UNIVERSITATEA POLITEHNICA TIMIŞOARA

Bridging the Gap Between Computational Network Science and Computer Engineering

Habilitation Thesis

Alexandru TOPÎRCEANU, PhD

2021

Acknowledgments

I remember myself being fascinated about the world of computers and engineering as early as my childhood. While my Bachelor and Master studies were barely discovering the tip of the iceberg – that was computer engineering and information technology at that time for me – my intuition told me to follow a path of innovation and breakthroughs by pursuing a PhD in this field. As such, I first received a warm welcome in the ACSA (Advanced Computing Systems and Architectures) research group, in the Department of Computer and Information Technology within Politehnica University Timisoara (UPT). The ACSA group was founded and led, back in 2012–2016, by Prof. Mircea Vladutiu, who also became my PhD advisor. I wish to thank him wholeheartedly, as well as Prof. Mihai Udrescu for guiding me along a path that would define my academic career. Mihai Udrescu, currently a good friend and collaborator, became a mentor for me in the field of Network Science, which I discovered during my late Master studies (2011). I quickly became fascinated by the interdisciplinarity of Network Science, with its roots embedded in computers, mathematics, physics, but also stretching over to sociology, biology, ecology, industry etc. Given the diverse interests of the ACSA group, I strengthened collaborations and obtained original contributions by bridging the computational network science approach with other topics of research – from engineering to sleep research, educational sciences, and pharmacology. Along the journey, I wish to thank several colleagues from ACSA: Lucian Prodan, Alexandru Iovanovici, Flavius Opritoiu, and former PhD/Master's collaborators: Gabriel Barina, Razvan Avram, Alexandra Duma. Additionally, I wish to thank close collaborators which I've met along my years of research, like Dr. Stefan Dan Mihaicuta et al. from the University of Medicine and Pharmacy of Timisoara (UMFT), Lucretia Udrescu et al. from the Faculty of Pharmacy at UMFT, Prof. Radu Marculescu from Carnegie Mellon University (former)/ University of Texas (current), Paul Bogdan from the University of Southern California, Gabriela Grosseck and Silvia Fierascu from West University Timisoara, and Ovidiu Sirbu from the Biochemistry department of UMFT. Overall, I am very lucky to have collaborated with specialists in diverse fields of research, witnessing the power of Network Science in fields that seemed out of reach for a computer engineer just one or two decades ago. All of these achievements would not have been so easy if it weren't for those who financed our research: UEFISCDI, Linde Healthcare and Horizon 2020.

Finally, I wish to conclude my acknowledgments by thanking my dear family for their support and encouragements along my career. My wife Mihaela, my parents Adrian and Rodica, my aunt Ligia, and my grandmother Viorica at 85, keep ensuring that my dynamical lifestyle, with endless days filled by research, is balanced with a rich emotional and spiritual family life.

Contents

Part I

Contributions

Chapter 1

Introduction

With the dawn of the 21^{st} Century, the novel field of *Network Science*, or the *Science of Complex Networks*, has surfaced and has gathered momentum to produce a prominent body of new multidisciplinary research. Network Science is an interdisciplinary field by nature, with roots embedded in mathematics, physics, statistics, computer science and information technology. Its applicability, however, stretches over to encompass the biological, pharmacological, social, economical, political sciences altogether.

From a historical standpoint, the study of networks has sparked the creativity of scientists for several centuries. Many consider Leonhard Euler as the grounding father of graph theory, back in the 18^{th} Century, who solved the problem of the seven bridges in Konigsberg. The Romanian-born psychologist Jacob Moreno introduced the sociogram – a methodology the to model relationships between individuals. Also noteworthy, is the Hungarian mathematician Denes Konig who formalized graph theory in the 20^{th} Century. However, it is not until the late 1950's when Alfred Renyi and Paul Erdos formalized the random graph model [1], which became the first conceptual complex network, possessing many of the properties found in Big Data today. The next major scientific leap is represented by two parallel joint efforts by Duncan Watts and Steven Strogatz, respectively Albert-Laszlo Barabasi and Reka Albert who uncovered the fundamental properties of *small-world*-ness [2] and *scale-free*-ness [3] that still drive our current understanding of the structure and dynamics of complex networks [4].

The formalization of complex Network Science – its scientific boundaries, methodologies, taxonomy, and real-world applications – have been gradually defined over the last two decades, based on both formalism and data-driven findings [5]. In particular, Network Science uses graphs of non-trivial size and non-trivial structure to model the structure and function of natural or man-made phenomena. The building blocks of these graphs are nodes (vertices) – e.g., representing entities like individuals, routers, genes, patients – and the relationship between nodes represented as edges (links) – e.g., representing friendships, acquaintances, cables, chemical interactions, binding forces. Complex network theory differentiates itself from classic graph theory through the size (i.e., often dealing with millions of nodes), and structure (i.e., dealing with heterogeneous topologies instead of regular ones) of complex networks. Furthermore, the dynamics enabled by such large networks are of higher complexity, requiring computer simulation rather than analytical approaches to understand and predict [6, 7].

Complex networks science covers an active area of scientific research inspired largely by

the empirical study of real-world networks with applicability in computer engineering and communication, computer science, health and pharmacological sciences, economics, politics, and even warfare. Given the diverse applicability of this emerging field, we note four major domains of interest driving the current state of the art:

- Biological networks dealing with the study of e.g., metabolic networks, transcription regulatory networks, protein-protein interaction networks, protein structure networks, neural networks, ecological networks, natural food chains [8, 9, 10];
- Social networks dealing with the study of e.g., friendship networks, citation networks, voter networks, world markets, political structures [11, 12, 13, 14];
- Technological networks dealing with the study of e.g., computer networks, the WWW, electrical circuits, road networks, power grids [15, 6, 16];
- Semantic networks dealing with the study of e.g., word nets, recipe networks, software projects structure [17, 18, 19].

The interdisciplinary nature of Network Science, found at the crossroads of Computer Science and Engineering, Mathematics and Physics, is detailed in Figure 1.1. Today, Network Science research can be classified into one or more of the following fundamental topics: social network analysis (SNA), biological network analysis, multilayer networks, dynamic network analysis, link (prediction) analysis, and centrality (influence) analysis.

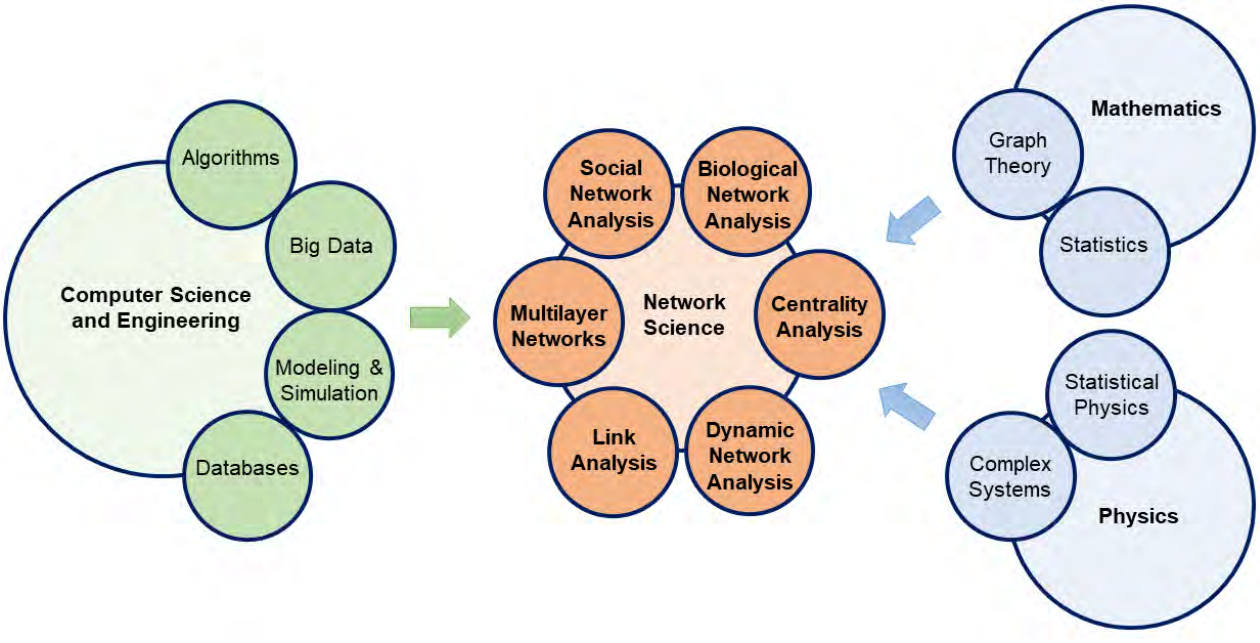

Figure 1.1: The interdisciplinary nature of Network Science found at the crossroads of Computer Science and Engineering (Algorithms, Big Data, Modeling & simulation, Databases), with Mathematics (Graph Theory, Statistics), and Physics (Statistical Physics, Complex Systems).

