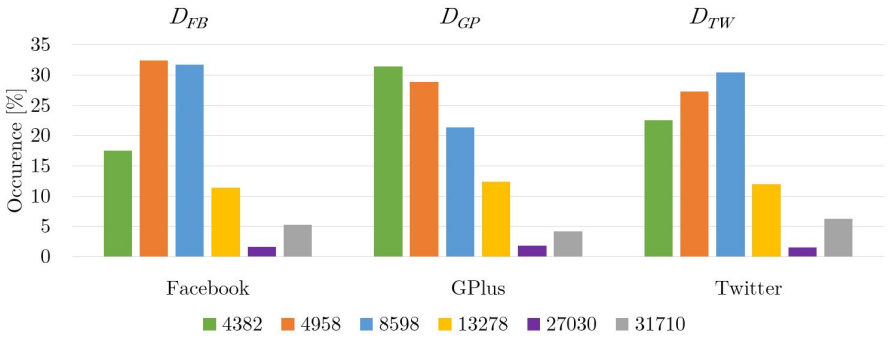


Figure B.5.: The resulting motif distributions on the online social networks: Facebook (D_{FB}), Google Plus (D_{GP}), and Twitter (D_{TW}). The occurrence of each motif is expressed in percentage. As can be seen, distinct motif patterns characterize each network class. The codes of each motif are the same as the ones used in Figure B.4.

By applying the same methodology on the empirical data, I obtain the distributions D_{FB}, D_{GP} and D_{TW}. These are depicted in Figure B.5 and also show very distinct fingerprints.

If I were to analyze the presented datasets from the conceptual perspective of social networks, there would be little to differentiate and conclude, since most online social networks serve a similar purpose. However, even at a first visual impression over Figures B.4 and B.5 it is interesting to point out how diverse the motif-based fingerprints of all 7 network types are. To facilitate the results discussion I also provide the numerical results in Table B.2

To begin with, the conclusions based on the obtained data are that each of the four topology classes has a distinct element in its motif-fingerprint. In my discussion I reference the fact whether networks favor the formation of triadic closures more, or keep triangles open. Looking at Figure B.3c, the six motifs can be divided in two categories: motifs with triads (2nd (4958), 4th (13278), and 5th (31710)) and with no triads (1st (4382), 3rd (8598), and 5th (27030)) in their structure. Triadic closures have been found to be one of the fundamental properties that give complexity and heterogeneity to social networks [145, 34]. This strongly impacts the communication through each network. By condensing the data from Table B.2 I present the occurrence of the two types of mentioned motifs in Table B.3. To ease the discussion based on each motif type I keep them highlighted in *italics* and redefine them using a short keyword as such:

- Motifs: 4382 - *star*, 8598 - *chain*, 27030 - *rectangle*.
- Motifs: 4958 - *one triad*, 13278 - *two triads*, 31710 - *four triads*.

Bridging my obtained motif distributions with the study of triads, I note that regular networks have the least characteristic mark, with a preference towards *chain* and also *star* and simple *one triad* constructs. The overall homogenous mixture reveals the fact that mesh networks keep a high local clustering (*one triad*). Overall, regular networks have 84.92% motifs that do not contain triads, and 15.08% motifs that contain them (from Table B.3), which replicates the state of the art experiments in this field [276]. Random networks have the same high occurrence of chains, and a very specific low

occurrence of *one-*, *two-* and *four triads*. Summing up the values, random networks have less than 6% triads in them, which again strengthens the known facts about low clustering in favor of a short path length. Small-worlds are a special case of empirically observed networks that lie their properties between the regular and random topologies. They favor high clustering and short path length. My analytical approach shows a fingerprint in terms of high density of *chains*, *one triads* and especially *four triads* (over 1%). Looking also at Table B.3 I notice that small-worlds are the most balanced type of topology with roughly 22% triads, and 78% no triadic formations. This balance gives them their realism in terms of replicating real social networks. Finally, scale-free networks have emerged to cover one shortcoming of small-worlds, namely the lack of preferential attachment and a power-law degree distribution, which are essential in modeling real world friendships. The scale-free network is characterized through many *chains*, but more interesting, many *stars*, and an extremely low number of *two-* and *four triads*. Added together, I can observe that there are only 1.78% motifs with triads in a scale-free network. The high occurrence of *stars* is correlated with the hub nodes with on top of the power-law degree distribution, which is specific only to this topology class.

Table B.2.: Numerical values for the distributions of the four topology classes (rows 1-4) and of the three online social networks (rows 5-7), expressed in percentages as to how often the respective size-4 motifs occur relative to the total number of recurring motifs. Each column highlights in bold the highest motif occurence for any of the four toplogy classes (1-4).

			Triads [%]			No triads [%]		
		Motif ID:	4958	13278	31710	4382	8598	27030
			one triad	*two triads*	*four triads*	*star*	*chain*	*rectangle*
1	Regular	D_{reg}	13.45	1.54	0.084	22.63	60.16	**2.14**
2	Random	D_{rnd}	5.613	0.26	0.004	23.25	**69.46**	1.41
3	S-World	D_{sw}	**17.46**	**3.51**	**1.08**	12.62	65.12	0.19
4	S-Free	D_{sf}	1.76	0.01	0.001	**54.39**	43.65	0.017
5	F-book	D_{FB}	32.44	11.41	5.25	17.49	31.75	1.66
6	GPlus	D_{GP}	28.86	12.33	4.14	31.48	21.34	1.84
7	Twitter	D_{TW}	27.33	11.94	6.23	22.50	30.43	1.53

Moving on to the empirical online social networks, I notice very distinct distributions of the six motifs (Figure B.5). Facebook friendship networks are characterized though a lower number of *stars*, but many *one triads* and *chains*. I can conclude that while there is a low tendency for hub formation (like in pure scale-free networks) the average path length is also maintained short. These remarks also coincide with the data presented in Table B.1. Google Plus one the other hand has a relatively lower number of *chains*, and a high number of *stars* and *one triads*. This network can be interpreted as one with higher clustering and and longer path lengths. Google Plus networks are known for their community (circle) based organization. Finally, Twitter networks are the most homogenous, with many *chains*, and an average-high number of *stars* and *one triads*. This fact translates into a more regular structure due to the concept of followers, which enable the creation on many random long-range links, with a disregard towards local clustering and triadic closure formation.

Table B.3.: Percentage of total motifs of size-4 that have triadic closures versus motifs that do not have any closed triangles in their structure, measured for each network type in part. The results are obtained through the condensation of the two sections in Table B.2.

	Triads [%]	No triads [%]
Regular	15.08	84.92
Random	5.88	94.12
S-World	22.06	77.94
S-Free	1.78	98.22
Facebook	49.1	50.9
Google Plus	45.33	54.67
Twitter	45.54	54.46

Taking the analysis beyond the mere topological level, I find a correlation between the characteristic graph metric values (see Table B.1) and the obtained distributions of motifs. To begin with, the prevalent occurrence of triads in the small-worlds can be explained by the higher clustering coefficient and higher modularity. These networks have a 15-1000 times higher concentration of *four triads* than all other topology classes. The low concentration of *stars* comes to support the lack of a power-law degree distribution. The small-world effect is mapped in the real-world network through the stronger community structure of Facebook and Google Plus networks. On the other hand, the lack of triads found in scale-free networks is a result of the power-law degree distribution. Its low clustering and relatively higher average path length are explained though the lower occurrence of *chains* and *rectangles*. The very low modularity of regular networks is correlated with the very high occurrence of *rectangles*, which suppress the formation of clear, distinguishable communities. On of the goals of social networks analysis is to create better generative models for real-world networks, thus my motif distribution - graph metric correlation may help improve the generation of specific synthetic networks. Based on these results, there are heuristic algorithms which can used to create synthetic networks with the required metric distributions. [216, 260].

To enhance the visual differentiation and similarity between the obtained motif patterns I provide a radar chart overview in Figure B.6. Notable in Figure B.6a are the higher occurrence of *stars* in scale-free networks, the low preference towards triads of the scale-free and random networks. In Figure B.6b I notice a good overlap between Facebook and Twitter networks, with high occurrences of *chains* and *one triads*, while Google Plus favors more *star* formations.

In order to further validate the insightful perspective revealed by motifs of size 4, I reapply the same methodology using motifs of size 3. In support of my claims, I briefly mention that the are only two types of motifs of size 3 in an undirected context. These can be seen in Figure B.3b and I will refer to them as *chain* (78) and *triangle* (238). Table B.4 contains the distribution data for each of the seven networks.

Even though motifs of size 3 have significantly less structural complexity compared to size 4, they do reveal and sustain out previous claims. Scale-free networks consistently favor open triangles to closed ones, with roughly 0.5% *triangles* in their structure. Small-worlds present the same balance between *chains* and *triangles* like in Table B.3. Regular networks have a notably higher occurrence

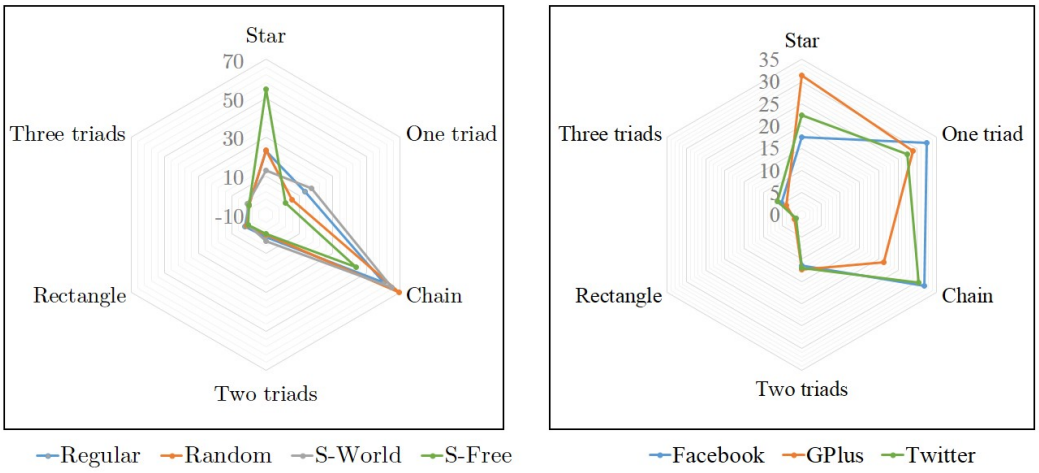


Figure B.6.: Radar chart showing the 2-dimensional distribution of motifs of size 4 for the topology classes **(a)** and the online social networks **(b)**.

of *triangles*, and random networks of *chains*, in conformity with previous claims. Finally, Facebook, Google Plus and Twitter networks share similar distributions of *chains* and *triangles*. I note that motifs of size 3 are insufficient to assess undirected friendship graphs.

For a final overview, I apply the φ-metric on the distribution vectors of motifs of size 4 and obtain the numerical data shown in Table B.5. A value of 1 means complete similarity, while a value of 0 means complete dissimilarity. The percentages of the fidelity are normalized into n-values which, summed on each column, add up to 1. The data is interpreted as, for example, Facebook can be mapped 26.6% over regular, 25.7% over random, 25.7% over small-world, and 21.9% over scale-free networks.

In interpreting the obtained fidelity results, I have to keep in mind the fact that the overall open- versus closed-triangles ratios are very similar. Specifically, this is displayed in the lower halves of Tables B.3 and B.4. Thus, the variations in terms of φ are small, but they map to significant structural differences [257]. Figure B.7a shows the 2-dimensional similarity mapping between the online social networks and the four topologies and Figure B.7b shows how much each topology contributes, in total, to the mapping of the three online social networks.

The fact that the highest overall occurrence is that of the regular topology, and the lowest, that of the scale-free topology, denotes an important real-world aspect of social networks: the formation of hubs is a rather exceptionally rare event, seemingly random long range links tend to form much more often, and the fundamental structure of social networks is based on mesh networks with a tendency towards local clustering. This observation sustains the fact that geographical proximity is indeed the main drive for friendships creation in society [58, 273]. Furthermore, the predominantly high occurrence of *chains* in all topology classes seems to be a natural facilitator of new friendships creation. A new study proves that new friendships are preferentially created between nodes located

Table B.4.: Numerical values for the distributions of the four topology classes and of the three online social networks, expressed in percentages as to how often the respective size-3 motifs occur relative to the total number of recurring motifs.

	Motif ID:	78 *chain*	228 *triangle*
Regular	D_{reg}	93.22	6.78
Random	D_{rnd}	97.37	2.63
S-World	D_{sw}	84.31	15.69
S-Free	D_{sf}	99.49	0.51
Facebook	D_{FB}	72.58	22.42
Google Plus	D_{GP}	76.87	23.13
Twitter	D_{TW}	75.28	24.72

Table B.5.: Similarity between the empirical network models and each topology class. The similarity is measured by applying the φ-metric on the distribution vectors as described in Equation 1. The columns labeled n display the normalized values for the obtained similarities, according to Equation 2. The sum of n-s is equal to 1 (100%) on each column.

	Facebook		Google Plus		Twitter	
	φ_{FB}	n	φ_{GP}	n	φ_{TW}	n
Regular	0.62	*.266*	0.61	*.269*	0.65	*.267*
Random	0.60	*.257*	0.58	*.255*	0.65	*.267*
Small-world	0.60	*.257*	0.56	*.247*	0.59	*.243*
Scale-free	0.51	*.219*	0.52	*.229*	0.53	*.219*

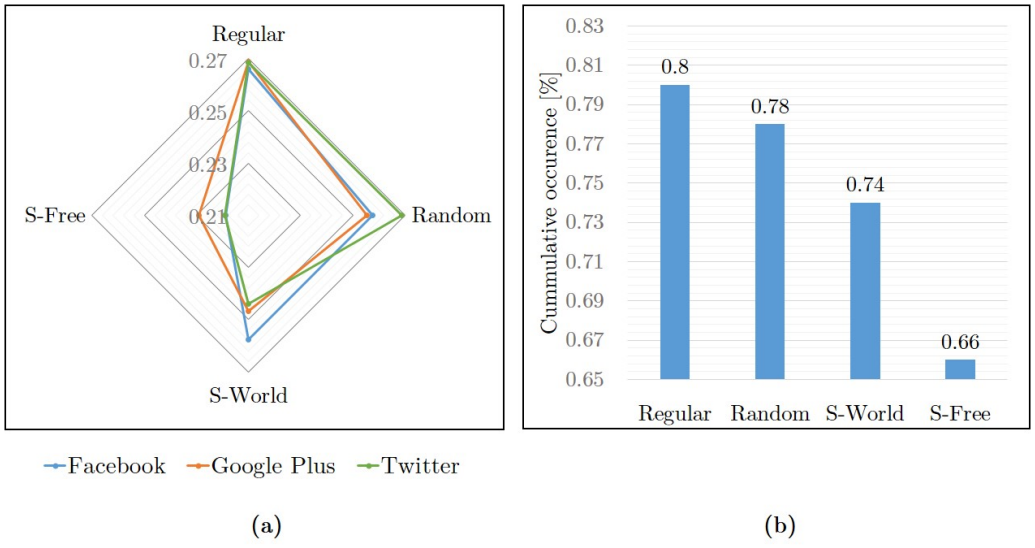


Figure B.7.: **a.** Radar chart showing the 2-dimensional mapping of the online social networks over the four topology classes. The mapping is done using the fidelity metric φ to assess the similarities based on the distribution of size 4 motifs. **b.** The cumulative occurrence of each topology class obtained by adding the normalized fidelities (n) on each row (from Table B.5). It shows how much each topology contributes overall to the three empirical networks.

at geodesic distances 2 and 3 in the social graph [65]. This conclusion strongly supports my results regarding *chains* which become natural pathways of length 3 between unconnected nodes.

To better interpret the similarity results I corroborate the results in Table B.1 with measurements of variance of the normalized fidelities (n) and conclude upon the following:

- Google Plus networks have the lowest variance (2.77-e4) showing a greater topological homogeneity. They have higher scale-free and regular appurtenances, which translates into a higher average path length (L) and a strong community structure (*Mod*). Empirically and intuitively, I explain this through the *circle* concept introduced by Google. Circles tend to offer better socializing within clusters of friends but they also limit external contacts. As most friendship clusters follow a normal distribution of contacts (degrees), the resulting model is classified as a *regular topology with preferential community formations*.