Finally, it is worth mentioning that many of the methodologies, involving large amounts of data and complex systems, are supported by Computer Engineering & Information Technology [20]. For instance, Information Technology offers approaches where computer algorithms, simulation tools and databases are used for the processing and understanding of biological, medical, pharmacological, or social data. Recent developments in personalized education and medicine are based on Big Data Analytics & Visualization [21], and computer science technologies such as Complex Network Analysis (CNA) and Machine Learning. Also, advancements in online social network technologies have enabled social physics and computational epidemics with global scale socio-economical impact [22].

1.1 Motivation

The main motivation of the research presented in this thesis, spanning from 2011–2021, is to bridge the main field of Computers & Information Technology with the multidisciplinary field of Network Science with real-world, impactful applicability. This fusion results in the newly coined term of *Computational Network Science* [23, 24], motivated by the fact that Computer Science and Engineering can change all other sciences through its data-driven approach [24].

In support of this recent trend, we find new journals being dedicated to Computational Network Science as well as Computational Social Science, a sister field focusing solely on social networks. Nature Publishing Group has launched a new journal named *Nature Computational Science* in 2021 encouraging the development and use of computational procedures and mathematical models, including their utilization to solve complex problems across various scientific disciplines. Also, IEEE has launched the journal *IEEE Transactions on Computational Social Systems* addressing the modeling, analysis, simulation and understanding of social systems from a quantitative or computational perspective.

When analyzing the list of the most prominent network scientists, we find these being divided across the following fields:

- Social and behavioral sciences (e.g., J. Fowler, M. Granovetter, M.O. Jackson, D. Lazer, D. Watts);
- Computers & Information sciences (e.g., L. Adamic, A. Clauset, J. Kleinberg, J. Leskovec, F. Menczer);
- Physics (e.g., A-L. Barabasi, S. Havlin, Y. Moreno, M. Newman, A. Vespignani);
- Biology (e.g., U. Alon, D. Bassett);
- Mathematics (e.g., L. Lovasz, S. Strogatz).

Among other important personalities, considered within the top 10 most powerful data scientists in the world, we enumerate: Larry Page (CEO, Google), Sebastian Thrun (Professor, Stanford University), Todd Park (CTO, Department of Health and Human Services of the USA), Alex "Sandy" Pentland (Professor, MIT). We conclude that Computational Network Science is a very dynamic research field, targeting many relevant social, technological and economical challenges in the present. Within this cross-disciplinary set of methods and

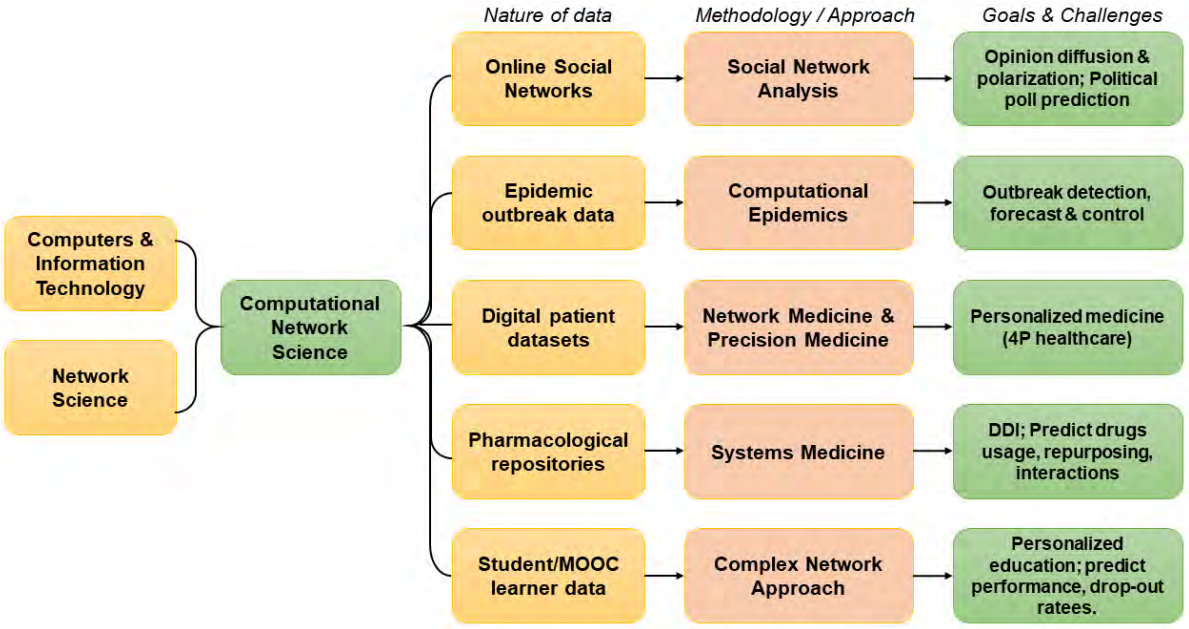

Figure 1.2: Main research paths presented in this thesis focusing on the cross-disciplinary use of Computers & Information Technology with Network Science represented by five distinct tracks backed up by a Computational Network Science approach. Each of the five tracks detail the nature of the data being used for modeling (*Nature of data*), the specific approach suited for each type of data (*Methodology/ Approach*), and the applicability of the results (*Goals & Challenges*).

applications, Computers & Information Technology have a fundamental and distinguished role.

In light of these premises, we started research in the field of Network Science in Politehnica University Timisoara around the year 2011 (from zero) when this novel concept was introduced to our research group by Prof. Mihai Udrescu, after a research visit and Carnegie Mellon University (USA). During my PhD studies (2012–2016) we published the first scientific results using a novel computational network approach, culminating with my PhD obtained in Computers. The successful defense of my PhD in February 2016 has laid the first bricks of a new School of Network Science within the Polytechnic context, and possibly the whole Romanian academic context.

Since 2012, our research has extended along multiple tracks, as summarized by Figure 1.2. Our first important approach is to study the generation of realistic complex network topologies, their growth in time, followed by the opinion diffusion over large evolving networks. For this, we analyze state of the art topological models, inspired by the small world [2] and scale free [3] networks, and aiming to reproduce empirically observed properties of real-world networks. To this end, we create a highly realistic social network model [25] using a multi-variable genetic algorithm approach. Also, in contrast to the fundamental Degree preferential attachment principle, advocated by Barabasi et al. [26], we further proposed the concept of Betweenness preferential attachment for better explaining the growth of social networks [27]. Subsequently, we studied opinion spreading models using discrete and continuous opinion [28, 29, 30], including ones which include stubborn agents [31, 32]. Given the complex interplay of agent nodes, we chose computer simulation as a valid research methodology to evaluate and quantify these opinion spreading models over large social networks, because an analytical approach is not able to handle the high complexity of social network interconnectivity. We find that, on small proof-of-concept networks, regular lattices or meshes as used for understanding dynamics on social networks [32]. Alternatively, random networks [1] are used to employ a fully statistical analysis [33]. However, in both cases, the possible quantitative and qualitative insights are limited. When applying the analytical power of computer modeling and simulation we show that our novel tolerance-based model for opinion interaction and spreading generates realistic, reproducible patterns in social networks [34]. We extend existing models [31, 32] by adding a tolerance parameter to each agent (node) which measures the degree of accepted influence by neighboring nodes' opinions. While several existing models have such trust parameters, we define tolerance as a time variable parameter dependent on the interaction patterns with neighboring nodes. Our inspiration for the evolution of tolerance derives from the idea that the dynamics towards tolerance and intolerance vary exponentially [35, 36], meaning that an agent under constant influence becomes indoctrinated at an increased rate over time. Along this track, we also introduced a statistical tool for measuring the structural similarity (fidelity) between any two complex networks [37]; modeled complex network antifragility under sustained attack [38]; and, introduced an original, reliable methodology for benchmarking node centrality measures in a competitive context [39]. There are over 50 different node centralities used for the selection of spreader nodes in networks, and our methodology can reliably determine the more efficient ones given a specific topology.

Another important approach detailed in this thesis is the application of Network Medicine (also Precision Medicine, Systems Medicine) providing computer-based solutions for medical

and pharmacological challenges [40]. On one hand, we worked for the past 8 years in sleep research, in order to offer computational solutions for predicting the severity and development of Obstructive sleep apnea (OSA) and Chronic obstructive pulmonary disease (COPD). The results are formed around a patient phenotype model [41], built through a dual clustering technique, and a score usable by doctors in every day patient monitoring [42]. More recently, we extended the state of the art in OSA severity monitoring by presenting a differentiated phenotype model for each gender [43], as well as analyzing the causes of improved CPAP treatment response [44]. All these results are aimed at developing personalized treatment and precise diagnostics (like 4P Healthcare). On the other hand, we use the public database *Drugbank* [45] in order to build a drug-drug interaction (DDI) network, where nodes represent drugs and links represent drug-drug interaction relationships between the drugs [46]. More recently, we have explored the potential of target based DDI [47], using the same dual clustering technique. These results help researchers estimate possible new interactions and repurposing alternatives for drugs, thus optimizing costly and time-consuming pharmacological studies.

Another undertaken track of interdisciplinary research has been that of adapting a complex network approach in the analysis of educational data. Specifically, we analyzed data of Romanian students/learners participating in MOOCs, and created a compatibility network of such students, based on their motivations, expectations and perceived difficulties throughout the courses. By applying clustering techniques on the network, we defined specific student archetypes (i.e., corresponding to communities) [48]. Furthermore, we developed a fully original gamification platform for student motivation in class [49], and conducted an exam cheating study [50].

Finally, we mention two more recent research tracks (2018-present), that of computational epidemics and political poll prediction. Given the COVID-19 pandemic, we were quickly motivated to collaborate (with the University of Texas) and focus our efforts on modeling and better understanding the impact of this outbreak. Ongoing research is under development on two tracks: that of understanding the impact of isolation strategies adopted in early 2020 [51], and improving the underlying population model used for epidemic simulations [52]. Also, given the turmoil caused by local and global elections over the past years, we focused on improving the accuracy of pre-election polls using time series analysis and network science. We obtained encouraging results [53, 54, 54] compared to existing state of the art methods, like Multilevel regression with poststratification.

Overall, we have presented a detailed motivation of the research path followed in this thesis, during the period 2011-2021, with the focus on impactful applications in Computational Network Science, thus bridging the field of Computers & Information Technology with the multidisciplinary field of Network Science.

1.2 Research Path and Contributions

The author has graduated the Faculty of Automation and Computers, Politehnica University Timisoara (UPT) in 2010, obtaining a degree in Computer Engineering. Two years later, the author obtained his Master's degree following the "Software Engineering" program. Between October 2012– February 2016, the author developed his PhD thesis, titled "Structural and Behavioral Analysis and Modeling of the Society" obtaining *Excellent* honors. Both the Mas-

ter's (2012) and PhD (2016) represent important milestones for Social Network Analysis in our University, as they were the first titles obtained in the field of Computers offered to contributions in the multidisciplinary field of Network Science. Since then, the author, alongside the ACSA research group, has worked on embedding Network Science into a representative research track within the Department of Computers and Information Technology of UPT.

The main fields of expertise in which the author has contributed with relevant scientific activity (publications, projects) are:

- Social Network Analysis (SNA) – generation of network topologies using genetic algorithms [25, 37, 55]; modeling network growth based on the original concept of Betweenness Preferential Attachment [27], quantifying network antifragility under topological attack [38]; community analysis [56, 57, 58]; centrality analysis, benchmarking and spreader selection strategies [39].
- Computational Social Networks (CSN) – modeling and simulation of opinion diffusion [34, 59], political poll prediction using time series and introducing temporal attenuation [60, 54, 53].
- Network Medicine – sleep research for assessing OSA severity [42, 44] and phenotyping patients [41, 43]; drug-drug interaction analysis [46, 47].
- Educational Science – MOOC student archetyping [48], gamification [49], and network analysis of student collaborations during exam [50].

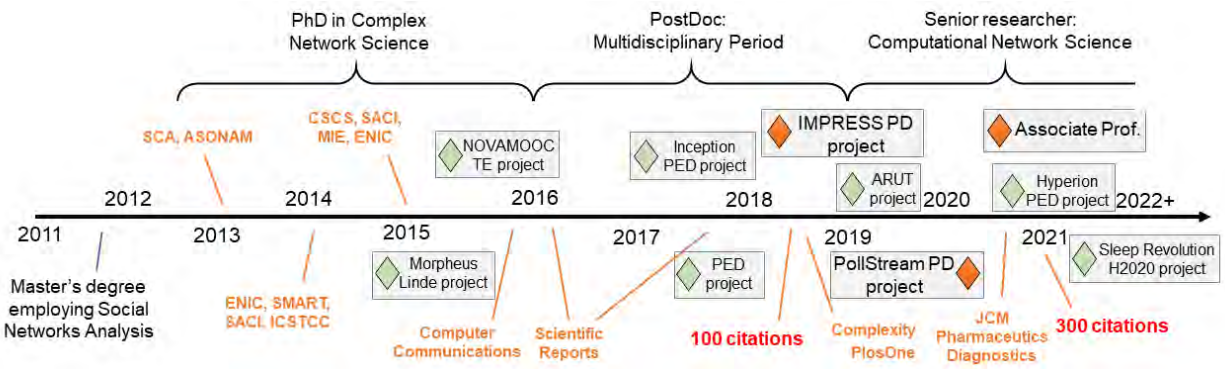

Figure 1.3: Timeline of the research path between 2011–2021. Diamonds highlight the most significant career events (i.e., director of projects with orange, member of projects with green). Important publication venues (e.g., conferences, journals) are highlighted with orange text and arrows, and citation milestones are depicted using red text.

These research tracks are supported by consistent scientific publications and projects (as director and team member, see Figure 1.3) with high impact. We summarize the overall results as follows:

- Director of 2 national research projects financed by UEFISCDI.
- Member of 2 international project teams (including one Horizon 2020 project), and member of 5 national project teams.

- Author of 2 books, and 5 international book chapters.
- Author of 53 Web of Science (WoS) indexed articles, out of which 16 journals, with 11 Q1 and 1 Q2 articles (first author in 8 of these journal articles).
- Cumulative impact factor of over 45, and journal impact factors in the range 1.05–4.53.
- H-index of 9 and 171 WoS citations[1]; H-index of 12 and 333 citations in Google Scholar[2].
- Reviewer for multiple journals and conferences, including Scientific Reports, Future Generation Computer Systems, Complexity, Mathematics, Online Social Media Analysis and Visualization, IEEE Transactions on Computational Social Systems, Physica A, ASONAM, ENIC etc.
- PC/EB member for Online Social Media Analysis and Visualization, European Network Intelligence Conference, International Conference on Engineering of Modern Electric Systems.
- One Best Paper Award at an IEEE conference (organized in Sweden, 2015).

In Figure 1.3 we represent the main scientific achievements of the author using a timeline from 2011–2021. During his PhD period (2012–2016) the author managed to publish a representative number of WoS proceedings papers at important venues in the field of Network Science (e.g., SCA, ASONAM, ENIC) and Computer Engineering & Information Technology (e.g., SACI, ICSTCC, CSCS, SMART). Also, he obtained his first journal publications (PeerJ CS, Computer Communications) and participated in two research projects as a team member. Later, between 2016–2019 the author managed to diversify his research tracks, published additional impactful journal papers (e.g., Scientific Reports, Complexity, Plos One), participated in two additional PED projects, and managed his own PD project (acronym IMPRESS). From 2019 until the present, the author was involved in multiple significant research projects (one ARUT, one PED, one Horizon2020), and is managing his second project as project director (acronym PollStream). Also during this period, the author surpassed the 300 citation milestone (in Google Scholar; 164 WoS citations), published additional journal papers of high impact (e.g., Journal of Clinical Medicine, Pharmaceutics, Diagnostics), and at important venues pertaining to Network Science (e.g., ASONAM, Complex Networks).

Our activity is supported by several national and international grants, including collaborations with the West University Timisoara and "Victor Babes" University of Medicine and Pharmacy Timisoara. We further detail the list of projects (P) in which the author was director, respectively member, and the full publication list, including books (B), book chapters (BC), in-extenso journals papers (J), and conference proceedings (C).

[1]Publons WoS profile: `https://publons.com/researcher/3545438/alexandru-topirceanu/`

[2]Google Scholar profile: `https://scholar.google.ro/citations?user=pHiA3SkAAAAJ&hl=en&oi=ao`

- Project director/responsible:

  P1 Alexandru Topîrceanu (director), "IMPRESS: Improving the prediction of opinion dynamics in temporal social networks: mathematical modeling and simulation framework", UEFISCDI PN-III-P1-1.1-PD-2016-0193, 28PD/2018 (Total value: 178.215 lei / 37.695 euro).

  P2 Alexandru Topîrceanu (director), "PollStream: Agent-based interaction models with temporal attenuation for opinion poll prediction", UEFISCDI PN-III-P1-1.1-PD-2019-0379, PD7/2020 (Total value: 246.410 lei / 50.711 euro).