- Twitter networks have the highest variance (5.63-e4) presenting the highest topological heterogeneity. They have more notable random and regular characteristics, which translates into a very short average path length (L) and a weak community structure (*Mod*). Intuitively, I explain this through the *follower* concept specific to the Twitter online platform. The act of following tends to omit local clusters formation, or be in any way linked to geographical proximity. On the other hand, many users follow distant celebrities and/or users with same interests that are evenly spread across the globe. Uncharacteristic for real tie formations, Twitter is classified as a *heterogeneous regular topology with random long range links*.

- Facebook networks have a variance situated between the other two (4.47-e4), presenting a good mixture of all topology types. Nonetheless, they have higher small-world and regular properties, which translates into a short average path length (L) and a strong community structure (*Mod*). Based on these observations, one could say that they lie between Twitter and Google Plus. Intuitively, but also backed up by other relevant research, Facebook friendships are considered the best replica and substitute for real-world friendships [125]. This idea is further supported by the fact that their fidelity distribution also coincides with the overall fidelity distribution depicted in Figure B.7b. The stronger community structure, but with low average path lengths, seems to be a natural emerging property of the society, modeled through the *friend*-ing concept on Facebook. In fact, this seems to narrow down the distances between communities until they start overlapping. With a very characteristic real tie formation process, I classify Facebook as a *regular topology with interspersed small-worlds*.

B.6. Discussion

This section has shown that studying complex networks from a topological perspective, though the insight offered by network motifs, is a new fundamental approach in understanding the emergence of social networks. Indeed, motifs highlight functional aspects of the driving forces behind online social network creation, ties formation, community emergence, and overall communication trends. My comprehensive social networks analysis, based on graph metric and fidelity assessments, has found a predisposition for characteristics of regular networks (geo-proximity drives tie formation), followed closely by random network aspects (long range link formation), then, with diminishing predisposition, by small-world properties (tendency to cluster and close triads), and, with very low

occurrence, characteristics of scale-free networks (hub formation). Finally, I have shown that each online social platform has quite distinct properties, which imply distinct motif fingerprints, and thus different communication mechanisms.

Based on my observations, and stemming from motif analysis, Facebook, Google Plus, and Twitter networks are not similar at all when it comes to mapping them over the fundamental topology classes. Each presented characteristic defines a different approach to dealing with processes like network growth, new tie formation, community formation, information diffusion and triadic closures. I believe my work will pave the way for a better understanding of the secrets that lie behind modeling and understanding dynamics in our societies.

C. A complex network approach to patient phenotyping

This appendix presents studies undertaken in medical science using to so-called network medicine paradigm, introduced in science in the last decade. I have modeled biological networks of patients with sleep disorders, respectively cardiovascular disorders. The graph based modeling consists of specific anthropometric risks which are associated with sleep apnea and heart disease. Through the means of graph based modeling and analysis I am able to offer better vision into specific phenotypes of patients, that would not have been detectable otherwise (i.e. through classic statistical methods). In this appendix I present the essence of both direction in which I have applied complex networks concepts to model patient networks and improve diagnosis accuracy in medical science,

"Declare the past, diagnose the present, foretell the future."

 Hippocrates

One important branch of network science, which has brought cutting edge contributions to medical science, is the field of network medicine. Along this direction, I am, and have been for over 3 years, a member of the research consortium between the ACSA team from the Department of Computer and Software Engineering, of the Politehnica University Timisoara, and the Department of Pneumology of the Victor Babes University of Medicine and Pharmacy. The team is led by by my PhD co-advisor Assoc. Prof. Mihai Udrescu together with Assoc. Prof. Dan Mihaicuta, who is an expert in sleep medicine. As a team, we have studied sleep apnea from a network science point of view, and are currently in a the *Realfund* project sponsored by Linde.

My proposed study sets out to identify specific patterns of developing obstructive sleep apnea (OSA), by taking into consideration the multiple connections between risk factors in a relevant population of patients. For this purpose, I create a social network of patients based on their common medical conditions and obtain a community-based society which pinpoints to specific - and previously uncharted - patterns of developing OSA. Eventually, this insight should create incentives for predicting the apnea stage for any new patient by evaluating its network topological position.

C.1. Obstructive sleep apnea

Sleep apnea is a disorder which consists of abnormal breathing pauses, irregular or superficial breathing that occurs during sleep [240]. It has often been indicated as a serious, frequent but mostly underrated clinical problem [239]. The reported incidence of apnea varies, due to different backgrounds of patient groups that were taken into account. Nonetheless, in [293] it was reported that there are 70 million people in USA with obstructive sleep apnea (OSA), so that 1 in 4 men and 1 in 10 women have developed this disorder. Other studies have reported estimates of 3% to 7% prevalence of OSA [224]. Obviously, these are high figures which indicate the magnitude of the situation; hence it comes as no surprise that the occurrence of sleep apnea is referred by many as epidemic [293, 224, 185]. The morbidity risks entailed by the fact that many sleep apnea cases are not discovered and treated in time are well documented by many comprehensive studies. Maybe the best known link is between sleep apnea and cardiovascular problems [183, 232, 266, 235], leading to hypertension, stroke and even death [289].

Also, there are studies that associate sleep apnea with obesity [43], the risk of developing diabetes mellitus [17], and even cancer [21]. These extreme risks were recently linked to the so-called very severe sleep apnea (i.e. AHI >= 60) [140]. This is one of the reasons which indicate the paramount importance of early diagnosis of these severe cases [218].

Apart from these problems, another perspective has recently been reported: the fact that OSA is responsible for serious perioperative risks which will put a supplementary strain on surgery procedure costs [185]. The extent of this risk is extremely worrying, as it was estimated that 80% of the patients in USA have undiagnosed OSA at the time of surgery [185]. The economic implications of all these risks are significant and must be dealt with, by adopting appropriate procedures for patient management [140, 213] which, in turn, are underpinned by methods for early detection of moderate and severe apnea cases. Taking a step forward, reference [185] argues that the available traditional methods for identifying patients at risk (based on randomized trials) are not efficient enough, and that new, innovative and ultimately better ways are required.

This project proposes such an innovative approach, based on identifying specific patterns of developing apnea by taking into consideration the multiple connections between risk factors in a relevant

population of patients. The main idea is based on the assumption that there is a connection between the way of acquiring apnea and the severity of this disease. Thus, by using tools that were put forward by the new network science [202, 276, 24] which spurred cutting-edge research in the field of network medicine [27, 170], this work proposes a methodology of associating apnea risk groups to each such apnea pattern. Eventually, this insight creates incentives for predicting the apnea severity for any new patient, by evaluating its network topological position, based only on simple clinical aspects such as sex, neck circumference, obesity, etc. This allows for introducing an easy-to-use scorecard that will accurately indicate the risk group that the potential patient pertains to.

C.1.1. Data acquisition

Part of the so-called *Morpheus* team in the Linde sponsored project, I processed data from our apnea database at Timisoara Pneumology Clinic, in Timisoara, Romania, gathered from March the 1st 2001 to March the 31st 2011. The cardiorespiratory polysomnography was performed using Poly-Mesam 4 1998, Alice 5 Respironics 2005, and Stardust Respironics 2005 devices.

The population group is represented by all consecutive persons referred to be evaluated in the sleep laboratory with suspicion of sleep breathing disorders (non-probability sample, convenience sample). Patients are not randomized but the statistical analysis shows that the study population is representative (stratified sample) with adequately represented subgroups with different severities, comorbidities and responses to therapies. Excluded patients have the same characteristics as the study group. 5,103 patients (1,426 females, mean±SD age 51.8±12.6 yrs, 79.4% with apnea/hypopnoea index (AHI) $\geq$15 events/h) were processed from March 15, 2007 to August 1, 2009. Morbid obesity (body mass index $\geq$35 kg/m^2) was present in 21.1% of males and 28.6% of females. Cardiovascular, metabolic and pulmonary comorbidities were frequent (49.1%, 32.9% and 14.2%, respectively). Patients investigated with a polygraphic method had a lower AHI than those undergoing polysomnography (23.2±23.5 versus 29.1±26.3 events/h, p<0.0001).

Conduct of the study

The study was conducted in accordance with the New England Journal of Medicine protocol (available at NEJM.org). The protocol was approved by the local ethics committee. The oral consent was obtained from each patient, as well as from their M.D.s who sent them for further investigations at TPC (Timisoara Pneumology Clinic). We used standard and non-invasive, effortless procedures only, which therefore did not require any kind of compensation or supplementary costs. Their identities and personal data were not used, thus assuring the complete confidentiality of our study. The authors vouch for the completeness and veracity of the reported work as well as the fidelity of the reported work to the protocol.

C.1.2. Network approach

A final database of 1367 consecutive patients from the sleep lab of Timisoara "Victor Babes" Hospital, with over 100 measured criteria, is used as input for a methodology inspired by the Network Medicine approach [27, 170]. Each patient is considered a node in a network where the link between two nodes is inserted if there is a risk compatibility relationship between the two corresponding patients. The risk compatibility exists if the two nodes have at least 5 out of 7 identical parameters: *sex* (male or

female), *age* (group 0: $\leq$20 yrs; group 1: 20-40 yrs; group 2: 40-60 yrs; group 3: >60 yrs), *blood pressure* (with hypertension; without hypertension), *obesity* (not obese: BMI$\leq$30; obese: BMI>30), *neck circumference* (normal circumference: <40cm for women, <43cm for men; large circumference), *mean-* and *desaturation index*. The graphical representation is generated with Gephi 0.8.1 [30], in order to extract the most important network attributes, as well as revealing the compatibility clusters. A compatibility cluster uniquely defines the specific pattern for acquiring apnea. I chose Gephi as it is the leading tool in large dataset visualization, and because it is open source, allowing us to create custom tools on top of the graph processing framework. In Figure C.1 I use the AHI parameter to classify the four stages of apnea:

- group 0: $0 \leq$ AHI < 5 (low risk)
- group 1: $5 \leq$ AHI < 15 (moderate risk)
- group 2: $15 \leq$AHI < 30 (high risk)
- group 3: $30 \leq$ AHI (very high risk)

The algorithm for mapping the patient database onto the apnea risk clusters is detailed below. The construction of the graph has a time complexity of O(n^2). The relevant health parameters of each two (distinct) patients are compared, and the compatibility degree is increased on each match. The functions displayed on lines 3-9 return discrete mappings of each parameter as previously described. The condition on line 10 acts as an edge weight filter, i.e. only edges with weight ≥ 5 out of 7 are actually added to the graph. The resulting graph is discarded of weights.

```
Patient database to complex network G:
1 :for each pair of patients (p_i, p_j):
2 :  compat = 0
3 :  if gender(p_i) = gender(p_j) then compat++
4 :  if ageGroup(p_i) = ageGroup(p_j) then compat++
5 :  if hypertension(p_i) = hypertension(p_j) then compat++
6 :  if obesity(p_i) = obesity(p_j) then compat++
7 :  if neck(p_i, gender(p_i)) = neck(p_j, gender(p_j)) then compat++
8 :  if meanDesatGroup(p_i) = meanDesatGroup(p_j) then compat++
9 :  if desatIndexGroup(p_i) = desatIndexGroup(p_j) then compat++
10:  if compat ≥5 then G.addEdge(p_i, p_j)
```

Once the patient graph is obtained, the nodes have to be classified into compatibility clusters. As such, I extract the modularity, a graph measure which is designed to measure the strength of division of a network into communities [205]. This is achieved using the modularity algorithm [38], with a corresponding default resolution of 1.0 [151]. Further, for each obtained community independently, I inventory the number of occurrences of each risk of each patient. Finally, the risks with a high occurrence per community are marked as significant for the risk cluster. A risk is considered significant if it is present in $\geq$75% of patients from that community.

The reason for choosing the limit of 4 out of 6 as the threshold t for adding edges to the graph was made empirically. The goal of the study was to clusterize the patient database based on ad-hoc properties, in such a way that the visualization is relevant for human understanding and clinical processing: not too many, nor too few clusters were required. Thus, Figure C.2 depicts the visualizations

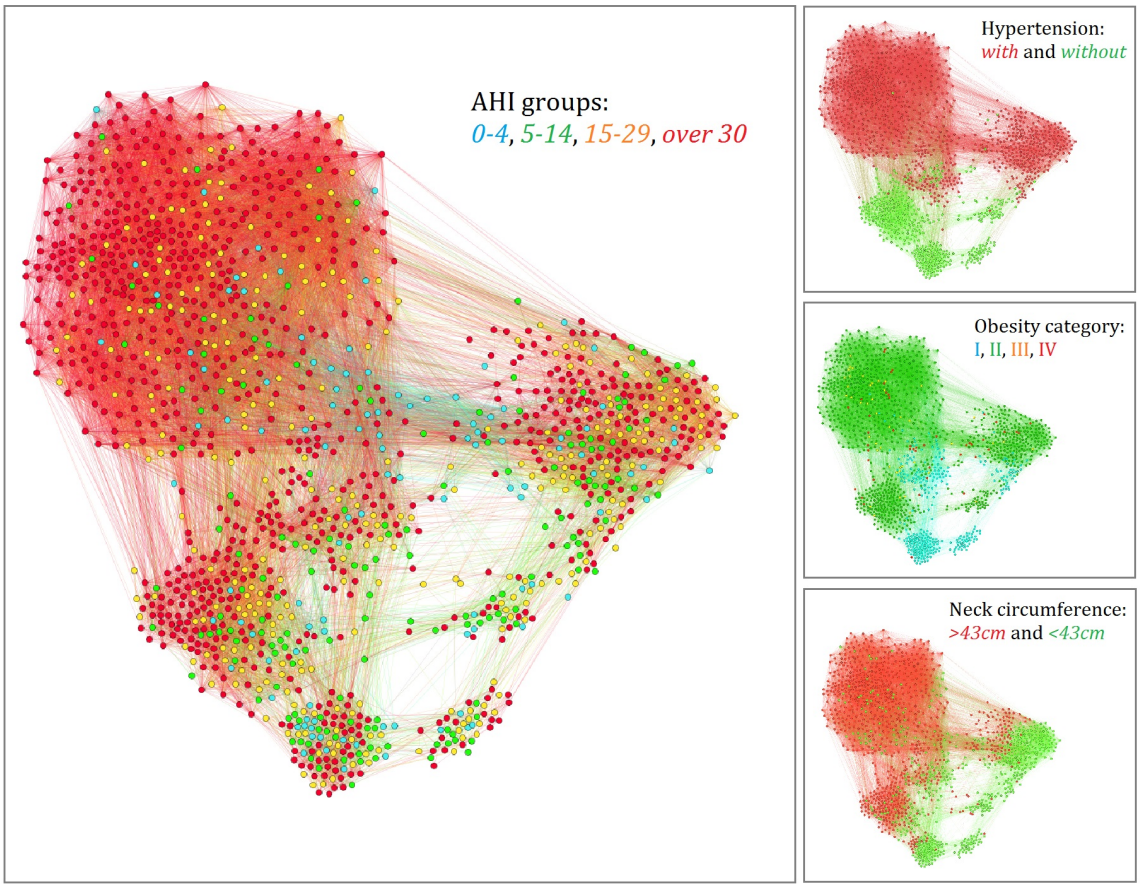
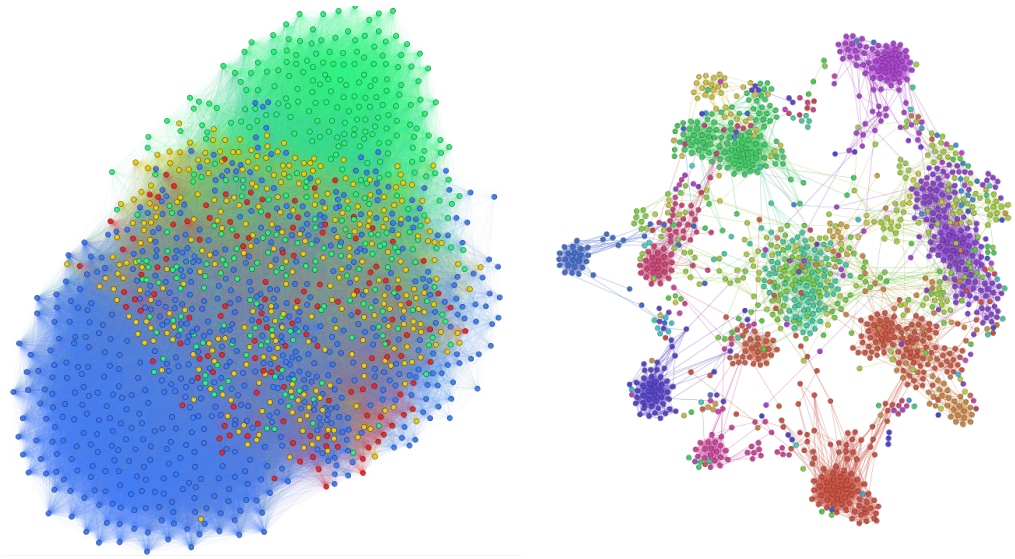


Figure C.1.: Graphical representation of the patient population with clinical apnea signs: node colors are assigned in order to depict, as indicated: AHI groups, hypertension, obesity and neck circumference.