- Project member:

  P3 Stefan Dan Mihaicuta (director), Universitatea de Medicina si Farmacie Victor Babes Timisoara, "Morpheus: A Screening and Monitoring System for Sleep Apnea Syndrome", Linde Realfund (2015–2016) - 75K euro.

  P4 Gabriela Grosseck (director), Universitatea de Vest Timisoara, "Novamooc: Innovative development and implementation of moocs in higher education", UEFISCDI PN-II-RU-TE-2014-4-2040 (2015–2017) - 120K euro.

  P5 Mihai Udrescu (director), Univeritatea Politehnica Timisoara, "Inception: Internet of things meets complex networks or early prediction and management of chronic obstructive pulmonary disease", UEFISCDI PN-III-P2-2.1-PED-2016-1145 (2017–2018) - 120K euro.

  P6 Lucian Prodan (director), Univeritatea Politehnica Timisoara, "Wikitraffic: Experimental Assessment of a Self-Adaptive Intelligent Transportation System", UEFISCDI PN-III-P2-2.1-PED-2016-1518 (2017–2018) - 100K euro.

  P7 Alexandru Iovanovici (director), Univeritatea Politehnica Timisoara, "Dormamu: Prediction and management of road congestion using machine learning", ARUT 1349/01.02.2019 (2019–2020) - 10K euro.

  P8 Lucretia Udrescu (director), Universitatea de Medicina si Farmacie Victor Babes Timisoara, "Hyperion: Știința complexității în farmacia de precizie: predicția interacțiunilor medicamentoase relevante folosind analiza rețelelor complexe", UEFISCDI PN-III-P2-2.1-PED2019-2842 (2020–2022) - 120K euro.

  P9 Stefan Dan Mihaicuta (director), Universitatea de Medicina si Farmacie Victor Babes Timisoara, "Sleep Revolution: Revolution of sleep diagnostics and personalized health care based on digital diagnostics and therapeutics with health data integration", 965417 / Horizon 2020 (2021–2025) - 15M euro (total value), 131K euro (local budget).

- Books

  B1 Alexandru Topîrceanu, Marius Marcu, "Introducere in programarea Android", Colectia "Programare", 137 pg., Editura Politehnica, Timisoara, 2015, ISBN 978-606-554-986-9.

B2 Alexandru Topîrceanu, "Hands-On Android Application Development with Google Firebase", Colectia "Calculatoare", 155 pg., Editura Politehnica, Timisoara, 2021, ISBN 978-606-35-0408-2.

- Book chapters (in chronological order)

BC1 Topirceanu, A., Udrescu, M., Vladutiu, M., "Genetically Optimized Social Network Topology Inspired by Facebook", Online Social Media Analysis and Visualization (pp. 163-179). Springer International Publishing, ISBN 978-3-319-13590-8, DOI:10.1007/978-3-319-13590-8_8, 2014.

BC2 Topirceanu, A., Udrescu, M., Avram, R., Mihaicuta, S. (2016). Data Analysis for Patients with Sleep Apnea Syndrome: A Complex Network Approach. In Soft Computing Applications (pp. 231-239), ISBN 978-3-319-18296-4, Springer International Publishing.

BC3 Topirceanu, A., Iovanovici, A., Cosariu, C., Udrescu, M., Prodan, L., Vladutiu, M. (2016). Social Cities: Redistribution of Traffic Flow in Cities Using a Social Network Approach. In Soft Computing Applications (pp. 39-49), ISBN 978-3-319-18296-4, Springer International Publishing.

BC4 Iovanovici, A., Topirceanu, A., Cosariu, C., Udrescu, M., Prodan, L., Vladutiu, M. (2016). Heuristic Optimization of Wireless Sensor Networks Using Social Network Analysis. In Soft Computing Applications (pp. 663-671), ISBN 978-3-319-18296-4, Springer International Publishing.

BC5 Topîrceanu A., Udrescu M. (2018) Strength of Nations: A Case Study on Estimating the Influence of Leading Countries Using Social Media Analysis. In: Alhajj R., Hoppe H., Hecking T., Bródka P., Kazienko P. (eds) Network Intelligence Meets User Centered Social Media Networks. ENIC 2017. Lecture Notes in Social Networks. Springer, Cham.

- Journal papers (in chronological order):

J1 Topirceanu, A., Duma, A., Udrescu, M. (2016). Uncovering the fingerprint of online social networks using a network motif based approach. Computer Communications, 73, 167-175 [**IF=3.338, Q1** - Computer Science, Information Systems; Engineering, Electrical & Electronic].

J2 Topirceanu, A., Udrescu, M., Vladutiu, M., Marculescu, R. (2016). Tolerance-based interaction: a new model targeting opinion formation and diffusion in social networks. PeerJ Computer Science, 2, e42. [**IF=3.091, Q1** - Computer Science, Theory & Methods].

J3 Suciu, L., Cristescu, C., Topîrceanu, A., Udrescu, L., Udrescu, M., Buda, V., Tomescu, M. C. (2016). Evaluation of patients diagnosed with essential arterial hypertension through network analysis. Irish Journal of Medical Science(1971-), 185(2), 443-451 [IF=1.224, Q3 - Medicine, General & Internal].

J4 Udrescu, L., Sbârcea, L., Topîrceanu, A., Iovanovici, A., Kurunczi, L., Bogdan, P., Udrescu, M. (2016). Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing. Scientific Reports, 6 [**IF=4.259, Q1** - Multidisciplinary Sciences].

J5 Topirceanu, A., Udrescu, M. (2017). Statistical fidelity: a tool to quantify the similarity between multi-variable entities with application in complex networks. International Journal of Computer Mathematics, 94(9), 1787-1805 [**IF=1.054, Q2** - Mathematics, Applied].

J6 Mihaicuta, S., Udrescu, M., Topirceanu, A., Udrescu, L. (2016). Network science meets respiratory medicine for OSAS phenotyping and severity prediction. PeerJ, 5:e3289 [**IF=2.183, Q1** - Multidisciplinary Sciences].

J7 Topîrceanu, A. (2017). Breaking up friendships in exams: A case study for minimizing student cheating in higher education using social network analysis. Computers & Education, 115, 171-187 [**IF=4.538, Q1** - Computer Science, Interdisciplinary Applications].

J8 Topirceanu, A., Udrescu, M., Marculescu, R. (2018). Weighted Betweenness Preferential Attachment: A New Mechanism Explaining Social Network Formation and Evolution. Scientific reports, 8(1), 10871 [**IF=4.122, Q1** - Multidisciplinary Sciences].

J9 Topîrceanu, A. (2018). Competition-Based Benchmarking of Influence Ranking Methods in Social Networks. Complexity, 2018 [**IF=2.591, Q1** - Mathematics, Interdisciplinary Applications; Multidisciplinary Sciences].

J10 Topîrceanu, A., Udrescu, M., Udrescu, L., Ardelean, C., Dan, R., Reisz, D., Mihaicuta, S. (2018). SAS score: Targeting high-specificity for efficient population-wide monitoring of obstructive sleep apnea. PloS one, 13(9), e0202042 [**IF=2.766, Q1** - Multidisciplinary Sciences].

J11 Fierăscu, S. I., Pârvu, M., Topîrceanu, A., Udrescu, M. (2018). Exploring Party Switching in the Post-1989 Romanian Politicians Networks from a Complex Network Perspective. Romanian Journal of Political Science, 18(1), 108-136 [IF=0.421, Q4 - Political Science].

J12 Barina, G., Udrescu, M., Barina, A., Topirceanu, A., Vladutiu, M. (2019). Agent-based simulations of payoff distribution in economic networks. Social Network Analysis and Mining, 9(1), 63.

J13 Topîrceanu, A., Precup, R. E. (2020). A framework for improving electoral forecasting based on time-aware polling. Social Netw. Analys. Mining, 10(1), 39.

J14 Udrescu, L., Bogdan, P., Chis, A., Sirbu, I. O., Topirceanu, A., Varut, R. M., Udrescu, M. (2020). Uncovering new drug properties in target-based drug-drug similarity networks. Pharmaceutics, 12(9), 879 [**IF=4.421, Q1** - Pharmacology & Pharmacy].

J15 Topîrceanu, A., Udrescu, L., Udrescu, M., Mihaicuta, S. (2020). Gender Phenotyping of Patients with Obstructive Sleep Apnea Syndrome Using a Network Science

Approach. Journal of Clinical Medicine, 9(12), 4025 [**IF=3.303, Q1** - Medicine, General & Internal].

J16 Mihaicuta, S., Udrescu, L., Udrescu, M., Toth, I. A., Topîrceanu, A., Pleavă, R., Ardelean, C. (2021). Analyzing Neck Circumference as an Indicator of CPAP Treatment Response in Obstructive Sleep Apnea with Network Medicine. Diagnostics, 11(1), 86 [**IF=3.11, Q1** - Medicine, General & Internal].