(a) (b)

Figure C.2.: The visualization of the patient graph with a threshold of: **a.** 4 out of 7 (4 communities, too dense), **b.** 6 out of 7 (162 communities, too sparse). The node color is according to the assigned community.

of the resulting graphs for a threshold t=4, and t=6 out of 7, while the ideal clustering for $t = 5$ is depicted in Figure C.3. A lighter condition (t<5) results in too many edges being added and the community structure disappears. A stronger condition (t>5) results in too few edges added, thus too many small communities. The threshold t=5 offers the best results, but it is no rule of thumb for other datasets. This aspect has to be tested empirically and adjusted accordingly for other studies.

In the current setting, choosing different values for t gives the following number of communities: t=1 $\rightarrow$ 2 communities, t=2 $\rightarrow$ 3 communities, t=3 $\rightarrow$ 3 communities, t=4 $\rightarrow$ 4 communities, t=5 $\rightarrow$ 7 communities, t=6 $\rightarrow$ 162 communities, t=7 $\rightarrow$ 1051 communities.

Using the complex network cluster analysis [293, 224] that I provide, 7 distinct compatibility clusters were found. Each of these clusters corresponds to a specific patient profile which leads to a certain probability of developing the disease, as shown in Figure C.3. There are 3 clusters of patients with severe apnea (1, 2 and 5) and 3 clusters which generally do not have severe apnea (3, 6, and 7). Cluster 4 is special in that it seems to reflect a transitory stage between the clusters indicating severe illness and the others; moreover, it shows that it cannot be characterized by a stable AHI group: ~47% are group 3, 26% group 2, ~17% group 1, and 7% pertain to group 0.

Each obtained community holds a particular distribution of the 7 parameters, however some of them are statistically relevant to a cluster - and thus representative - and some are not relevant. In order to extract the most relevant features for each cluster I have developed a data mining tool in Gephi which extracts these relevant features. I consider these results to pave the way for defining the characteristic patterns of each cluster of patients to develop sleep apnea. Also, each cluster can be correlated with a type of prevention and treatment scheme. The characteristic features for each

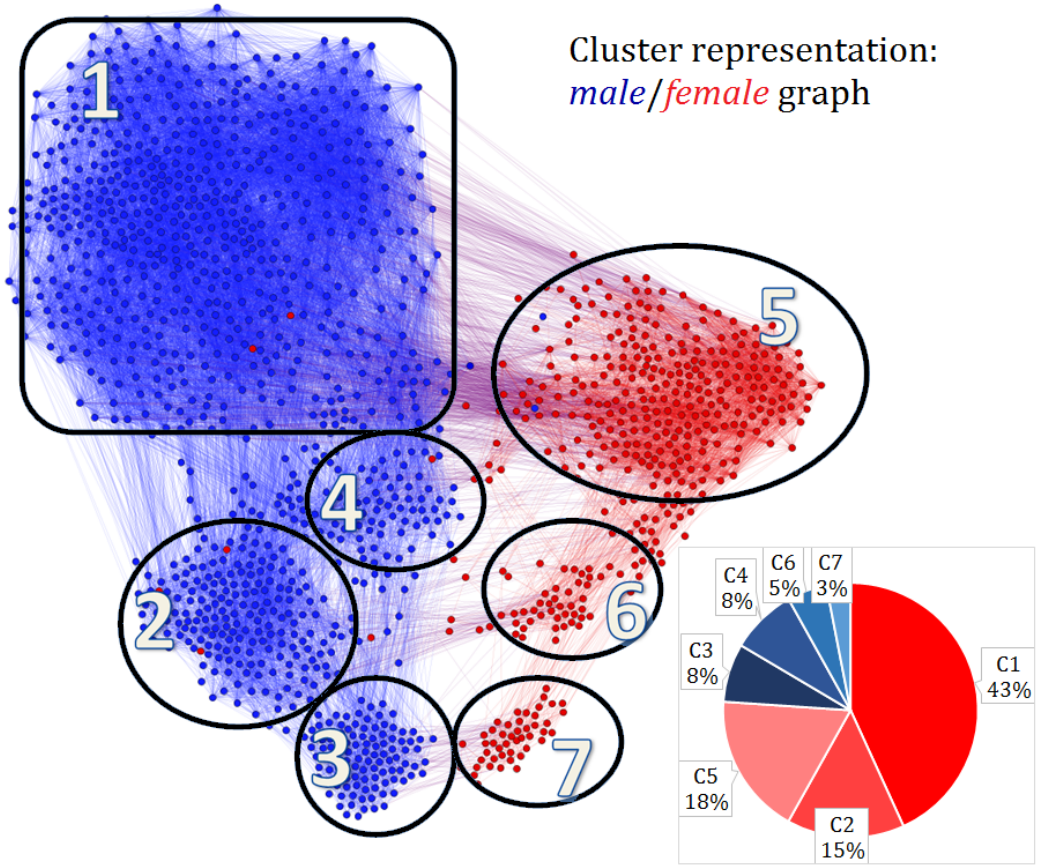


Figure C.3.: Graphical representation of clustering in the patient population with clinical apnea signs. Lower-right corner: the population distribution over the 7 clusters. Red depicts very severely sick patients, blue depicts clusters with moderate to high severity of OSA.

obtained cluster are:

- cluster 1: male, obese, with hypertension, neck circumference >43cm, desaturation<90 (73% have AHI>30);
- cluster 2: male, obese, no hypertension, neck circumference >43cm, desaturation 90-95 (55% have AHI>30);
- cluster 3: male, no hypertension, neck circumference <43cm, desaturation>95 (39% have AHI>30);
- cluster 4: male, not obese, all other risk factors are variable (56% have AHI>30);
- cluster 5: female, obese, with hypertension, neck circumference <43cm, desaturation 93-97, (52% have AHI>30);
- cluster 6: female, obese, without hypertension, neck circumference <43cm, normal desaturation 93-97, (33% have AHI>30);
- cluster 7: female, not obese, no hypertension, neck circumference <43cm, desaturation >95, (39% have AHI>30).

In line with the goal of this study, namely to facilitate the efficiency of diagnosing OSA, I have developed the apnea risk matrix based on the cluster analysis. Using the apnea risk matrix a doctor may identify patients with a possible risk of apnea through a survey, without the need for an initial specialized control. Once the measurements are done on a patient the doctor will use the matrix in Table C.1 to fit the results into one of the seven clusters (columns). When the cluster corresponding to the patient is found, the doctor will use the apnea risk percentages to formulate the diagnosis. If the patient's risk towards apnea results as statistically high, then specialized control is recommended.

C.1.3. Improving the network model: from AER score to SAS score

The aim of the project is to accelerate and simplify the diagnosis of OSA in the general population. As such, I propose to quantitatively predict the risk of obstructive sleep apnea (OSA) without measuring the patient's AHI, based solely on the anthropometric risk factors in a relevant population, validated with our network-based methodology [265].

The *AER* score predictor emerges from the statistical analysis of the obtained clusters and helps easily and rapidly assess the risk of OSA of a new patient. Using it to prioritize patient treatment/evaluation I manage to improve to overall process efficacy by 53%, in terms of cumulative AHI diagnosed, as compared to the first-come, first-served (non-prioritized) method currently used. Looking at Figure C.4 we see the difference between a randomly ordered first-come-first-served patient queue (unordered) and an ideal but impossible scenario of sick-first patient queue (optimal). Figure C.4 shows the accumulated diagnosed AHI, which is, of course, much faster in case of the optimal ordering. However, in between, there is the scenario made possible by the *AER* score, which is 53% closer to the optimal ordering than the random one.

After over one year of testing the *AER* score, we have considered to drop the desaturation measurements (which are harder to obtain in a non-specialized doctor's office), and have included the

Table C.1.: The apnea risk matrix is a table which facilitates a statistical diagnosis of apnea patients. It is based on the following simple measurable criteria: gender (M/F), hypertension (0/1), obesity (0/1), neck (0/1 = ≥43cm men, ≥40cm women), mean desaturation (0-100%). Each of the seven resulting clusters can be described by the set of characteristic features represented in the table. We mark with 'x' the fact that a criteria can take both values (i.e. is irrelevant). In the lower part of the table we represent the apnea risk probability that a patient included in either one of the clusters will have. These values result from the analysis of the database with over 1300 patients.

Gender	M	M	M	M	F	F	F
Hypertension	1	0	0	X	1	0	0
Obesity	1	1	0	0	1	X	0
Neck	1	1	0	X	X	X	0
Desaturation	<90	90-95	>95	X	93-97	93-97	>95
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Normal (0-4)	5%	6%	15%	7%	10%	10%	9%
Moderate (5-14)	6%	14%	22%	14%	13%	28%	22%
High (15-29)	16%	25%	24%	23%	25%	29%	30%
Very high (≥30)	73%	55%	39%	56%	52%	33%	39%

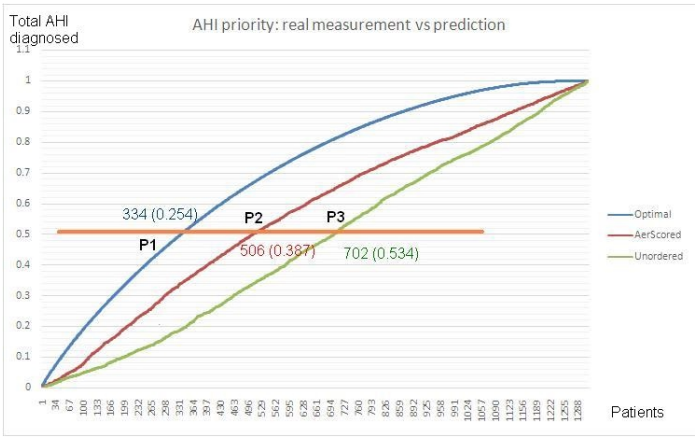


Figure C.4.: Cumulative AHI diagnosed on the dataset of 1367 patients. The unordered scenario considers that random patients are assessed one a a time; the optimal scenario assumes that we first diagnose the most "sick" patients in terms of AHI, so the accumulation of total AHI is faster; the *AER* score scenario is the one made possible with the prediction offered by our score.

Epworth sleepiness score. This score is highly relevant to OSA, and is easy to measure using a questionnaire. As such, the new score is called *SAS* score (sleep apnea syndrome score) and the new clusters which emerge are depicted in Figure C.5. Currently, we use this score as a state of the art method for assessing patient risk towards OSA. A patient may be a member of one of the 8 detected communities with their respective risks towards specific commorbidities and health complications.

The clustering is done in a similar manner, but instead of desaturation index and mean desaturation, the Epworth sleepiness score is used. This also reduces the maximum patient compatibility from 7 (maximum common metrics) to 6. The resulting graph and its communities are based on 4 out of 6 filtering, as described above.

As part of the Morpheus team, I am still highly active in this research, as it represented the main side project which I undertook during my PhD. I have extensively used the know-hows acquired from social networks analysis, and vice-versa, I have used the modeling know-how in the other topics related to my doctoral goals.

C.2. Evaluation of patients diagnosed with arterial hypertension through network analysis

Along the same direction of using network medicine, my proposed study [247] sets out to identify specific patterns of treatment response to arterial hypertension, by taking into consideration the multiple connections between risk factors in a relevant population of hypertensive patients. For this purpose, I create a network of hypertensive patients based on their common medical conditions and obtain a community-based society which pinpoints to specific - and previously uncharted - patterns of developing hypertension. Patients are nodes in a network and are linked whether they have a degree of risk factor compatibility greater than an imposed threshold. Distinct communities are detected from the emerging graph, and are used to describe different patient profiles. Eventually, this insight should create incentives for predicting the treatment efficacy for any new patient by evaluating its network topological position.

The study includes 289 patients diagnosed with essential arterial hypertension. They were monitored and evaluated at the Cardiology Clinic from the Municipal Hospital in Timisoara, Romania. The patients were randomized in three groups, according to the type of medication administered: group A, comprising 106 patients treated with a beta blocker (nebivolol), group B including 104 patients treated with an inhibitor of angiotensin conversion enzyme (perindopril), and group C consisting of 79 patients who had an inhibitor of the receptor AT1 of angiotensin II (candesartan cilexetil) in the therapeutic regimen. From the study were excluded patients having the following diagnoses: abnormal heart rhythm (atrial fibrillation), an ejection fraction (EF) below 50%, heart failure (NYHA II, III, IV), history of myocardial infarction, or elevated creatinine values.

By modeling three groups of study patients (group A treated with nebivolol, group B treated with perindopril and group C treated with candesartan cilexetil), I was able to create "patient communities", based on common risk factors. For each such community of patients, the medical approach should be different. Furthermore, by applying the insights gained from the network study I am able to improve the efficacy of the established treatments.

A so-called community of patients is characterized by all patients who share a number of common features. In the network-based approach, I build a network in which patients are represented by

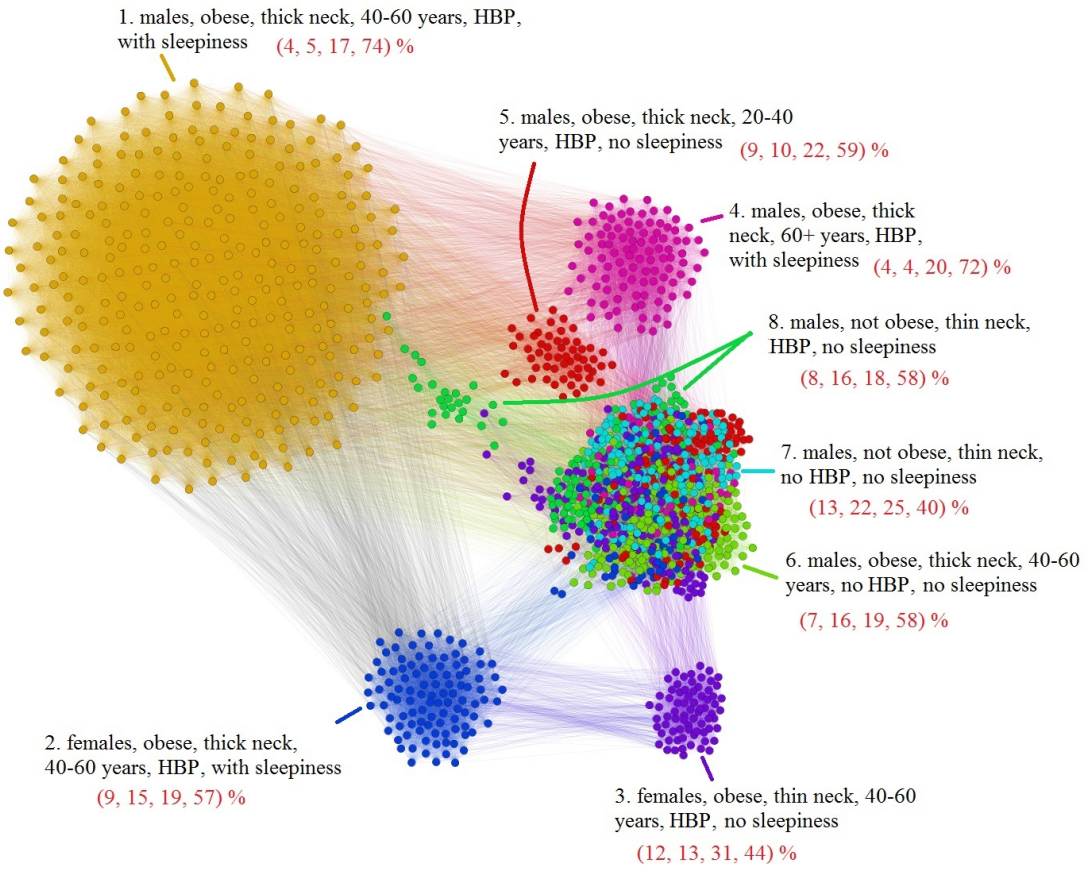


Figure C.5.: Graphical representation of the patient population with clinical apnea signs, by including the Epworth sleepiness score. Node colors correspond to one of the 8 detected communities.

nodes, and the link between any two nodes is created if there is a compatibility relationship between them. In my study, the compatibility relationship exists if 10 well-defined patient characteristics are similar. These characteristics are: age category, gender, the degree of hypertension, cholesterol level, the cholesterol fractions HDL-C and LDL-C, triglycerides, LDL-C/HDL-C ratio, diabetes and cardiovascular risk classes. After building the network according to the above-described methodology, a topological community detection algorithm, namely ForceAtlas2 [132], is used in order to graphically render the corresponding topological patient communities.