- Conference proceedings papers (indexed in WoS)

C1 Topirceanu, A., Udrescu, M., Vladutiu, M. (2013, September). Network fidelity: A metric to quantify the similarity and realism of complex networks. In Cloud and Green Computing (CGC), 2013 Third International Conference on (pp. 289-296). IEEE.

C2 Barina, G., Topirceanu, A., Udrescu, M. (2014, May). MuSeNet: Natural patterns in the music artists industry. In Applied Computational Intelligence and Informatics (SACI), 2014 IEEE 9th International Symposium on (pp. 317-322). IEEE.

C3 Duma, A., Topirceanu, A. (2014, May). 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). IEEE.

C4 Alexandru Topîrceanu, Cezar Fleșeriu, Mihai Udrescu, "Gamified: An Effective and Innovative Approach to Student Motivation Using Gamification". The 2nd International Conference on Social Media in Academia: Research and Teaching (SMART 2014), pp. 41-44.

C5 Alexandru Topîrceanu, Dragoș 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 (SMART 2014), pp. 355-360.

C6 Topirceanu, A., Barina, G., Udrescu, M. (2014, September). Musenet: Collaboration in the music artists industry. In Network Intelligence Conference (ENIC), 2014 European (pp. 89-94). IEEE.

C7 Topirceanu, A., Udrescu, M. (2015, September). FMNet: Physical Trait Patterns in the Fashion World. In Network Intelligence Conference (ENIC), 2015 Second European (pp. 25-32). IEEE.

C8 Iovanovici, A., Topirceanu, A., Udrescu, M., Prodan, L., Mihaicuta, S. (2015). A high-availability architecture for continuous monitoring of sleep disorders. In MIE (pp. 729-733).

C9 Topirceanu, A., Udrescu, M. (2015, May). Measuring realism of social network models using network motifs. In Applied Computational Intelligence and Informatics (SACI), 2015 IEEE 10th Jubilee International Symposium on(pp. 443-447). IEEE.

C10 Udrescu, M., Topîrceanu, A. (2015, May). What Drives the Emergence of Social Networks?. In 2015 20th International Conference on Control Systems and Computer Science (pp. 999-999). IEEE.

C11 Topirceanu, A., Garcia, J., Udrescu, M. (2016, September). UPT. Social: The Growth of a New Online Social Network. In Network Intelligence Conference (ENIC), 2016 Third European (pp. 9-16). IEEE.

C12 Udrescu, M., Topirceanu, A. (2016, September). Probabilistic Modeling of Tolerance-Based Social Network Interaction. In Network Intelligence Conference (ENIC), 2016 Third European (pp. 48-54). IEEE.

C13 Topîrceanu, A. (2017). Gamified learning: A role-playing approach to increase student in-class motivation. Procedia Computer Science, 112, 41-50.

C14 Topîrceanu, A., Grosseck, G. (2017). Decision tree learning used for the classification of student archetypes in online courses. Procedia Computer Science, 112, 51-60.

C15 Barina, G., Udrescu, M., Topirceanu, A., Vladutiu, M. (2018, August). Simulating Payoff Distribution in Networks of Economic Agents. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 467-470). IEEE.

C16 Topirceanu, A., Udrescu, M. (2018, August). Topological Fragility Versus Antifragility: Understanding the Impact of Real-Time Repairs in Networks Under Targeted Attacks. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1215-1222). IEEE.

C17 Topîrceanu, A., Precup, R. E. (2019, August). A novel methodology for improving election poll prediction using time-aware polling. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 282-285).

- Conference proceedings papers (indexed in other databases)

C18 Marcu, M., Stangaciu, C., Topirceanu, A., Volcinschi, D., Stangaciu, V. (2010, December). Wireless Sensors Solution for Energy Monitoring, Analyzing, Controlling and Predicting. In International Conference on Sensor Systems and Software (pp. 1-19). Springer Berlin Heidelberg [SpringerLink].

C19 Iovanovici, A., Topirceanu, A., Udrescu, M., Vladutiu, M. (2014, October). 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). IEEE [IEEE Xplore].

C20 Topirceanu, A., Iovanovici, A., Udrescu, M., Vladutiu, M. (2014, October). 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). IEEE [IEEE Xplore].

C21 Topirceanu, M., Topirceanu, A., Udrescu, M. (2019, May). Exploring currency exchange dynamics from a complex network perspective. In 2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI) (pp. 63-68). IEEE [IEEE Xplore].

C22 Topîrceanu, A., Udrescu, M., Mărculescu, R. (2020, January). Complex Networks Antifragility under Sustained Edge Attack-Repair Mechanisms. In International Conference on Network Science (pp. 185-199). Springer, Cham [SpringerLink].

C23 Muntea, D., Giurgiu, M., Topirceanu, A. (2020, May). Network-based clustering of book genres based on the connection between books bought together. In 2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI) (pp. 000041-000046). IEEE [IEEE Xplore].

C24 Topîrceanu, A. (2020, December). Analyzing the Impact of Geo-Spatial Organization of Real-World Communities on Epidemic Spreading Dynamics. In International Conference on Complex Networks and Their Applications (pp. 345-356). Springer, Cham [SpringerLink].

In addition to the list of enumerated journal papers (J) and proceedings papers (C), we have published a number of 11 WoS & PubMed indexed medical congress abstracts (1–2 pages) indexed in journals of high visibility (e.g., Chest, European Respiratory Journal).

Moreover, as an appreciation for contributions in the field of Network Science, we received a **Best Paper Award** for our paper [C6] at the 2nd European Network Intelligence Conference, ENIC, Karlskrona, Sweden, 21-22 Sep, 2015: Alexandru Topîrceanu and Mihai Udrescu, "FMNet: Physical Trait Patterns in the Fashion World" [58].

1.3 Theoretical Foundations of Complex Networks

This section introduces the basic theoretical elements and taxonomy that is used throughout the rest of the thesis to refer to our Complex Network research. Please note that a full introductory coverage of the field of Network Science is beyond the goal of this thesis. As such, we provide references pinpointing to the detailed literature regarding the introduced topics [4, 9, 61, 5, 10, 62].

1.3.1 Graphs as Complex Networks

Modeling a complex system often starts from identifying its components and possible interaction types. As such, graphs capture the building blocks of such systems conveniently, by making use of nodes (vertices) and edges (links). Even though complex systems vary greatly in terms of structure, function and goal, graphs offer a common modeling paradigm, and enable the study of graph-specific properties. The data structure commonly used in mathematics, computer science and engineering to model pairwise relations between objects is done by defining a complex network $G = (N, E)$, which consists of the set N of nodes, which are interconnected via the set E of edges. We symbolize a node as $n_i \in N$ and any edge $e_{ij} \in E$ connects two nodes n_i and n_j. Based on the nature of the problem, and the set E, the graph may be undirected, meaning that edges are equivalent in terms of the two ends ($e_{ij} = e_{ji}$),

or directed from one node to another ($e_{ij} \neq e_{ji}$). In the directed scenario, we can state that there is a path d_{ij} from n_i to n_j, but not necessarily vice-versa. In an undirected scenario all paths are bidirectional.

Nodes

Nodes represent the abstraction of a natural or synthetic entity stemming from a process (complex system) for which network science is employed. At the most basic level, each node possesses an identity (id) and a set of edges through which it connects to other nodes, forming its vicinity N_i. Often, nodes possess context-specific properties which are used to characterize emergent clusters, using bipartite graphs [63], and community detection methods [64, 65, 62].

Formally, N represents the number of components in the modeled system, and is referred to as the *size* of the network. Intrinsically, each node n_i possesses a position $x(n_i) \in a2 - dimensionalEuclideanspace\mathbb{R}^2$ (in some cases, networks can be modeled in higher degree dimensions).

Edges

Edges represent a relationship between two nodes, connecting them in graph G. Edges can be directed or undirected, respectively weighted or unweighted. Intuitively, if edge $e_{i,j}$ is undirected then both nodes can be reached following the edge from the other end. Conversely, if $e_{i,j}$ is directed then the edge can only be followed from n_i to n_j. Of course $e_{i,j}$ does not imply the existence or absence of another edge $e_{ji} \in E$. In the context of undirected graphs, an edge $e_{i,j} \in E$ may be associated a weight equal to $w_{ij} = 1$ when computing paths d_{ij} (or costs). If the relationship between nodes implies different magnitudes, then weights may be assigned to edges ($w_{i,j} > 0$). A special case for edges are the self-loops $e_{i,i}$, in which a node redirects to itself (e.g., a web page having a link that redirects to itself).

Formally, each pair of nodes (n_i, n_j) can assign a Euclidean distance $\|x(n_i) - x(n_j)\|$ to each existing edge $e_{i,j} \in E$. The number of edges E represents the total number of interactions between all the nodes N in the graph G. More often, edges are identified through the nodes which they connect rather than using a distinct label.

1.3.2 Metrics of Complex Networks

Node Degree and Network Average Degree

The degree of a node is the number of nodes with which it is connected through graph edges. The degree can represent the number of acquaintances in a social network, the number of emails received from a particular contact, the number of medications on a prescription, the number of coauthors of a paper etc.

We use the notation k_i for the degree of a node n_i. In an undirected network we can express the total number of edges E as the sum of node degrees as: $E = 1/2 \sum_i k_i$. In directed graphs (also digraphs), a node has two degrees: an out-degree k_i^{out} for edges exiting the node, and an in-degree k_i^{in} for incoming edges. The sum $k_i^{out} + k_i^{in}$ is considered the total node degree k_i. In digraphs, the number of edges is expressed as $E = \sum_i k_i$ (without the division by 2).

The average degree $\langle k \rangle$ of an undirected network G is computed as:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2E}{N} \tag{1.1}$$

For directed networks, we express the average in-degree $\langle k^{in} \rangle$ and average out-degree $\langle k^{out} \rangle$ separately, as:

$$\langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} = \langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = \frac{E}{N} \tag{1.2}$$

A popular alternative to the symbol $\langle k \rangle$ used in state of the art papers is simply AD or $AvgDeg$.

Degree Distributions

The degree distribution $P(k)$ of a network is a statistical function describing the probability that random node $n_i \in N$ has exactly degree k. $P(k)$ is used to describe distribution of node degrees over the whole network. Since the degree of any connected node is $k_i \leq 1$, but we are using a probability distribution, the function $P(k)$ must be normalized such that $\sum_i^N P(k) = 1$.

The degree distribution P(k) is defined as the ratio between the number of nodes with degree k and the total number of nodes [10]:

$$P(k) = \frac{N_k}{N} \tag{1.3}$$

where N_k is the number of nodes with degree k. The function $P(k)$ describes the probability that a randomly selected node has degree k. For example, a regular lattice (i.e., a chessboard with only vertical and horizontal links), with most nodes having similar degree ($k = 4$), will have a distribution P(k) characterized by spike at the exact $k = 4$. The more randomness (irregularity) is added to the network connections, the broader the spike becomes. Conversely, 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 [8]. Such a distribution is often expressed as:

$$P(k) \approx k^{-\beta} \tag{1.4}$$

where β has been observed to be between 2 and 3 in a large variety of real-world networks. As this form of distribution is not subject to network scale, it is a signature characteristic for scale-free networks.

Nodes with a higher degree than other are called hubs, as they tend to facilitate communication for non-local nodes (i.e., nodes outside the vicinity). Also, according to the scale-free network model, a more connected node has a higher chance of becoming even more connected – the basis of Degree Preferential Attachment [3].

Network Density

Based on the introduced notations, we call graph G a dense graph when the number of edges E is close to the maximal number of edges $E \to E_{max}$. The opposite of a dense graph is a sparse graph, having $E \ll E_{max}$. The maximum number of edges is expressed as $E_{max} = N(N-1)/2$. Complete graphs are very rare in nature, as most real-world networks are sparse. The exact threshold between sparse and dense graphs is rather vague, and depends on the context. Two different definitions exists for density Dns whether we refer to undirected or directed graphs. For undirected graphs, the graph density Dns is defined as:

$$Dns = \frac{2|E|}{|N| \cdot (|N| - 1)} = \frac{|E|}{|E_{max}|} \tag{1.5}$$

Using equation 1.5 Dns 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 equation is slightly modified to:

$$Dns = \frac{|E|}{|N| \cdot (|N| - 1)} = \frac{|E|}{|E_{max}|} \tag{1.6}$$

Equation 1.6 lacks the $2\times$ multiplier because in directed graphs we have a double number of maximum edges compared to undirected graphs. The minimum density Dns is 0, and the maximum Dns is 1.

Average Clustering Coefficient

The clustering coefficient C_i of a node n_i is a measure of the nodes' tendency to cluster together. This real-world property can be exemplified with a friendship network, where there is a high probability that one person's friend of a friend is also a direct friend of that person. Rephrased, it is very likely that two friends of a person are also friends with one another, forming a triadic closure. Thus, the clustering coefficient can be defined as the ratio between the existing number of links between a node and his neighbors, and the total number of links that can exist between the neighbors. More precisely, a node n_i with degree k_i has $|N_i| = k_i$ neighbors. The maximum number of links between all neighboring nodes is $|N_i|(|N_i| - 1)/2$. As such, we express C_i in an undirected network as:

$$C_i = 2 \cdot \frac{E_i}{|N_i|(|N_i| - 1)} \tag{1.7}$$

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 average clustering coefficient C (or ACC) of the network:

$$C = \frac{\sum_i^N C_i}{N} \tag{1.8}$$

From equation 1.8 we conclude that the maximum value of C is 1. A network with $C = 1$ is a fully connected graph with point-to-point connections, while a completely random network has $C \approx 1/N$. However, the clustering measured in random networks is much smaller than

compared to observable networks which have their clustering coefficient satisfy the following relationship:

$$\frac{1}{N} \ll C < 1 \tag{1.9}$$

This empirical observation means that most networks are neither random, nor fully connected, and thus the triadic closure is a very important aspect of real networks. An illustration of the clustering coefficient is depicted in Figure 1.4a.

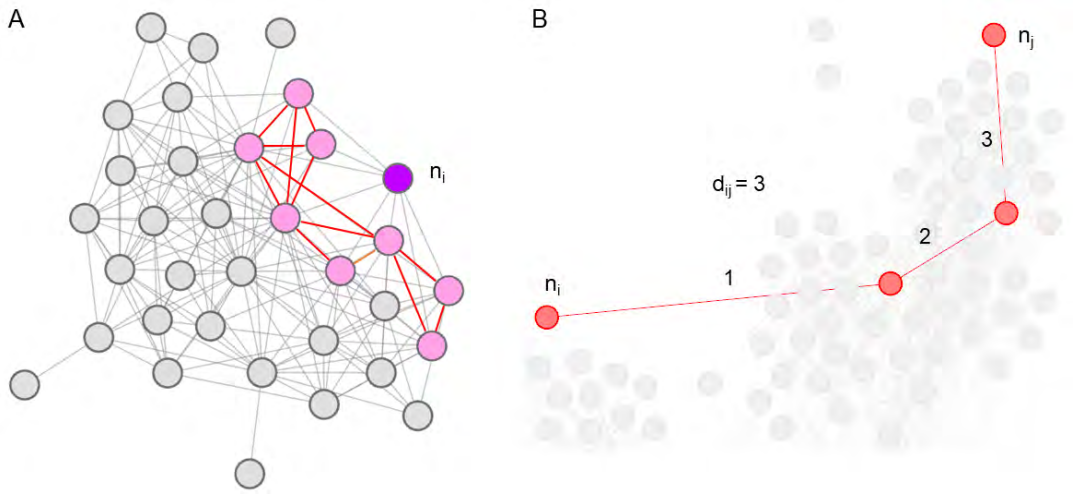

Figure 1.4: (A) Illustration of the clustering coefficient of a node n_i (violet). The size of its vicinity is $|N_i| = 8$, and the neighbors have a number of $E_i = 13$ links between them. Given $E_{max} = 28$ for 8 nodes, we compute $C_i = 13/28 = 0.464$. (B) Illustration of the shortest distance between two nodes in a network. The nodes are connected via the red path with length $d_{ij} = 3$.

1.3.3 Paths and Distances in Networks

The physical distance between entities is replaced by path length between nodes in a complex network. A path connects any two nodes via a set of links, and its length is given by the number of links contained in the path. The distinction between a path and the more generic term walk, is that a walk may contain cycles, whereas a path may not contain a node twice. Paths are highly studied in Network Science because most types of processes (diffusion) are influenced by the general length of paths between nodes.

The distance between two nodes n_i and n_j in a network G is given by the number of links of the shortest path between n_i and n_j denoted as d_{ij}. Note that in an undirected network $d_{ij} \neq d_{ji}$, while in a directed network they are equal.

The diameter Dmt of a network is expressed as the longest shortest path in G. Intuitively, Dmt measures the distance between the two furthest nodes.

Average Path Length

The average path length (L, or APL) is one of the fundamental graph metrics used for characterizing a network topology. The average path length L is the sum of all paths between all pairs of nodes, divided by the number of all possible paths in the network:

$$L = \frac{2}{|N| \cdot (|N| - 1)} \sum_{i \neq j} d_{ij} \tag{1.10}$$

where $|N|$ is the size of the graph, and (i, j) are distinct node indices. For example, in a network of friends, L is the average number of friends that form up the shortest way connecting any two friends [10]. 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 (e.g, brain neural networks), even though having significantly more edges than some simple synthetic networks (e.g., computer network), still have a very small average path length L. This is the property known as small world effect found in small-world networks [2, 66]. An illustration of the shortest path length is depicted in Figure 1.4b.