If the patients from study groups A, B and C are submitted to my network-based topological analysis, according to the ten characteristics (to which I add the type of used medication), six communities of patients (denoted as T0 - T5) are rendered, as represented in Figure C.6. Communities T0, T1 and T3 contain patients treated with candesartan. Patients from these communities are characterized by the presence of diabetes, pathologically altered fractions of cholesterol (total cholesterol, LDL-C, the ratio of LDL/HDL-C), age over 60 years and a very high cardiovascular risk class. In these communities, there are both male and female patients. For these patients, after the treatment, I notice a tendency towards improving diastolic blood pressure values. Community T2 is characterized by: mainly female patients, a stage II of hypertension, changes in LDL-C and increased cardiovascular risk class. These patients are treated with perindopril. Community T4 is characterised by hypertensive male patients, without any specification regarding the age, which are in an early stage of hypertension and therefore without a high cardiovascular risk. These patients do not present pathological values of biochemical tests and are not diabetics. They are treated with nebivolol. Community T5 patients are characterized by stage II of hypertension and, therefore, moderate cardiovascular risk. This community cannot be associated with any of the three specific antihypertensive drugs which are taken into consideration by my study. From a total of 289 study patients, 82 (28.37%) are included in this community.

The evolution of blood pressure values in a network-based representation, for all study patients, after 12 months of antihypertensive treatment, is given in Figure C.7.

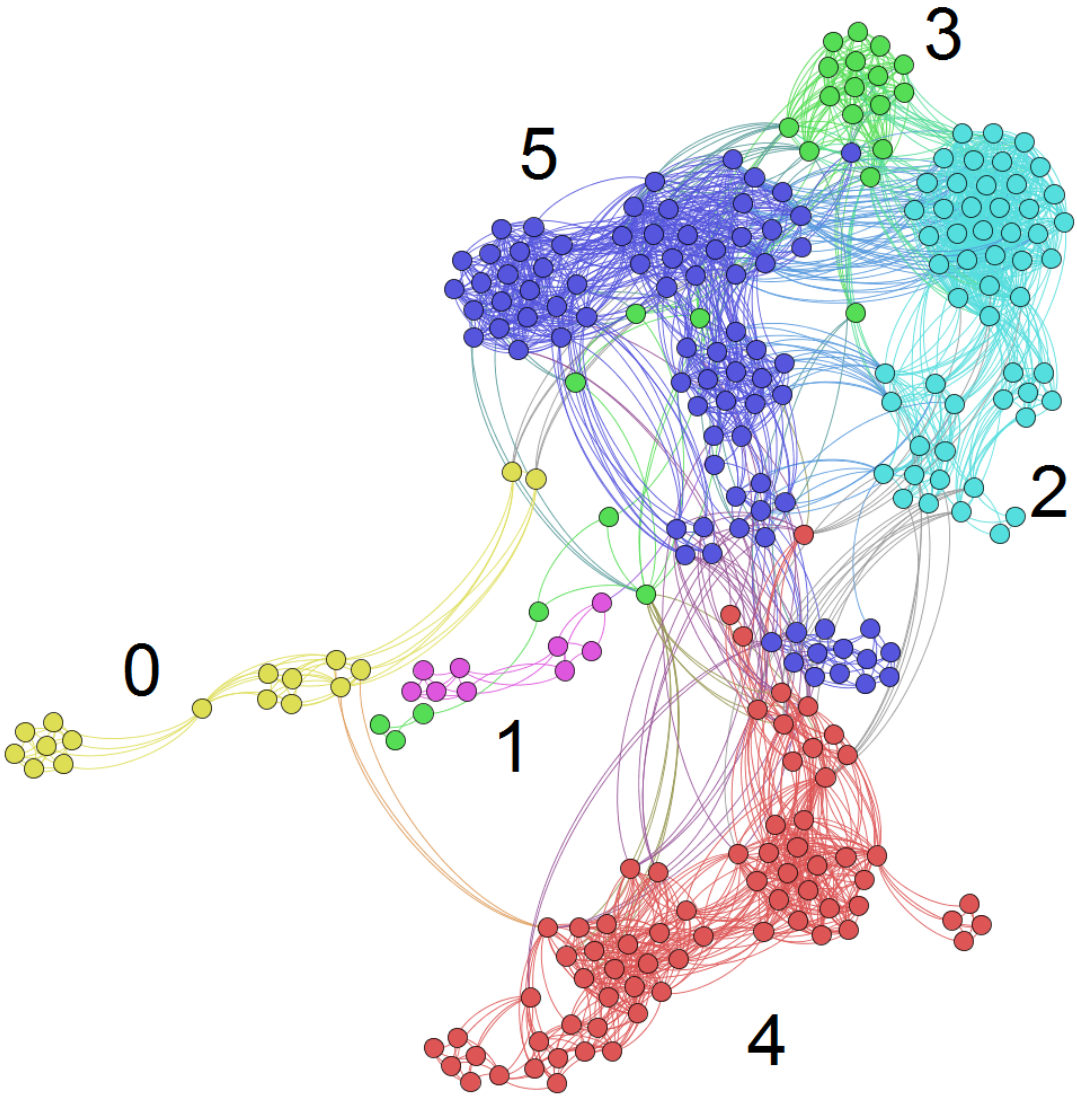


Figure C.6.: Communities of patients from all study groups. Colors are assigned in order to visually identify the communities.

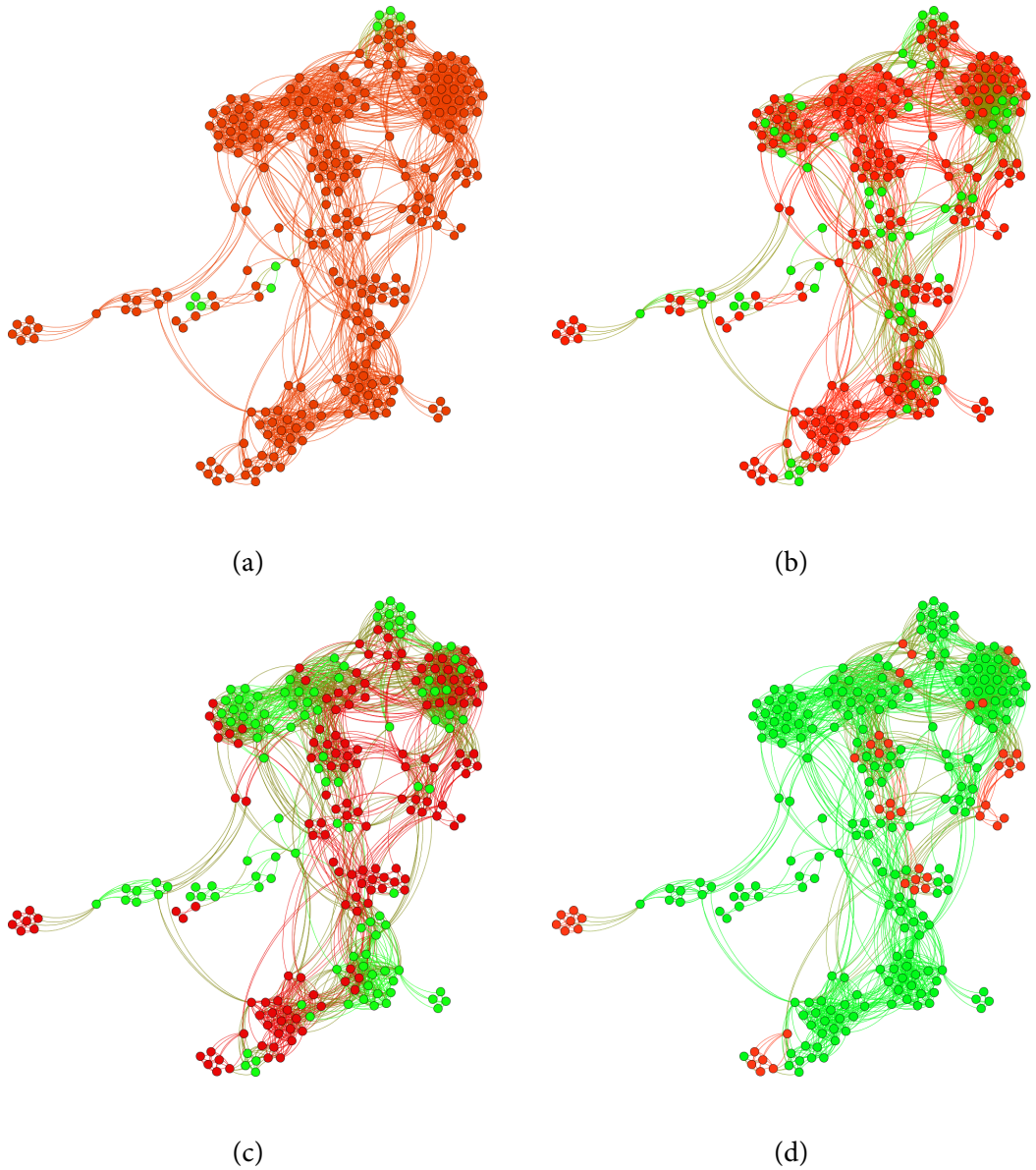


Figure C.7.: Graphical representation of network analysis results of the variation in systolic blood pressure of the study patients before (a) and after the treatment (b), as well as diastolic before (c), and after the treatment (d). Colors represent normal (green) and high (red) values of blood pressure

D. Technological communication optimizations using complex networks

This appendix presents studies undertaken on communication in technological context using to so-called technological networks, as one of the four categories of complex networks. I have modeled technological communication networks of urban roads and wireless sensors. Through the means of graph based modeling and assessment, I am able to offer a different perspective over classic approaches, which has led to optimizations in the cost and effectiveness of communications. These observations would not have been detectable otherwise (i.e. through classic statistical methods). In this chapter I describe how I analyzed and optimized urban traffic networks , and developed an algorithm for placing relays and a central sink in a wireless sensor network, in order to balance cost versus latency in such networks where communication timing is essential.

"Everything you can imagine is real."

 Pablo Picasso

D.1. Road network optimizations using network analysis

Witnessing the large real-world applicability of SNA, and complex networks in general, I, together with my collaborators from the ACSA research group, model and analyze the network formed by road networks in cities from an innovative perspective. Inspired by similar approaches of comparing networks, I create a methodology that proposes the assessment of city road networks based on their motif distributions [189, 190]. To the best of my knowledge, I am the first to fully interpret the roads infrastructure by using network motifs. Based on the similarity of the motif distributions, I choose diverse city topologies, create a similarity graph, and discuss the urban influences one has on each other. Through my analysis, I coin the title of *Social City* to any city which meets particular criteria in terms of optimal roads distribution.

The motivation behind this set of studies is to create incentives for studying road networks from the pure topological perspective of complex networks. Through intensive data mining from online repositories, through network analysis methodologies, and motif distribution analysis, I have created a traffic quality metric, which represents a state-of-the-art analysis of this kind. This study, along with its results allow us to elucidate the mechanisms of urban infrastructure emergence and the way new roads are built to serve adjacent areas. The correlation with network analysis, and especially social networks analysis, has made me attribute the term *Social City* to any city that meets particular requirements to street homogeneity in terms of its topological layout.

Additionally, this study presents a novel perspective on how different cities can be differentiated using a network motif approach corroborated with the state of the art similarity assessment. I quantify the topological differentiation - based on motif analysis - using the network fidelity metric [257]. Even though similar in nature, it is shown in this study that all studied urban networks have specific properties which make them unique in terms of traffic throughput.

D.1.1. Methodology

Significant research has been carried towards finding alternative approaches in analyzing the structure of cities and especially good patterns of roads, which maximize the car traffic throughput. Using graph theory was a clear choice based on the clear historic affiliation of the domain and the suitability of representing the relationships between intersections (nodes) and streets (edges). Much work was put into this segment and there are even some far fetched investigations into creating for example most real artificial driver", which acts as close to a real driver as possible [139]

Understanding how drivers interact and how road networks are created around specific points of interest (schools, shopping centers, concert halls, sports arenas) could lead to identifying the patterns - motifs [189, 190] - that can apply at different scales over several road networks to achieve increased traffic flow and consequently, less congestion.

Jiang and his team identify in [137, 136] a classical 80/20 behavior because roughly 20% of the streets account for more than 80% of the urban traffic. There is a clear distinction between some important streets (few) and some which are less important (many) which leads us to what could be easily presented as a *hierarchical view* of the urban structure [252].

Porta et al. present in [223] a methodology and a framework for analysis of urban environments emphasizing on the layout of the urban roads, in terms of both classical graph theory but also using complex networks specific metrics and algorithms.

There is a trend in measuring the *optimality* of a specific street layout and even finding algorithms and methodologies of dynamically reassigning the traffic light signaling policies in order to indirectly "reconfigure" the network as to maximize various throughput metrics and an approach based on complex networks analysis is used for this.

I obtain the road information data from the online repository OpenStreetMap. This data is parsed into a *gexf* file format using a customly implemented python plugin. From there, I have all intersections as a node list, and all streets inside the city as edges between nodes. The edge data is further parsed to extract only a plain edge list text file. This serves as an input for FANMOD [283]. It is a fast and lightweight motif detection tool based on the state of the art algorithm RAND-ESU [283]. It takes as input an edge list in text format, and offers a detailed motif distribution statistic based on the analysis using the RAND-ESU algorithm.

Due to the complexity of the calculations, I limit my research to the motif distributions for motifs of sizes 3, 4, 5, and 6.

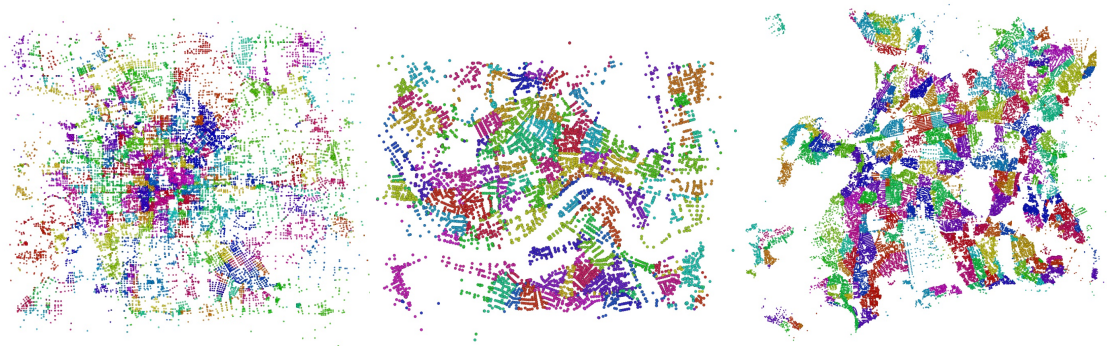
The above explained process - which is partly automatized, partly manual - is repeated for each selected city. Specifically, 16 diverse cities were selected for analysis. Each city corresponds to one of three topographies:

- *Compact* topography: any city that is well structured and non-divided by any major river. The overall structure is regular, and there are no significant bottlenecks in the road infrastructure. These cities are usually found inside the continent, on even terrain.
- *River* topography: any city that is clearly divided by a large river. The impact of such a river is that it creates the need for a few, large bridges which connect the two or more parts of the city. These rivers act as bottlenecks.
- *Seaside* topography: any city that is built along the seaside. While they may maintain a compact structure similar to the first category, these cities are usually spread along the coastline and have a natural bottleneck in the sense that the coastline dictates the layout of streets.