1.3.4 Community Formation

One of the main motivations behind graph modeling of natural or man-made complex systems is to analyze how the different entities (i.e., nodes) connect and cluster together - both quantitatively and qualitatively [67, 65]. Most network-based modeling results in an emergent community structure that has substantial importance in further understanding the dynamics of the network. For instance, a highly clustered social community will imply a faster rate of transmission of information or rumor among them than loosely connected communities [62]. Thus, if a network is represented by a number of nodes connected by edges, which signify a certain degree of interaction between those nodes, then communities are defined as groups of densely interconnected nodes that are only sparsely connected with the rest of the network [68]. Hence, it is useful to identify the communities in networks since these can have different graph properties (e.g., average node degree, clustering coefficient, and other centralities) than the larger network G [69]. Consequently, community detection and analysis have received much attention over the last two decades [70, 64, 65, 62, 9, 58, 41, 43, 63].

There are two parallel, non-exclusive, approaches to community detection and analysis. First, we mention node partitioning into distinct clusters. There are multiple numerical methods which imply assigning a community label to each node. We mention here two of the more popular partitioning methods: the modularity algorithm [71] and the Louvain method [72]. Second, given the visual nature of networks, force-directed layout algorithms are further used to place the nodes in an intuitive manner in a 2-dimensional space. As such, nodes that are clustered together tend to be more similar, and have an increased interconnectedness compared to other distant nodes.

Network Modularity

Modularity (Mod) is a measure of the interconnectedness of networks. Mod is a quantitative measure of the strength of division of a network into groups (also called modules, communities

or clusters). The value of Mod suggests that networks with high modularity have denser connections between the nodes within communities, but sparser connections between nodes in different communities. Numerically, Mod is defined as the ratio of edges which exist within a given community minus the expected such ratio if edges were distributed at random. The maximum value for modularity is 1 [73].

In a complex network G, a clustering algorithm via modularity is an assignment A_m of each node n_i in one of the clusters C_j, with $\cup_{k=1}^{m} C_j = N$ [44]. As such, when modularity determines the assignment of nodes to their corresponding clusters $A_m = \{C_1, C_2, ..., C_m\}$, the algorithm maximizes the modularity of clustering A_m as follows:

$$Mod_{A_m} = \sum_{C \in A_m} \left(\frac{|E_{C_j}|}{|E|} - \frac{\frac{1}{2}k_{C_j}^2}{\frac{1}{2}k^2} \right) \tag{1.11}$$

where $|E|$ is the number of edges in G, $|E_{C_j}|$ is the number of edges in cluster C_j, k is the accumulated degree of nodes in G, and k_{C_k} is the accumulated degree of nodes solely in cluster C_j [47]. As such, $|E_{C_j}|/|E|$ translates to the edge density of cluster C_j relative to the density of the whole network, and the term $\frac{1}{2}k_{C_j}^2/\frac{1}{2}k^2$ represents the expected relative density of C_j [64].

There are different methods for calculating modularity [65]; in the most common version of the metric, the randomization of the edges is done so as to preserve the degree of each vertex. An example of two graphs with different modularity are generated and depicted in Figure 1.5a.

Force Directed Layout Algorithms

A force-directed layout algorithm assigns each node $n_i \in N$ a coordinate in a 2-dimensional space $\delta_i = (x_i, y_i) \in R^2$. Consequently, each edge has a length defined by the Euclidean distance $\delta_{i,j} = |\delta_i - \delta_j|$. A force-directed (or energy-based) layout generates the δ_i for each n_i using a dynamic, emergent process, where any two adjacent nodes n_i and n_j attract each other, and any two non-adjacent nodes n_i and n_k repulse each other. We express such attraction/repulsion forces as $|\delta_i - \delta_j|^{\Phi} \cdot \overrightarrow{\delta_i \delta_j}$, where $\Phi = a$ for attraction, $\Phi = r$ for repulsion, and $\overrightarrow{\delta_i \delta_j}$ is the unit vector. The attraction between adjacent nodes decreases and the repulsion between non-adjacent nodes increases with the Euclidean distance between them; therefore, we have $a \geq 0$ and $r \leq 0$ [44].

One of the most popular force directed layout algorithms is Force Atlas 2 [74] which employs a dynamic complex process based on interacting attraction and repulsion forces to attain minimal energy in the layout:

$$min\left\{ \sum_{(n_i,n_j), i \neq j} \left(\frac{|\delta_i - \delta_j|^a}{a+1} - \frac{|\delta_i - \delta_j|^r}{r+1} \right) \right\} \tag{1.12}$$

As such, force directed layouts can generate topological clusters, as some specific network regions have higher than average edge densities. Noack [69] has demonstrated that modularity-based and force-directed layout communities/clusters are equivalent when $a > -1$ and $r > -1$, which, indeed, is the case for Force Atlas 2 [44].

Noack et al. [69] demonstrated that force directed layouts and modularity are analogous when $a > -1$ and $r > -1$ (which is the case for Force Atlas 2). As such, we will often refer to the dual clustering technique in this thesis, which means applying both modularity (numerically – quantitative) and a layout algorithm (visually – qualitative). The practical impact of applying a force directed layout on a network dataset is exemplified in Figure Figure 1.5b.

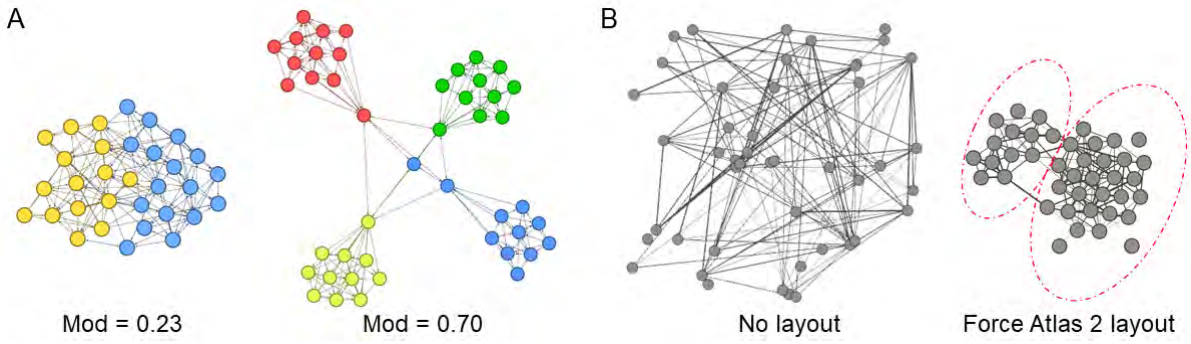

Figure 1.5: (A) An overview of community detection in networks using a modularity based approach. First, we depict a graph with a weak community structure, quantified by the relatively small modularity $Mod = 0.23$. Second, we depict a graph with visibly stronger community structure, and a modularity of $Mod = 0.70$. All nodes are colored according to the community to which they belong. (B) The visual impact of applying a layout algorithm on the same dataset.

The visualizations in Figure 1.5 are based on network models generated using proprietary code implemented as a Java plugin in Gephi [75], an open source graph visualization tool.

1.4 Thesis Outline

The thesis is organized in six subsequent chapters, as follows. Chapter 2 outlines the algorithmic contributions aimed at improving our structural understanding of complex networks, with direct applicability in social network analysis. Chapter 3 describes contributions in modeling and computer simulation of processes over social networks, representing opinion formation, dynamics and convergence. Chapter 4 summarizes two important tracks of contributions using algorithmic methods in sleep research and pharmacology, with direct applicability in patient phenotype definition and drug repurposing. Chapter 5 describes the applicability of network science in educational Big Data.

The second part of the thesis, starting with Chapter 6, outlines future research directions in computational epidemics. Indeed, there has been a recent rise in efforts to better model and predict epidemic outbreak dynamics, and we are proposing robust computer simulation based approaches for large scale analyses with high societal impact. The last chapter of the thesis outlines the main obtained results from research projects, and sketches the perspectives for future research projects in the context of Computers & Information Technology using Complex Systems and Big Data. Finally, we list all relevant references from our work, other related works, and from relevant state-of-the-art literature.

Chapter 2

Contributions in Social Network Analysis

2.1 Introduction

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 real or virtual societies [76, 77, 78] implying e.g., friendship, collaboration, competition, economical, political ties. The structure of social networks can be further generalized to multiple layers [79], represented by groups of persons, clusters, cities, states etc., each layer with a particular set of defining characteristics [80]. The area of Social Network Analysis (SNA) is relatively new, with its fundamentals starting from the 1970s [81] and was initially based on empirical observations of computer networks and human networks, with many ideas coming from the more distant field of sociology. Even though 50 years old, only more recently (2010s) has this field started to attract massive interest from universities and researchers around the world. Many important Computer Science and Engineering Universities in Europe, North America and Asia have a dedicated department to Network Science, Social Computing, or any related branches (e.g., ETH Zurich, Oxford, Cambridge, CEU, MIT, Stanford, CMU, Columbia, Penn State, USC etc.).