To showcase the topological differences of each urban category, Figure D.1 displays the following cities: Beijing (26215 nodes and 36178 edges), Rotterdam (7828 nodes and 10709 edges), and Cape Town (61058 nodes and 82247 edges).

Further, I obtain the motif distributions D^{city} for the mentioned cities, and then quantify the similarity of each pair of motifs of the same size. The problem is solved through the correlation of any two resulting motif vectors. For each city I obtain a distribution D_3^{city},D_4^{city} ,D_5^{city} ,D_6^{city} , for the 16 cities. Each distribution D_i^{city} is a vector of normalized percentages corresponding to the occurrences of each individual motif of size i. To correlate any two vectors of same size, for different cities, D_i^{city1} and D_i^{city2}, I make use of the existing fidelity metric φ [257]. It is worth to be noted that I use each city in turn as a reference for calculating the similarities of the other cities to it. I repeat this process by taking each city as a reference. The analysis is undertaken for each topographic category independently.

The resulting mutual similarities between any to cities can be modeled as directed weighted edges in a digraph. The information I have extracted using OpenStreetMap offers information about the direction of the street as well. As directed motifs offer more insight regarding functional patterns than undirected ones, I have decided to use the street networks in a directed context [190]. The



(a) (b) (c)

Figure D.1.: City topographies visualized using Gephi. **a.** Beijing (compact) **b.** Rotterdam (river) **c.** Cape Twon (seaside). All nodes are colored according to the community they belong to. This community corresponds to the local neighborhood, and was determined using an existing clustering algorithm built in Gephi [38].

graph is created from the numerical φ-values and imported in Gephi where it is visualized. To better assess the road similarity, two different perspectives are presented:

- Nodes' indegree: the sum of all weights of all incoming edges represents the traffic similarity of other cities towards one particular city. Such a city with a high indegree can be considered a model in terms of road infrastructure.
- Nodes' outdegree: the sum of all weights of all outgoing edges represents the traffic similarity of one particular city to other cities. Such a city with a high outdegree can be considered a homogeneous replica of other cities.

D.1.2. Results

Based on the presented approach, I present the similarity assessment results and graph interpretation. Table D.1 contains the similarity values φ for comparing each of the six cities to each other on the motif vectors of same size. The acronyms of the cities in Table D.1 correspond to: Beijing (Bei), Bucharest (Buc), Johannesburg (Joh), Madrid (Mad), Mexico City (Mex), New Delhi (Del), Budapest (Bud), Cairo (Cai), Rotterdam (Rot), Barcelona (Bar), Buenos Aires (Bue), Cape Town (Cap), Mumbai (Mum), Rio de Janeiro (Rio), Singapore (Sin), and Vancouver (Van). The table contains the results for comparing the empirical data in terms of motif sizes: 3, 4, 5, and 6, averaged respectively, for each of the three defined topographic categories, from top to bottom: compact, river, and seaside. The reference models are the cities written in the column headers, to which each city on each row is compared. As an example, the fidelity of Bucharest (Buc) to Beijing (Bei) is 0.747. Is is also worth mentioning that the φ-function is not symmetric if the reference is interchanged with the compared model. Following the same example, the fidelity of Beijing (Bei) to Bucharest (Buc) is 0.555. This translates into the fact that the motif distribution of streets in Bucharest has meaningful characteristics which better map on the road network found in Beijing, but not vice-versa. Beijing has the

same characteristic motif representatives in particular, but it also has (many) other relevant motifs characterizing the city. This is why I need to compare each city to each other to better understand the structure.

From Table D.1 I can extract some conclusions regarding individual city similarities. In terms of compact cities, the highest similarity outdegree (on rows) is that of Bucharest, while the highest similarity indegree (on columns) is that of Johannesburg. I can immediately observe that Johannesburg is the city with the most similarity oriented towards it. On the other hand, Bucharest has the highest outgoing similarity, suggesting that it resembles the most to other city road networks. To enhance the numerical interpretation I chose to use a graphical representation of mutual similarity. For river cities, Budapest has the highest outdegree and Rotterdam the highest indegree. Finally, for seaside cities, Mumbai and Buenos Aires have the highest outdegrees, while Cape Town has the highest indegree.

Figure D.2 shows the distributions of similarity towards the reference cities for each topography. In Figure D.2 one can denote the histograms of motif similarities averaged over all motif sizes (3 to 6). Comparing the histograms, I can conclude which city has the highest overall fidelity in each category. For example, Mexico City has a poor urban infrastructure in terms of a compact city, and Rio de Janeiro is not a typical seaside city. On the other hand, Bucharest is a typical social city for compact topographies, Budapest for river cities, and Mumbai is by far the best candidate for social city, in terms of resemblance to Cape Town.

What is noteworthy is that by following an innovative methodology of comparing city road infrastructures I am able to obtain quantifiable scores for assessing the road quality and resemblance to any other city. This is in turn matched to the geographical layout similarities between cities, like I obtained for Johannesburg, Rotterdam and Cape Town.

The work I have undertaken together with my collaborators represents a novel approach in comparing the road infrastructure of cities, and is a continuation of the work started in [259]. Moving from a generic comparison between diverse cities, to topographically similar cities, has bridged the gap between understanding how the geography affects the emergent infrastructure of a city. Combining concepts from the area of network analysis and mapping them to road networks, I succeed to add improvements in existing techniques that address the analysis of traffic control. My research is focused around a network motif-based algorithm which assigns cities a so called *Social City*-score.

D.2. Performance versus cost optimizations of wireless sensor networks

Applying complex network analysis principles in order to analyze and optimize sensor networks is nothing but natural as the network perspective provides an innovative means of analyzing the structure of entities with a social-like structure [278]. Thus, I can detect influential nodes, patterns of communication and also study dynamics inside the network. This strongly relates to wireless sensor networks as it is important to disseminate which sensor nodes are critical for the data throughput, which are more central so that relays can be placed at those positions, and also model growth as the network coverage spreads in time.

In this discussion, I are going to differentiate between regular nodes, responsible for gathering data and/or acting upon received commands and relay nodes which collect data from the nodes in

Table D.1.: Motif-based network fidelity of cities (rows) using each other as reference models (columns), averaged over motifs of sizes 3-6. A lower value of φ means a lower resemblance to the reference motif distribution. The similarity is computed based on each of the three topographic categories.

Compact	References					
	Bei	Buc	Joh	Mad	Mex	Del
Bei	1	0.555	0.652	0.672	0.502	0.556
Buc	0.747	1	0.713	0.745	0.576	0.679
Joh	0.546	0.449	1	0.555	0.465	0.472
Mad	0.641	0.552	0.636	1	0.472	0.523
Mex	0.509	0.42	0.573	0.521	1	0.492
Del	0.677	0.611	0.7	0.669	0.603	1

River	References		
	Bud	Cai	Rot
Bud	1	0.525	0.72
Cai	0.431	1	0.607
Rot	0.477	0.435	1

Seaside	References						
	Bar	Bue	Cap	Mum	Rio	Sin	Van
Bar	1	0.0.42	0.679	0.627	0.715	0.671	0.532
Bue	0.608	1	0.621	0.599	0.663	0.587	0.539
Cap	0.426	0.279	1	0.472	0.5	0.545	0.33
Mum	0.613	0.41	0.76	1	0.677	0.729	0.477
Rio	0.502	0.33	0.593	0.488	1	0.518	0.407
Sin	0.487	0.297	0.651	0.548	0.523	1	0.369
Van	0.624	0.443	0.639	0.585	0.663	0.607	1

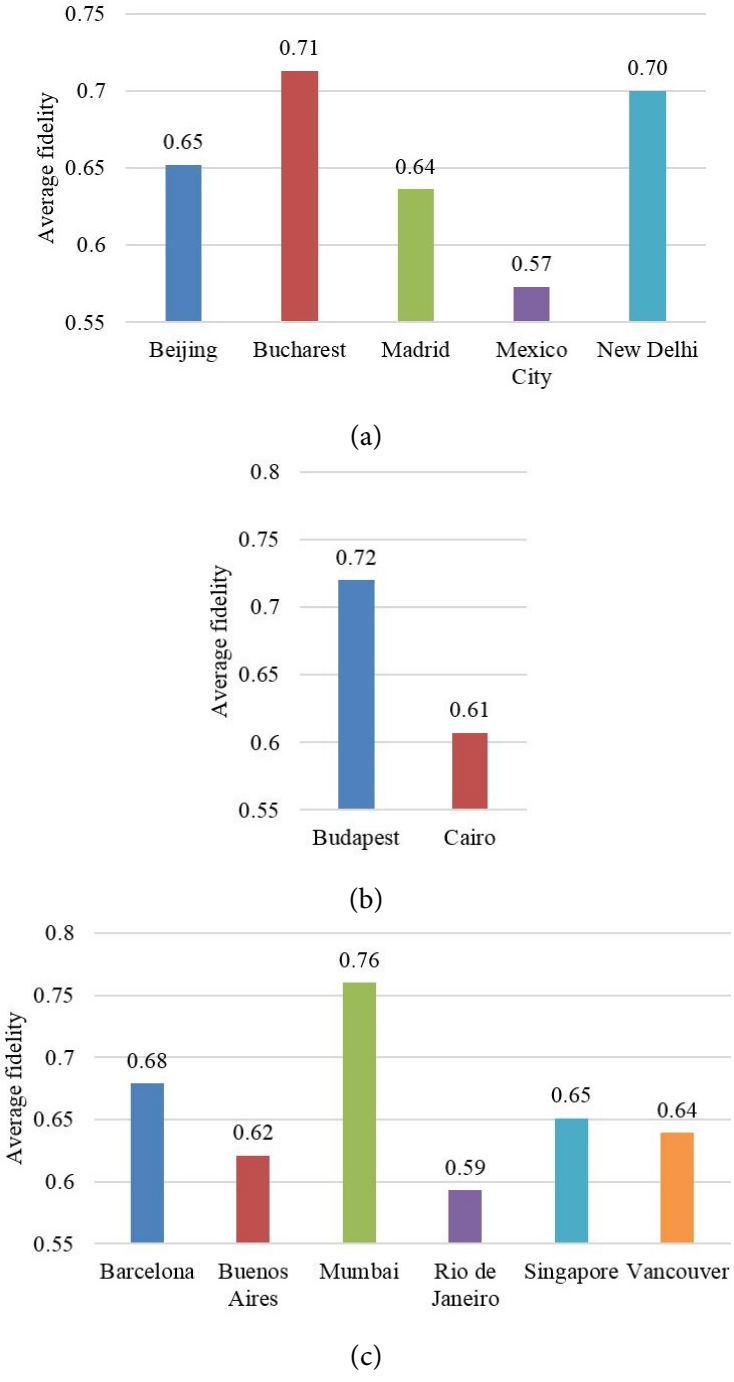


Figure D.2.: Network fidelity of city road networks compared to the Social City of each topographic category: **a.** Johannesburg (compact). **b.** Rotterdam (river). **c.** Cape Town (seaside). A lower fidelity (column height) means less resemblance to the reference model.

the direct area of coverage and send them upstream to the network sink. The aim of this chapter is to propose an optimization solution for choosing the number of required relays and their optimal position so that I maximize the performance of the network while keeping the overhead at a minimum.

In the design of a WSN the practitioner has to balance the costs involved with the solution, with the performance, and one of the key performance metrics is the average delay from node to sink. My research is part of a larger endeavor of designing and deploying a near real-time sensor network for monitoring and reporting data regarding road traffic conditions and consequently dynamically adapt the state of the traffic lights.

Given any two-dimensional WSN, I model it as graph $G = \{V,\ E\}$, composed out of nodes (vertices) V and edges E. The set of edges consists of all wireless links between all pairs of sensors inside each node's coverage area, like in an ad-hoc network. The requirements are as follows:

- assign one sink for the network: $n_s \in V$,
- assign an optimal number of relay nodes: $r_i \in R,\ R \subset V,\ |R| \ll |V|\ ,\ n_s \in R$,

in order to balance a maximal performance and a minimal cost for G. I consider the relay nodes r_i interconnected using cable links with negligible latency and infinite power supply. The performance is expressed in terms of number of hops required to reach the nearest relay (relay-to-sink communication is considered negligible) and the cost is expressed in the number of required relays $|R|$.

D.2.1. Background

Mostly dependent on the task of the network but also of the particular conditions and the type of sensors there are two major strategies in placing the nodes of a sensor network: deterministic and random. The first one, when possible, can ensure great coverage with careful placement of the nodes and even the logical topology of the network can be established at deployment time [168].

Because of the adverse condition on the field there are situations where the single possible option for deploying nodes is in a random manner. This has adverse effects on the main metrics of a WSN [163]. In any situation where there is a large distance between two adjacent nodes, I witness a low throughput and high energy consumption.

Rich literature exists on the topic of optimal node placement [130], which is considered an NP-hard problem [28] and some non-deterministic approaches were proposed, which provide sub-optimal results [222].

Much because the current approaches in deterministic placement of the nodes proven themselves problematic but also because some of the typical WSN deployment scenarios presented both in the literature and also in the real life scenarios, such as wild fire prevention, battlefield monitoring or disaster rescue, require a random distribution of the nodes, even if there are some possibility of controlling the density of the nodes [28] I decided to investigate the problem of relay placement strategies in this case.

D.2.2. Methodology

In order to generate the input data I use a WSN topology generator which produces a topology of nodes (sensors) with 2D geographical data. I convert the information into *gdf* file format which

can be imported in Gephi [30], the leading tool in large graph data visualization. Any layout of sensors in a geographic space can be processed by my algorithm by importing it in Gephi, where my developed plugin can be used from. Further, the enhancement algorithm, called SIDeWISE, processes the topological data. The result is the initial sensor network with an additional overlapping layer of optimally placed relays which are all connected to a sink through a heuristically obtained minimum cost spanning tree.

I further present the SIDeWISE (SocIally enhanceD WIreless Sensor nEtwork) algorithm step by step. I start with a topology of given sensors $|V|$which all have positional data attached. Also, a wireless coverage range r is given, and a $resolution$ parameter which controls the density of required relays to be placed. In step A of the given algorithm, a graph $G = \{V,\ E\}$ is formed by connecting each two sensors that are within each other's coverage range r. The distance $(n_i,\ n_j)$ is defined as the Euclidean distance between the two points $(x_i,\ y_i)$ and $(x_j,\ y_j)$.

Step B of the algorithm implies determining which node would best fit as being the (single) sink of the network. Once the graph is obtained, the Eigenvector centrality algorithm is run on G [206]. The Eigenvector centrality was chosen as it is considered a well predictor of a node's influence in a network [276, 153]. The centrality algorithm is defined in such a way that only one node - the most central - has a centrality of 1.0. All other nodes have their centralities between 0 and 1. Consequently I define node n_s with maximal centrality as the sink for G, in contrast to the geographically centered approach proven less effective [288, 294]. Once this node is determined it becomes the sink n_s, and is added to the set of relays R.

Step C determines the clusters of sensors which are relevant to the network from the throughput perspective. While it is not a common practice to determine communities in sensor networks, communities are highly relevant in social networks [205]. As such, in order to determine the optimal number of needed relays I run a community detection algorithm on the network G, by measuring its modularity. A community detection algorithm is a method for grouping individuals (nodes) into clusters in which all elements share one or more common properties [38]. In this case, the commonly shared property is the position of each sensor. A parameter named *resolution* can influence the number of detected communities. In comparison to the default *resolution* value of 1.0, a custom *resolution* < 1.0 will determine smaller/more communities, and a *resolution* > 1.0 will determine larger/less communities [151]. I discuss the impact of using a custom *resolution* in the next chapter. Measuring the modularity of a realistic network (i.e. not regular, not evenly spread) results in a high number of communities with various sizes. As there are always small communities formed out of several stranded nodes, I ignore all communities with a total size smaller than a fraction λ of the total population. It is important to mention that discarding does not mean the sensors are removed from the network, it means that those groups of sensors will be considered irrelevant for the next step of the relay placement algorithm.

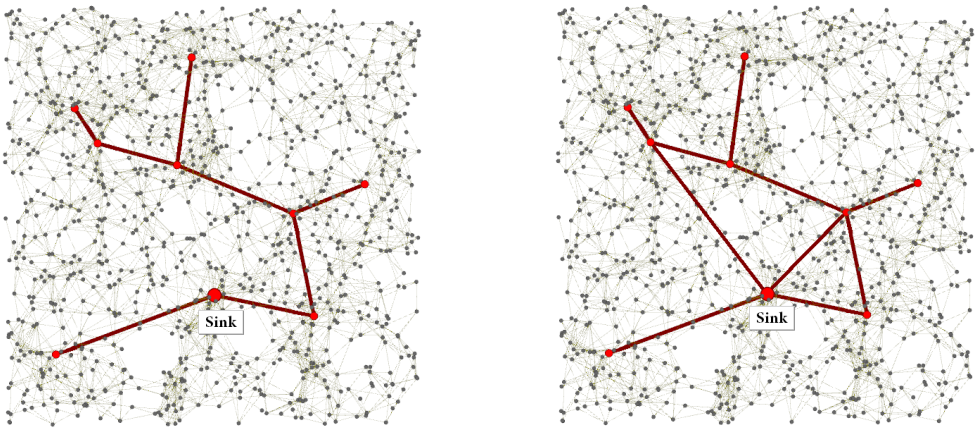
Step D is an iterative process similar to step B, but it is applied on each individual community previously determined. The number of relays is determined by the number of relevant communities (i.e. size > λ fraction of the population) during step C while the relays themselves are chosen during this step. Measuring the centrality distribution of each community, I choose the most central node as a relay. As mentioned before, the central node is the closest to all other nodes in its community. This is relevant to wireless sensors because the existing edges are determined by position, and so it becomes straightforward and efficient to choose a relay to whom any sensor requires the minimum number of hops to reach.

Algorithm D.1 SIDeWISE pseudocode for a cost-throughput-optimal network

```
Input: raw wireless sensor network with |V| nodes with positional data (x_i, y_i), and
wireless coverage range r.

A: Link all nodes of G in wireless range r
1 :E ← {}
2 :for each pair of nodes (n_i, n_j) in V:
3 :  if distance(n_i, n_j) < r:  e_ij ← create edge between (n_i, n_j)
4 :  E = E ⋃ e_ij
B: Assign the sink n_S to G
5 :for each node n_i in V:
6 :  C[n_i] ← compute centrality {G, n_i}
7 :find n_S in V where C[n_S] = 1.0 (maximal)
8 :R = {n_S}
C: Detect and filter communities Com
9 :Com ← community detection algorithm G, resolution
10:for each com_i in Com:
11:  if size(com_i) < λ × |V|:  (5% of population)
12:    Com = Com \ {com_i}
D: Assign relays R for each community com_i
13:for each community com_i in Com:
14:  for each node n_ik in com_i:
15:    C[n_ik] ← compute centrality {com_i, n_ik}
16:  find n_ik in com_i where C[n_ik] = 1.0 (maximal)
17:  R = R ⋃ n_ik
E: Create an MST for relay-graph G_R = (R, MSTE_R)
18:E_R ← {}
19:for each pair of relays (r_i, r_j) in R:
20:  e_ij ← create edge between (r_i, r_j)
21:  E_R = E_R ⋃ e_ij  (complete graph)
22:G_R R, MSTE_R ← Kruskal-MST R, E_R
23:E_R = E_R \ MSTE_R  (complete graph minus MST)
F: Centralize sink n_S
24:while C(n_S) < 1.0:
25:  for each edge e_i in E_R:
26:    MSTE_R = MSTE_R ⋃ e_i
27:    C(n_S) ← compute centrality G_R, n_S
28:    fitness(e_i) ← C(n_S)
29:    MSTE_R = MSTE_R \ e_i
30:  find e_r in E_R where fitness(e_r) is maximal
31:  MSTE_R = MSTE_R ⋃ e_r
32:  E_R = E_R \ e_r
33:  C(n_S) ← fitness(e_r)

Output: wireless sensor network G with cost-optimal overlapping relay and sink
physical network MSTG_R = {R, MSTE_R}.
```



(a) (b)

Figure D.3.: Heuristic optimization of the MST to increase throughput of relays (red nodes). **a.** A relay network connected with an MST. **b.** The same MST but with an additional two edges so that the sink (bigger red node) becomes the central node of the network.

Considering that now all relevant wireless sensors have a relay in their vicinity, step E processes set R in order to create a secondary, overlapped graph of edges that connect all relays and the sink. The edges represent physical links, like broadband cable. Coverage of the set with edges is done using Kruskal's minimum spanning tree (MST) algorithm and I obtain $G_R = (R,\ MSTE_R)$ [149].

The final step of the algorithm is step F. At this point I have a cost-optimal overlapping tree of relays, $MSTG_R$, but from a throughput perspective it may result as highly latent. This problem is depicted in Figure D.3. The smallest (gray) nodes represent wireless sensors, the red nodes represent relays and the single larger red node is the sink. The red edges are the physical links connecting the relays. The next optimization tries to heuristically lower the average number of hops required for relays to reach the sink. As such, an iterative process of maximizing the sink's centrality is proposed. Even though the ideal result would be to connect all relays with the sink directly, this would alter the cost-optimality. Thus, a trade-off solution like the one depicted in Figure D.3b is preferred. It shows how the sink is made more "central" by adding two more edges to the MST resulting from Kruskal's algorithm. Most relays now have a distance of one or two hops with the addition of only two edges to the MST. The algorithm tries to add another edge to the MST, measures the sink's new resulting centrality and keeps this as a fitness for the added edge. It does this for all candidates. After one iteration, the edge with the best fitness is kept in the MST. This is repeated until the centrality of the sink becomes 1.

D.2.3. Discussion

Based on the full simulation results described in [129], I make an analogy with the small-world property which represents an ideal balance between the characteristics of a regular network and a random network [276], Figure D.4 suggests the same principle: the *socially* enhanced wireless sensor networks lie at the ideal crossroads between cost and performance. One the left side is a network with

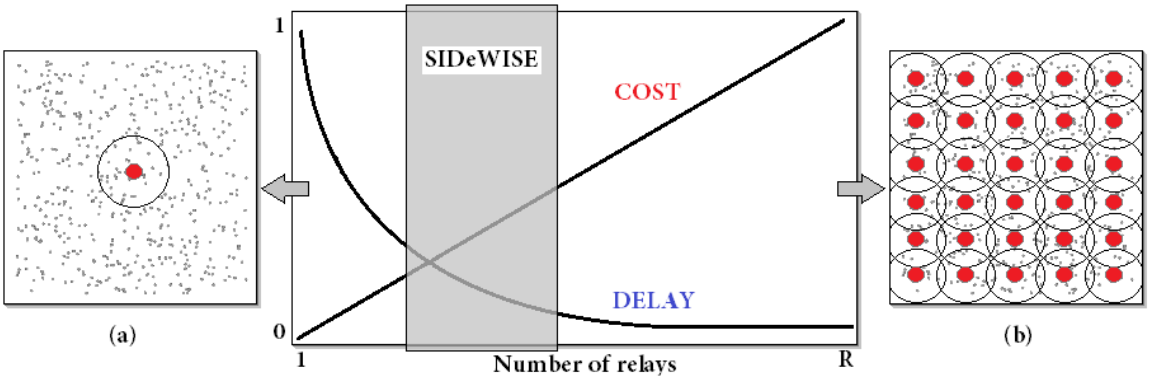


Figure D.4.: The SIDeWISE algorithm balances cost and propagation delay by optimizing the placement of the relays in a WSN. The two extreme cases are represented by a single-sink network (a) and a network fully covered by relays (b).

just one sink and no relays. While being cost-optimal ($1r$ (relay)), it offers the worst performance as the propagation delay is maximal (6.98τ (average number of hops from any sensor the the nearest relay)). On the right side is a network fully covered by relays. In this case the delay is optimal (1τ) but the cost is maximized ($100r$). As the graphics of the delay and cost show, there is a window in which I can create a network with the best possible trade-offs: a relatively low delay (i.e. high performance) and a low cost. This is the type of enhancement which the SIDeWISE algorithm facilitates. Experimentally I prove that my proposed solution has a cost of $7r$ with a delay of only 3.62τ. This yields a 92% performance improvement over (a) and and only uses 7% of the relays required for (b).

This work represents a novel approach in designing the placement of relay nodes in a sensor network. By using concepts from the area of social network analysis and mapping them to the already classical field of sensor networks I succeed to add improvements to the costs implied with deploying the infrastructure. This research is focused around the algorithm I have devised with my close collaborators, called SIDeWISE.

Bibliography

[1] A. Abbasi, L. Hossain, and L. Leydesdorff. Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, 6(3): 403–412, 2012.

[2] D. Acemoglu and A. Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1):3–49, 2011.

[3] D. Acemoglu, A. Ozdaglar, and E. Yildiz. Diffusion of innovations in social networks. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pages 2329–2334. IEEE, 2011.

[4] D. Acemoğlu, G. Como, F. Fagnani, and A. Ozdaglar. Opinion fluctuations and disagreement in social networks. *Mathematics of Operations Research*, 38(1):1–27, 2013.

[5] L. Adamic, O. Buyukkokten, and E. Adar. A social network caught in the web. *First monday*, 8(6), 2003.

[6] H. B. Adamic L. Comment to "emergence of scaling in random networks". *arXiv:cond-mat/0001459*, 2000.

[7] E. Adar. Guess: a language and interface for graph exploration. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 791–800. ACM, 2006.

[8] R. Alberich, J. Miro-Julia, and F. Rosselló. Marvel universe looks almost like a real social network. *arXiv preprint cond-mat/0202174*, 2002.

[9] J. S. Albert and R. E. Reis. *Historical biogeography of Neotropical freshwater fishes*. Univ of California Press, 2011.

[10] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[11] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8 (6):450–461, 2007.

[12] C. Alt, O. Astrachan, J. Forbes, R. Lucic, and S. Rodger. Social networks generate interest in computer science. *ACM SIGCSE Bulletin*, 38(1):438–442, 2006.

[13] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000.

[14] T. Arentze, P. van den Berg, and H. Timmermans. Modeling social networks in geographic space: approach and empirical application. *Environment and Planning-Part A*, 44(5):1101, 2012.

[15] S. E. Asch. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9):1, 1956.

[16] D. Auber. Tulip: A huge graph visualization framework. In *Graph Drawing Software*, pages 105–126. Springer, 2004.

[17] R. N. Aurora and N. M. Punjabi. Obstructive sleep apnoea and type 2 diabetes mellitus: a bidirectional association. *The Lancet Respiratory Medicine*, 1(4):329–338, 2013.

[18] R. Axelrod. The dissemination of culture a model with local convergence and global polarization. *Journal of conflict resolution*, 41(2):203–226, 1997.

[19] R. Axelrod and W. D. Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

[20] R. Axtell. Why agents?: on the varied motivations for agent computing in the social sciences. 2000.

[21] J. P. Bakker, S. B. Montesi, and A. Malhotra. Obstructive sleep apnoea: new associations and approaches. *The Lancet Respiratory Medicine*, 1(1):e15–e16, 2013.

[22] S. Bandyopadhyay, A. R. Rao, B. K. Sinha, and B. K. Sinha. *Models for Social Networks with Statistical Applications*, volume 13. Sage, 2011.

[23] A.-L. Barabasi. Linked: How everything is connected to everything else and what it means. *Plume Editors*, 2002.

[24] A.-L. Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

[25] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439): 509–512, 1999.

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

[27] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[28] A. Bari, Y. Chen, A. Jaekel, and S. Bandyopadhyay. A new architecture for hierarchical sensor networks with mobile data collectors. In *Distributed Computing and Networking*, pages 116–127. Springer, 2010.

[29] G. Barina, A. Topirceanu, and M. Udrescu. Musenet: Natural patterns in the music artists industry. In *Applied Computational Intelligence and Informatics (SACI), 2014 IEEE 9th International Symposium on*, pages 317–322. IEEE, 2014.

[30] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. In *ICWSM*, 2009.

[31] V. Batagelj and A. Mrvar. Pajek-program for large network analysis. *Connections*, 21(2):47–57, 1998.

[32] A. Bavelas. Communication patterns in task-oriented groups. *Journal of the acoustical society of America*, 1950.

[33] W. L. Bennett and A. Segerberg. Digital media and the personalization of collective action: Social technology and the organization of protests against the global economic crisis. *Information, Communication & Society*, 14(6):770–799, 2011.

[34] G. Bianconi, R. K. Darst, J. Iacovacci, and S. Fortunato. Triadic closure as a basic generating mechanism of the structure of complex networks. *arXiv preprint arXiv:1407.1664*, 2014.

[35] A. Bigdeli, A. Tizghadam, and A. Leon-Garcia. Comparison of network criticality, algebraic connectivity, and other graph metrics. In *Proceedings of the 1st Annual Workshop on Simplifying Complex Network for Practitioners*, page 4. ACM, 2009.

[36] K. R. Bisset, J. Chen, X. Feng, V. Kumar, and M. V. Marathe. Epifast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *Proceedings of the 23rd international conference on Supercomputing*, pages 430–439. ACM, 2009.

[37] S. Biswas, A. K. Chandra, A. Chatterjee, and B. K. Chakrabarti. Phase transitions and non-equilibrium relaxation in kinetic models of opinion formation. In *Journal of Physics: Conference Series*, volume 297, page 012004. IOP Publishing, 2011.

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

[39] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[40] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Physical review E*, 70(5):056122, 2004.

[41] M. Bojanowski and R. Corten. Measuring segregation in social networks. *Social Networks*, 39: 14–32, 2014.

[42] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the african web. In *The Eleventh International WWW Conference*, volume 66, 2002.

[43] M. R. Bonsignore, W. T. McNicholas, J. M. Montserrat, and J. Eckel. Adipose tissue in obesity and obstructive sleep apnoea. *European Respiratory Journal*, 39(3):746–767, 2012.

[44] J. Borondo, F. Borondo, C. Rodriguez-Sickert, and C. Hidalgo. To each according to its degree: The meritocracy and topocracy of embedded markets. *Scientific reports*, 4, 2014.

[45] M. E. Brashears. Humans use compression heuristics to improve the recall of social networks. *Scientific reports*, 3, 2013.

[46] J. A. C. Brown. *Techniques of persuasion: From propaganda to brainwashing*, volume 604. Penguin books Middlesex, England, 1963.

[47] R. S. Burt. Attachment, decay, and social network. *Journal of Organizational Behavior*, 22(6): 619–643, 2001.

[48] P. Cano, O. Celma, M. Koppenberger, and J. M. Buldu. Topology of music recommendation networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16(1):013107, 2006.

[49] N. Caseiro and P. Trigo. Comparing complex networks: An application to emergency managers mental models. 2012.

[50] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Physical Review E*, 71(2):027103, 2005.

[51] D. Centola and M. Macy. Complex contagions and the weakness of long ties1. *American Journal of Sociology*, 113(3):702–734, 2007.

[52] H. Chang, B.-B. Su, Y.-P. Zhou, and D.-R. He. Assortativity and act degree distribution of some collaboration networks. *Physica A: Statistical Mechanics and its Applications*, 383(2):687–702, 2007.

[53] K. Channakeshava, K. Bisset, V. A. Kumar, M. Marathe, and S. Yardi. High performance scalable and expressive modeling environment to study mobile malware in large dynamic networks. In *Parallel & Distributed Processing Symposium (IPDPS), 2011 IEEE International*, pages 770–781. IEEE, 2011.

[54] H. Chau, C. Wong, F. Chow, and C.-H. F. Fung. Social judgment theory based model on opinion formation, polarization and evolution. *Physica A: Statistical Mechanics and its Applications*, 415:133–140, 2014.

[55] G. Chen, X. Wang, and X. Li. *Fundamentals of Complex Networks: Models, Structures and Dynamics*. John Wiley & Sons, 2014.

[56] W.-K. Chen. *Graph theory and its engineering applications*, volume 5. World Scientific, 1997.

[57] Y. Chen, L. Zhang, and J. Huang. The watts–strogatz network model developed by including degree distribution: theory and computer simulation. *Journal of Physics A: Mathematical and Theoretical*, 40(29):8237, 2007.

[58] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[59] N. A. Christakis and J. H. Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Hachette Digital, Inc., 2009.

[60] F. Chung. Open problems of paul erdos in graph theory. *Journal of Graph Theory*, 25(1):3–36, 1997.

[61] N. R. Clark, K. Hu, E. Y. Chen, Q. Duan, and A. Maayan. Characteristic direction approach to identify differentially expressed genes. *arXiv preprint arXiv:1307.8366*, 2013.

[62] J. J. Clarkson, Z. L. Tormala, D. D. Rucker, and R. G. Dugan. The malleable influence of social consensus on attitude certainty. *Journal of Experimental Social Psychology*, 49(6):1019–1022, 2013.

[63] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[64] S. Coles, J. Bawa, L. Trenner, and P. Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[65] F. Collet and P. Hedström. Old friends and new acquaintances: Tie formation mechanisms in an interorganizational network generated by employee mobility. *Social Networks*, 35(3): 288–299, 2013.

[66] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[67] G. Csányi and B. Szendrői. Structure of a large social network. *Physical Review E*, 69(3):036131, 2004.

[68] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5), 2006.

[69] P. Csermely, T. Korcsmáros, H. J. Kiss, G. London, and R. Nussinov. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology & therapeutics*, 138(3):333–408, 2013.

[70] A. Cuevas, M. Febrero, and R. Fraiman. An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122, 2004.

[71] K. Cukier. *Data, data everywhere: A special report on managing information*. Economist Newspaper, 2010.

[72] A. Curtis, S. Lambert, and R. Television. *The century of the self*. bnpublishing. com, 2006.

[73] L. Daqing, K. Kosmidis, A. Bunde, and S. Havlin. Dimension of spatially embedded networks. *Nature Physics*, 7(6):481–484, 2011.

[74] A. Das, S. Gollapudi, and K. Munagala. Modeling opinion dynamics in social networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 403–412. ACM, 2014.

[75] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

[76] L. Deng, Y. Liu, and F. Xiong. An opinion diffusion model with clustered early adopters. *Physica A: Statistical Mechanics and its Applications*, 392(17):3546–3554, 2013.

[77] N. K. Denzin and Y. S. Lincoln. *The SAGE handbook of qualitative research*. Sage, 2011.

[78] A. Duma and A. Topirceanu. A network motif based approach for classifying online social networks. In *Applied Computational Intelligence and Informatics (SACI), 2014 IEEE 9th International Symposium on*, pages 311–315. IEEE, 2014.

[79] R. I. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992.

[80] W. R. Duncan, B. Jancar-Webster, and B. Switky. *World Politics in the 21st Century: Student Choice Edition*. Cengage Learning, 2008.

[81] E. Durkheim. *Durkheim: The Rules of Sociological Method: And Selected Texts on Sociology and Its Method*. Palgrave Macmillan, 2013.

[82] D. Easley and J. Kleinberg. *Networks, crowds, and markets*, volume 8. Cambridge Univ Press, 2010.

[83] F. Echenique and R. G. Fryer Jr. A measure of segregation based on social interactions. *The Quarterly Journal of Economics*, pages 441–485, 2007.

[84] D. Elkind. Egocentrism in adolescence. *Child development*, pages 1025–1034, 1967.

[85] J. Ellson, E. Gansner, L. Koutsofios, S. C. North, and G. Woodhull. Graphviz: open source graph drawing tools. In *Graph Drawing*, pages 483–484. Springer, 2002.

[86] P. Erdos and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[87] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM, 1999.

[88] H. Fang, J. Zhang, and N. M. Thalmann. A trust model stemmed from the diffusion theory for opinion evaluation. In *Proceedings of the 2013 international conference on autonomous agents and multi-agent systems*, pages 805–812. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[89] J.-q. Fang, X.-f. Wang, Z.-g. Zheng, Q. Bi, Z.-r. Di, and L. Xiang. New interdisciplinary science: Network science (1). *PROGRESS IN PHYSICS-NANJING-*, 27(3):239, 2007.

[90] C. Fass, M. Ginelli, and B. Turtle. *Six Degrees of Kevin Bacon*. Plume, 1996.

[91] E. Ferrara and G. Fiumara. Topological features of online social networks. *arXiv preprint arXiv:1202.0331*, 2012.

[92] A. Fonseca. Modeling political opinion dynamics through social media and multi-agent simulation. In *First Doctoral Workshop for Complexity Sciences*, 2011.

[93] G. Fowler. Facebook: One billion and counting. *The Wall Street Journal*, page B1, 2012.

[94] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[95] L. C. Freeman. *The development of social network analysis: A study in the sociology of science*, volume 1. Empirical Press Vancouver, 2004.

[96] P. Fu and K. Liao. An evolving scale-free network with large clustering coefficient. In *Control, Automation, Robotics and Vision, 2006. ICARCV'06. 9th International Conference on*, pages 1–4. IEEE, 2006.

[97] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423(6937):267–270, 2003.

[98] L. K. Gallos, F. Q. Potiguar, J. S. Andrade Jr, and H. A. Makse. Imdb network revisited: unveiling fractal and modular properties from a typical small-world network. *PloS one*, 8(6):e66443, 2013.

[99] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. In *Proceedings of the 3rd conference on Online social networks*, volume 39, page 3âAS3, 2010.

[100] M. Gerber and J. Linda. Sociology 7th canadian ed, 2010.

[101] S. Geven, J. Weesie, and F. van Tubergen. The influence of friends on adolescents behavior problems at school: The role of ego, alter and dyadic characteristics. *Social Networks*, 35(4): 583–592, 2013.

[102] G. Giaquinto, C. Bledsoe, and B. McGuirk. *Influence and Similarity between Contemporary Jazz Artists, plus Six Degrees of Kind of Blue*. PhD thesis, Citeseer, 2007.

[103] A. Gionis, E. Terzi, and P. Tsaparas. Opinion maximization in social networks. *ArXiv. Prepr. ArXiv*, 1301, 2013.

[104] L. M. Given. *The Sage encyclopedia of qualitative research methods*. Sage Publications, 2008.

[105] P. M. Gleiser and L. Danon. Community structure in jazz. *Advances in complex systems*, 6(04): 565–573, 2003.

[106] J. Golbeck. *Analyzing the Social Web*. Access Online via Elsevier, 2013.

[107] J. Golbeck and D. Hansen. A method for computing political preference among twitter followers. *Social Networks*, 2013.

[108] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[109] S. Goyal and S. Joshi. Networks of collaboration in oligopoly. *Games and Economic behavior*, 43(1):57–85, 2003.

[110] M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

[111] T. Gross and B. Blasius. Adaptive coevolutionary networks: a review. *Journal of the Royal Society Interface*, 5(20):259–271, 2008.

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

[113] K. Hampton, L. S. Goulet, L. Rainie, and K. Purcell. Social networking sites and our lives. *Retrieved July 12, 2011 from*, 2011.

[114] H. L. A. Hart. Positivism and the separation of law and morals. *Harvard law review*, pages 593–629, 1958.

[115] S. Harvey. Rosie the riveter: Real women workers in world war ii. In *Journeys & Crossings, Library of Congress. http://www. loc. gov/rr/program/journey/rosie-transcript. html (accessed September 15, 2012)*, 2007.

[116] A. Hashmi, F. Zaidi, A. Sallaberry, and T. Mehmood. Are all social networks structurally similar? In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 310–314. IEEE Computer Society, 2012.

[117] M. T. Heaney. Multiplex networks and interest group influence reputation: An exponential random graph model. *Social Networks*, 2012.

[118] R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[119] F. Herrera, E. Herrera-Viedma, et al. A model of consensus in group decision making under linguistic assessments. *Fuzzy sets and Systems*, 78(1):73–87, 1996.

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

[121] A. O. I. Hoffmann, W. Jager, and J. H. Von Eije. Social simulation of stock markets: Taking it to the next level. *Journal of Artificial Societies & Social Simulation*, 10(2), 2007.

[122] R. A. Holley and T. M. Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, pages 643–663, 1975.

[123] P. Holme and B. J. Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65(2):026107, 2002.

[124] J. A. Hołyst, K. Kacperski, and F. Schweitzer. Phase transitions in social impact models of opinion formation. *Physica A: Statistical Mechanics and its Applications*, 285(1):199–210, 2000.

[125] T. Hossmann, F. Legendre, G. Nomikos, and T. Spyropoulos. Stumbl: Using facebook to collect rich datasets for opportunistic networking research. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE International Symposium on a*, pages 1–6. IEEE, 2011.

[126] B. A. Huberman and L. A. Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131–131, 1999.

[127] N. E. Humphries, N. Queiroz, J. R. Dyer, N. G. Pade, M. K. Musyl, K. M. Schaefer, D. W. Fuller, J. M. Brunnschweiler, T. K. Doyle, J. D. Houghton, et al. Environmental context explains lévy and brownian movement patterns of marine predators. *Nature*, 465(7301):1066–1069, 2010.

[128] O. Hussain, Z. Anwar, S. Saleem, F. Zaidi, et al. Empirical analysis of seed selection criterion in influence mining for different classes of networks. *ASONAM*, pages 1–8, 2013.

[129] A. Iovanovici, A. Topirceanu, M. Udrescu, and M. Vladutiu. Design space exploration for optimizing wireless sensor networks using social network analysis. In *System Theory, Control and Computing (ICSTCC), 2014 18th International Conference*, pages 815–820. IEEE, 2014.

[130] M. Ishizuka and M. Aida. Performance study of node placement in sensor networks. In *Distributed Computing Systems Workshops, 2004. Proceedings. 24th International Conference on*, pages 598–603. IEEE, 2004.

[131] M. O. Jackson. An overview of social networks and economic applications. *The handbook of social economics*, 1:511–85, 2010.

[132] M. Jacomy, S. Heymann, T. Venturini, and M. Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization. *Medialab center of research*, 2011.

[133] M. A. Javarone and T. Squartini. Conformism-driven phases of opinion formation on heterogeneous networks: the q-voter model case. *ArXiv e-prints*, Oct. 2014.

[134] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[135] L. Jian-Guo, D. Yan-Zhong, and W. Zhong-Tuo. Multistage random growing small-world networks with power-law degree distribution. *Chinese Physics Letters*, 23(3):746, 2006.

[136] B. Jiang. Street hierarchies: a minority of streets account for a majority of traffic flow. *International Journal of Geographical Information Science*, 23(8):1033–1048, 2009.

[137] B. Jiang, Y. Duan, F. Lu, T. Yang, and J. Zhao. Topological structure of urban street networks from the perspective of degree correlations. *arXiv preprint arXiv:1308.1533*, 2013.

[138] S. Johnson, J. J. Torres, J. Marro, and M. A. Munoz. Entropic origin of disassortativity in complex networks. *Physical review letters*, 104(10):108702, 2010.

[139] O. Juhlin. Traffic behaviour as social interaction-implications for the design of artificial drivers. In *PROCEEDINGS OF 6TH WORLD CONGRESS ON INTELLIGENT TRANSPORT SYSTEMS (ITS), HELD TORONTO, CANADA, NOVEMBER 8-12, 1999*, 1999.

[140] D. Jurcevic, Z. Shaman, and V. Krishnan. A new category: Very severe obstructive sleep apnea has worse outcomes on morbidity and mortality. *Chest*, 142(4_MeetingAbstracts): 1075A–1075A, 2012.

[141] B. Kantarci and V. Labatut. Classification of complex networks based on topological properties. *Proceedings of the 3rd International Conference on Social Computing and Its Applications*, 2013.

[142] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[143] A. Klaus, S. Yu, and D. Plenz. Statistical analyses support power law distributions found in neuronal avalanches. 2011.

[144] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[145] P. Klimek and S. Thurner. Triadic closure dynamics drives scaling laws in social multiplex networks. *New Journal of Physics*, 15(6):063008, 2013.

[146] D. Knoke and S. Yang. *Social network analysis*, volume 154. Sage, 2008.

[147] D. E. Knuth. *The Stanford GraphBase: a platform for combinatorial computing*, volume 37. Addison-Wesley Reading, 1993.

[148] D. Krackhardt. The strength of strong ties: The importance of philos in organizations. *Networks and organizations: Structure, form, and action*, 216:239, 1992.

[149] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

[150] J. Kunegis, D. Fay, and C. 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.

[151] R. Lambiotte, J.-C. Delvenne, and M. Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008.

[152] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[153] A. N. Langville and C. D. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[154] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[155] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 251–260. ACM, 2012.

[156] E. Leicht, P. Holme, and M. Newman. Vertex similarity in networks. *Physical Review E*, 73(2): 026120, 2006.

[157] J. Leskovec. Stanford large network dataset collection. *URL http://snap. stanford. edu/data/index. html*, 2011.

[158] J. Leskovec and J. J. Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[159] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.

[160] L. Leydesdorff. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9): 1303–1319, 2007.

[161] L. Li, A. Scaglione, A. Swami, and Q. Zhao. Trust, opinion diffusion and radicalization in social networks. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 691–695. IEEE, 2011.

[162] L. Li, A. Scaglione, A. Swami, and Q. Zhao. Phase transition in opinion diffusion in social networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3073–3076. IEEE, 2012.

[163] L. E. Li and P. Sinha. Throughput and energy efficiency in topology-controlled multi-hop wireless sensor networks. In *Proceedings of the 2nd ACM international conference on Wireless sensor networks and applications*, pages 132–140. ACM, 2003.

[164] Q. Li, L. A. Braunstein, S. Havlin, and H. E. Stanley. Strategy of competition between two groups based on an inflexible contrarian opinion model. *Physical Review E*, 84(6):066101, 2011.

[165] W. Li and J.-Y. Yang. Comparing networks from a data analysis perspective. In *Complex Sciences*, pages 1907–1916. Springer, 2009.

[166] Y. Li, X. Qian, and D. Wang. Extended hk evolving network model. In *Control and Decision Conference (CCDC), 2012 24th Chinese*, pages 4095–4097. IEEE, 2012.

[167] Y. Li, W. Chen, Y. Wang, and Z.-L. Zhang. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 657–666. ACM, 2013.

[168] X. Liu and M. Haenggi. Toward quasiregular sensor networks: Topology control algorithms for improved energy efficiency. *Parallel and Distributed Systems, IEEE Transactions on*, 17(9): 975–986, 2006.

[169] G. Long and X. Cai. The fractal dimensions of complex networks. *Chinese Physics Letters*, 26 (8):088901, 2009.

[170] J. Loscalzo and A.-L. Barabasi. Systems biology and the future of medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(6):619–627, 2011.

[171] R. D. Luce and A. D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

[172] I. Lunden. 73popularity, facebook stays on top. *techcrunch.com*, 2013.

[173] M. Magnani and L. Rossi. The ml-model for multi-layer social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 5–12. IEEE, 2011.

[174] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[175] B. B. Mandelbrot. *The fractal geometry of nature*. Macmillan, 1983.

[176] L. Manovich. Trending: the promises and the challenges of big social data. *Debates in the digital humanities*, pages 460–475, 2011.

[177] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[178] A. Masoudi-Nejad, F. Schreiber, and Z. Kashani. Building blocks of biological networks: a review on major network motif discovery algorithms. *IET systems biology*, 6(5):164–174, 2012.

[179] N. Masuda. Voter models with contrarian agents. *Physical Review E*, 88(5):052803, 2013.

[180] J. C. Maxwell. *Developing the leader within you*. Thomas Nelson Publishers, 1993.

[181] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton. Big data. *The management revolution. Harvard Bus Rev*, 90(10):61–67, 2012.

[182] M. McDonald and H. Wilson. *Marketing plans: How to prepare them, how to use them*. John Wiley & Sons, 2011.

[183] W. McNicholas, M. Bonsignore, et al. Sleep apnoea as an independent risk factor for cardiovascular disease: current evidence, basic mechanisms and research priorities. *European Respiratory Journal*, 29(1):156–178, 2007.

[184] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[185] S. G. Memtsoudis, M. C. Besculides, and M. Mazumdar. A rude awakening - the perioperative sleep apnea epidemic. *N Engl J Med*, 368(25):2352–2353, 2013.

[186] S. Mihaicuta, R. Avram, A. Topirceanu, and M. Udrescu. A network-based approach to sleep apnea syndrome. *European Respiratory Journal*, 42(Suppl 57):P2046, 2013.

[187] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[188] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

[189] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[190] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[191] M. Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

[192] W. G. Moons, D. J. Leonard, D. M. Mackie, and E. R. Smith. I feel our pain: Antecedents and consequences of emotional self-stereotyping. *Journal of Experimental Social Psychology*, 45(4): 760–769, 2009.

[193] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[194] T. Nakada, Y. Kato, and S. Kunifuji. A study on the dynamics of friendship network formation using a directed network model. 2007.

[195] S. Narayanan. *The betweenness centrality of biological networks*. PhD thesis, Citeseer, 2005.

[196] A. Newell, H. A. Simon, et al. *Human problem solving*, volume 104. Prentice-Hall Englewood Cliffs, NJ, 1972.

[197] M. Newman. *Networks: an introduction*. Oxford University Press, 2010.

[198] M. Newman, A.-L. Barabási, and D. J. Watts. *The structure and dynamics of networks*. Princeton University Press, 2006.

[199] M. E. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, 64(1):016131, 2001.

[200] M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[201] M. E. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

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

[203] M. E. Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

[204] M. E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5): 323–351, 2005.

[205] M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[206] M. E. Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2 (2008):1–12, 2008.

[207] A. Noack. Modularity clustering is force-directed layout. *Physical Review E*, 79(2):026102, 2009.

[208] B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[209] J. G. Oliveira and A.-L. Barabási. Human dynamics: Darwin and einstein correspondence patterns. *Nature*, 437(7063):1251–1251, 2005.

[210] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis. Geographic constraints on social network groups. *PLoS one*, 6(4):e16939, 2011.

[211] T. Opsahl and P. Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009.

[212] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[213] G. Parati, C. Lombardi, J. Hedner, M. R. Bonsignore, L. Grote, R. Tkacova, P. Levy, R. Riha, C. Bassetti, K. Narkiewicz, et al. Position paper on the management of patients with obstructive sleep apnea and hypertension: Joint recommendations by the european society of hypertension, by the european respiratory society and by the members of european cost (cooperation in scientific and technological research) action b26 on obstructive sleep apnea. *Journal of hypertension*, 30(4):633–646, 2012.

[214] P. Parigi and L. Sartori. The political party as a network of cleavages: Disclosing the inner structure of italian political parties in the seventies. *Social Networks*, 2012.

[215] K. Park, Y. Han, and Y.-K. Lee. An efficient method for computing similarity between frequent subgraphs. *Proceedings of the 3rd International Conference on Social Computing and Its Applications*, 2013.

[216] M. Q. Pasta, Z. Jan, A. Sallaberry, and F. Zaidi. Tunable and growing network generation model with community structures. In *Cloud and Green Computing (CGC), 2013 Third International Conference on*, pages 233–240. IEEE, 2013.

[217] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[218] N. Pelletier-Fleury, N. Meslier, F. Gagnadoux, C. Person, D. Rakotonanahary, H. Ouksel, B. Fleury, and J. Racineux. Economic arguments for the immediate management of moderate-to-severe obstructive sleep apnoea syndrome. *European Respiratory Journal*, 23(1):53–60, 2004.

[219] A. Pentland. Reinventing society in the wake of big data. *Edge. Available online at: http://www. edge. org/conversation/reinventing-society-in-the-wake-of-big-data*, 2012.

[220] M. Piraveenan, M. Prokopenko, and A. Zomaya. Local assortativeness in scale-free networks. *EPL (Europhysics Letters)*, 84(2):28002, 2008.

[221] S. Plous. *The psychology of judgment and decision making*. Mcgraw-Hill Book Company, 1993.

[222] W. Y. Poe and J. B. Schmitt. Sink placement without location information in large-scale wireless sensor networks. In *Asian Internet Engineering Conference*, pages 69–76. ACM, 2009.

[223] S. Porta, P. Crucitti, and V. Latora. The network analysis of urban streets: a dual approach. *Physica A: Statistical Mechanics and its Applications*, 369(2):853–866, 2006.

[224] N. M. Punjabi. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, 5(2):136, 2008.

[225] W. Quattrociocchi, G. Caldarelli, and A. Scala. Opinion dynamics on interacting networks: media competition and social influence. *Scientific reports*, 4, 2014.

[226] O. F. Rana, A. Akram, and S. J. Lynden. Building scalable virtual communities: infrastructure requirements and computational costs. In *Socionics*, pages 68–83. Springer, 2005.

[227] O. Reichman, M. B. Jones, and M. P. Schildhauer. Challenges and opportunities of open data in ecology. *Science*, 331(6018), 2011.

[228] B. Rieder. Studying facebook via data extraction: the netvizz application. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 346–355. ACM, 2013.

[229] R. L. Riolo, M. D. Cohen, and R. Axelrod. Evolution of cooperation without reciprocity. *Nature*, 414(6862):441–443, 2001.

[230] S. Roccas and A. Amit. Group heterogeneity and tolerance: The moderating role of conservation values. *Journal of Experimental Social Psychology*, 47(5):898–907, 2011.

[231] E. M. Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[232] V. A. Rossi, J. R. Stradling, and M. Kohler. Effects of obstructive sleep apnoea on heart rhythm. *European Respiratory Journal*, 41(6):1439–1451, 2013.

[233] Z. Ruan, G. Iniguez, M. Karsai, and J. Kertesz. Kinetics of social contagion. *arXiv preprint arXiv:1506.00251*, 2015.

[234] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda. Learning diffusion probability based on node attributes in social networks. In *Foundations of Intelligent Systems*, pages 153–162. Springer, 2011.

[235] M. Sánchez-de-la Torre, F. Campos-Rodriguez, and F. Barbé. Obstructive sleep apnoea and cardiovascular disease. *The Lancet Respiratory Medicine*, 1(1):61–72, 2013.

[236] C. Scholz, M. Atzmueller, M. Kibanov, and G. Stumme. Predictability of evolving contacts and triadic closure in human face-to-face proximity networks. *Social Network Analysis and Mining*, 4(1):1–17, 2014.

[237] V. Schwämmle, A. Auto-Moreira, J. Andrade, M. Gonz'alez, and H. Herrmann. The spread of opinions in a model with different topologies. Technical report, 2007.

[238] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[239] S. K. Sharma, S. Agrawal, D. Damodaran, V. Sreenivas, T. Kadhiravan, R. Lakshmy, P. Jagia, and A. Kumar. Cpap for the metabolic syndrome in patients with obstructive sleep apnea. *New England Journal of Medicine*, 365(24):2277–2286, 2011.

[240] S. Simon and N. Collop. Latest advances in sleep medicinelatest advances in obstructive sleep apneaobstructive sleep apnea. *CHEST Journal*, 142(6):1645–1651, 2012.

[241] J. M. Smith, D. S. Halgin, V. Kidwell-Lopez, G. Labianca, D. J. Brass, and S. P. Borgatti. Power in politically charged networks. *Social Networks*, 2013.

[242] C. Song, S. Havlin, and H. A. Makse. Self-similarity of complex networks. *Nature*, 433(7024): 392–395, 2005.

[243] C. Song, L. K. Gallos, S. Havlin, and H. A. Makse. How to calculate the fractal dimension of a complex network: the box covering algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(03):P03006, 2007.

[244] E. Spertus, M. Sahami, and O. Buyukkokten. Evaluating similarity measures: a large-scale study in the orkut social network. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 678–684. ACM, 2005.

[245] S. M. Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, 4(2): 73–79, 1989.

[246] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[247] L. Suciu, C. Cristescu, A. Topîrceanu, L. Udrescu, M. Udrescu, V. Buda, and M. Tomescu. Evaluation of patients diagnosed with essential arterial hypertension through network analysis. *Irish Journal of Medical Science (1971-)*, pages 1–9, 2015.

[248] K. Sznajd-Weron and J. Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, 2000.

[249] P.-N. Tan et al. *Introduction to data mining*. Pearson Education India, 2007.

[250] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic. Recipe recommendation using ingredient networks. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 298–307. ACM, 2012.

[251] R. Toivonen, J.-P. Onnela, J. Saramäki, J. Hyvönen, and K. Kaski. A model for social networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):851–860, 2006.

[252] M. Tomko, S. Winter, and C. Claramunt. Experiential hierarchies of streets. *Computers, Environment and Urban Systems*, 32(1):41–52, 2008.

[253] A. Topirceanu and M. Udrescu. SocialSim: A framework for opinion dynamics simulations, September 2014. URL `http://cs.upt.ro/~alext/socialsim`.

[254] A. Topirceanu and M. Udrescu. Fmnet: Physical trait patterns in the fashion world. In *Network Intelligence Conference (ENIC), 2015 Second European*, pages 25–32. IEEE, 2015.

[255] A. Topirceanu and M. Udrescu. Measuring realism of social network models using network motifs. In *Applied Computational Intelligence and Informatics (SACI), 2015 IEEE 10th Jubilee International Symposium on*, pages 443–447. IEEE, 2015.

[256] A. Topirceanu and M. Udrescu. Statistical fidelity: A tool to quantify the similarity between multi-variable entities with application in complex networks. *International Journal of Computer Mathematics*, 2016.

[257] A. Topirceanu, M. Udrescu, and M. Vladutiu. Network fidelity: A metric to quantify the similarity and realism of complex networks. In *Cloud and Green Computing (CGC), 2013 Third International Conference on*, pages 289–296. IEEE, 2013.

[258] A. Topirceanu, G. Barina, and M. Udrescu. Musenet: Collaboration in the music artists industry. In *Network Intelligence Conference (ENIC), 2014 European*, pages 89–94. IEEE, 2014.

[259] A. Topirceanu, A. Iovanovici, M. Udrescu, and M. Vladutiu. Social cities: Quality assessment of road infrastructures using a network motif approach. In *System Theory, Control and Computing (ICSTCC), 2014 18th International Conference*, pages 803–808. IEEE, 2014.

[260] A. Topirceanu, M. Udrescu, and M. Vladutiu. Genetically optimized realistic social network topology inspired by facebook. In *Online Social Media Analysis and Visualization*, pages 163–179. Springer, 2014.

[261] A. Topirceanu, A. Duma, and M. Udrescu. Uncovering the fingerprint of online social networks using a network motif based approach. *Computer Communications*, 73:167–175, 2016.

[262] A. Topirceanu, M. Udrescu, M. Vladutiu, and R. Marculescu. Tolerance-based interaction: A new model targeting opinion formation and diffusion in social networks. *PeerJ Computer Science*, 2:e42, 2016.

[263] M. Tsvetovat and K. M. Carley. Generation of realistic social network datasets for testing of analysis and simulation tools. Technical report, DTIC Document, 2005.

[264] M. Udrescu and A. Topirceanu. What drives the emergence of social networks? In *Control Systems and Computer Science (CSCS), 2015 20th International Conference on*, pages 999–999. IEEE, 2015.

[265] M. Udrescu, A. Topirceanu, R. Avram, and S. Mihaicuta. Aer score: A social-network-inspired predictor for sleep apnea syndrome. *CHEST Journal*, 145:609A–609A, 2014.

[266] K. T. Utriainen, J. K. Airaksinen, O. Polo, O. T. Raitakari, M. J. Pietilä, H. Scheinin, H. Y. Helenius, K. A. Leino, E. S. Kentala, J. R. Jalonen, et al. Unrecognised obstructive sleep apnoea is common in severe peripheral arterial disease. *European Respiratory Journal*, 41(3):616–620, 2013.

[267] T. W. Valente, K. Fujimoto, J. B. Unger, D. W. Soto, and D. Meeker. Variations in network boundary and type: A study of adolescent peer influences. *Social Networks*, 2013.

[268] S. Valenzuela, N. Park, and K. F. Kee. Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation1. *Journal of Computer-Mediated Communication*, 14(4):875–901, 2009.

[269] J. Van Der Schalk, A. Fischer, B. Doosje, D. Wigboldus, S. Hawk, M. Rotteveel, and U. Hess. Convergent and divergent responses to emotional displays of ingroup and outgroup. *Emotion*, 11(2):286, 2011.

[270] M. Vidal, M. E. Cusick, and A.-L. Barabasi. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

[271] B. Viswanath, A. Mislove, M. Cha, and K. 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.

[272] C. S. Wagner and L. Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research policy*, 34(10):1608–1618, 2005.

[273] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

[274] J. Wang and L. Rong. Evolving small-world networks based on the modified ba model. In *Computer Science and Information Technology, 2008. ICCSIT'08. International Conference on*, pages 143–146. IEEE, 2008.

[275] P. Wang, J. Lü, and X. Yu. Identification of important nodes in directed biological networks: A network motif approach. *PloS one*, 9(8):e106132, 2014.

[276] X. F. Wang and G. Chen. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE*, 3(1):6–20, 2003.

[277] S. Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[278] S. Wasserman and J. Galaskiewicz. *Advances in social network analysis: Research in the social and behavioral sciences*, volume 171. Sage Publications, 1994.

[279] D. J. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton university press, 1999.

[280] D. J. Watts. *Six degrees: The science of a connected age*. WW Norton & Company, 2004.

[281] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *nature*, 393 (6684):440–442, 1998.

[282] W. Weidlich. Sociodynamics-a systematic approach to mathematical modelling in the social sciences. *NONLINEAR PHENOMENA IN COMPLEX SYSTEMS-MINSK-*, 5(4):479–487, 2002.

[283] S. Wernicke. Efficient detection of network motifs. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 3(4):347–359, 2006.

[284] S. Wernicke and F. Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[285] P. D. Windschitl, J. P. Rose, M. T. Stalkfleet, and A. R. Smith. Are people excessive or judicious in their egocentrism? a modeling approach to understanding bias and accuracy in people's optimism. *Journal of personality and social psychology*, 95(2):253, 2008.

[286] S. Wuchty, Z. N. Oltvai, and A.-L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature genetics*, 35(2):176–179, 2003.

[287] W. Xiao-Yan and L. Zong-Hua. Epidemic diffusion on complex networks. *Chinese Physics Letters*, 24(4):1118, 2007.

[288] K. Xu, H. S. Hassanein, G. Takahara, and Q. Wang. Relay node deployment strategies in heterogeneous wireless sensor networks: multiple-hop communication case. In *SECON*, volume 5, pages 575–585, 2005.

[289] H. K. Yaggi, J. Concato, W. N. Kernan, J. H. Lichtman, L. M. Brass, and V. Mohsenin. Obstructive sleep apnea as a risk factor for stroke and death. *New England Journal of Medicine*, 353(19):2034–2041, 2005.

[290] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.

[291] E. Yildiz, D. Acemoglu, A. Ozdaglar, A. Saberi, and A. Scaglione. Discrete opinion dynamics with stubborn agents. *Available at SSRN 1744113*, 2011.

[292] E. Yildiz, A. Ozdaglar, D. Acemoglu, A. Saberi, and A. Scaglione. Binary opinion dynamics with stubborn agents. *ACM Transactions on Economics and Computation*, 1(4):19, 2013.

[293] T. Young, P. E. Peppard, and D. J. Gottlieb. Epidemiology of obstructive sleep apnea: a population health perspective. *American journal of respiratory and critical care medicine*, 165(9): 1217–1239, 2002.

[294] M. Younis and K. Akkaya. Strategies and techniques for node placement in wireless sensor networks: A survey. *Ad Hoc Networks*, 6(4):621–655, 2008.

[295] G. U. Yule. A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, pages 21–87, 1925.

[296] F. Zaidi. Small world networks and clustered small world networks with random connectivity. *Social Network Analysis and Mining*, pages 1–13, 2013.

[297] D. Zhou, H. E. Stanley, G. DAgostino, and A. Scala. Assortativity decreases the robustness of interdependent networks. *Physical Review E*, 86(6):066103, 2012.