In general, it is considered that the evolution of companies like Facebook and Google has generated the increased interest of computer science in social networks [82]. Not only does the available Big Data offer data scientists valuable feedback on their assumptions and validations, but the ubiquitous usage of social sites is attracting more Computer Science and Engineering students into this area of science.

A social network is a graph model G consisting of individuals (i.e., actors, agents) and connections (i.e., relationships) between these individuals resembling a real social structure of people. The scientific value of such a network model is to provide information on how relationships evolve and how information is transmitted within the society, as determined by the interactions, i.e., topology. The two important 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 [3, 83]. Emergence refers to the characteristic that patterns observed at a small scale, inside a small group of agents, replicate at a greater scale.

However, with increasing network size (e.g., million of agents), the dynamics of the model become overwhelming for analytical approaches, and computer simulation is adopted. Also, studies are aimed at groups of interest, rather than at cities or country levels.

Classically, there are three levels of the social network composition that are being studied separately [62]:

- The micro level – studying an individual and relationships, and modeling its interaction patterns,
- The middle (meso) level – studying small groups of individuals, like triadic closures and cliques, where specific dynamics and polarization are observable.
- The macro-level – studying emergence of large populations and the global effects of processes on these large topologies, regardless of the individual scale.

This thesis deals with all three levels of SNA and this section is oriented towards summarizing contributions on the topological level of network. The rest of this section introduces (i) fundamental complex network topologies used as inspiration for state of the art topological research, (ii) network centralities used for studying influence ranking methods, and (iii) network motifs used for the characterization of different types of social networks.

2.1.1 State-of-the-art Complex Network Topologies

A network topology is a term used to describe the interconnection pattern of the entities composing the network. Linking elements can be done physically or logically. As social networks describe combined human relationships, knowledge or collaboration, the links are purely logical. Analyzing topologies is done with the help of graph theory, a mathematical theory used to describe relationships between objects. Network Science divides the types of topologies in two main categories [10]:

- Regular, or basic topologies -– are mostly find are configurations of simple (small-scale, local) technological networks, studied by Computer Engineering and Communication. These networks have simple, generally symmetric layouts of nodes with simple patterns of interconnectivity [22]. Moreover, regular networks are also called non-complex networks because of the reduced number of nodes (e.g., tens of nodes instead of thousands).
- Complex network topologies -– are a more comprehensive set of interconnections that bind larger numbers of nodes. Complex networks are characterized by high to very high number of nodes (e.g., up to billions) which possess numerous links, both with local neighbors, as well as links to distant nodes [5, 66]. Natural and man-made systems have but recently been modeled and studied as networks [8].

The research goal of Network Science is to offer models which characterize real-world systems in an accurate manner. As such, the analysis of complex systems which can be modeled as graphs has revealed that real-world networks (technological, biological, social or semantic [10]) are generally sparse with low density (Dns), have a short average path length (APL), high clustering (CC), and complex, non-uniform degree distribution, such as

power-law or Gaussian [26]. Also, many real-world complex networks also exhibit complex distributions of other centralities like: betweenness Btw, closeness Cls, Eigenvector EC, or PageRank PR [84, 70].

Random Networks

Random networks were first formalized by Paul Erdos and Alfred Renyi in the Erdos-Renyi (ER) model [1]. Nodes in the ER model are randomly connected with a given probability p, regardless of spatial localization. This algorithm results in the creation of a high proportion of so-called long range links across the network. On inspection, random networks, show a drastic decrease in the average path length APL, as long-range links are randomly inserted in the network. On the other hand, the clustering coefficient CC remains low as there is no rule to back up the creation of triadic closures. The resulting average degree $\langle k \rangle$ is pN, $APL \approx lnN/\langle k \rangle$, and $CC = p = \langle k \rangle / N$.

Even though random networks differ from most real-world network models, ER is a valuable complex network model because it first proved an existing phase transition from a disconnected to a connected random, based solely on the interconnectivity $\langle k \rangle$, given by p. Hence, it is considered that networks become connected when $p > lnN/N$. ER are also used to bridge regular networks with small-worlds [2], and to integrate statistics with graph theory [10]. Also, ER models are frequently used in null hypothesis testing.

Regular Networks (Meshes)

Meshes are commonly used to describe any type of interaction bounded by geography or location. Human society itself – in the pre-Internet Era – was best described by meshes, where people had to travel in order to create now links. In meshes, each node n_i may be connected to one or more neighboring nodes, forming the vicinity N_i, within a close proximity (usually given by a maximum range δ). This type of network allows advanced routing of information along multiple paths, from a destination to a source, with increased reliability. In theory, any node can have any number of connections ($0 \leq k_i \leq N(N-1)/2$); in practice, the average degree of nodes is determined by the maximum connection distance δ and the link probability p, and results in the so-called forest fire, or mesh topology [85].

Meshes have relatively large average path lengths (i.e., a linear relationship between APL and N), simple sequences for degree distributions, high clustering coefficients CC, similar to real-world networks (e.g., road networks).

Small-World Networks

The "small world effect" was first experimented by s. Milgram [86] who concluded that we live in a world with unexpected short average paths. The term "six degrees of separation" was also introduced to quantify the expected APL over the entire planet. The small-world (SW) model of Watts-Strogatz (WS) was introduced in 1998 and represent a fundamental topology which possesses properties found in real societies and systems [2]. The WS 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 APL only grows at a logarithmic rate relative N.

Small-world networks represent a trade-off between the random network's short path length, and the mesh network's high clustering. Given a rewiring probability p, of transforming a mesh into a ER network, we find the SW network, starting with small values of p. In this sense, SW networks have high APL and high CC. Nevertheless, the majority of real-world networks have power-law degree distributions $P(k)$, i.e., not normally distributed as is the case of SW (WS including) networks.

Overall, the Watts-Strogatz network model introduces a complex topology that encompasses one important real world property, namely the triadic closure with high CC; combined with the small APL, WS is an appropriate model for many technological networks, such as power grids, road networks, brain neural networks, food chains, the WWW router network. However, as WS does not create a heterogeneous degree distribution it cannot be used alone for representing social networks.

Scale-Free Networks

The second fundamental complex network topology was introduced by Albert-Laszlo Barabasi and Reka Albert (BA) in 1999 [3]. The scale-free (SF) networks are dynamic, modeling growth, and can describe many observable real world systems based on degree preferential attachment (DPA). DPA produces the observed power-law degree distribution, based on the idea that nodes with higher degree attract more nodes (rich get richer phenomenon). The probability of a new node n_j to connect to any of the existing nodes $n_i \in N$ in the network is $p_j i = k_i / \sum k_n$. The resulting degree distribution is $P(k) \approx k^{-\beta}$, with $-2 < \beta < -3$, such that networks are further considered scale-free if their degree distribution slope β falls within the interval (2, 3). The average path length is short, with $APL \approx lnN/ln(lnN)$, and the clustering coefficient is also low, and scales with the network size $CC \approx N^{-3/4}$ [8].

In conclusion, the BA network model introduces 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, such as the Internet (links between pages), collaboration relationships, airline networks, protein interactions etc. However, the BA topology creates a homogeneous clustering coefficient, which scales with the degree and network size, so it cannot be used alone for representing social networks.

2.1.2 Network Centralities

Quantifying node influence can lead to an improved understanding of the interaction patterns in complex systems. The applicability of metrics for measuring the influence potential of nodes has wide-ranging interdisciplinary applications [39]. The reviewed measures of centrality, referred to in this thesis, are classified in one of three categories, based on the amount of topological information needed, and on their deterministic nature, namely [87, 39]: structure-based, location-based, and diffusion-based centralities.

Structure-based measures

Structure-based measures require the topological information of the graph - either local (e.g, ego-network, vicinity) or global (*e.g.*, path-based). Under local measures we first mention

degree centrality (Deg) k_i of a node v_i; it is easy to use and efficient, but less relevant in some real-world scenarios [88, 89], as some studies show that Deg fails to identify influential nodes because it is limited to the ego-network of each node [39, 90, 88].

The local centrality measure (LC) was introduced as a trade-off between the low-relevant degree centrality and other time-consuming measures [88]. LC of node v_i considers both the nearest and the next nearest neighbors, and is defined as:

$$LC(v_i) = \sum_{v_j \in N_i} Q(v_j), Q(v_j) = \sum_{v_k \in N_j} N(v_k) \tag{2.1}$$

where N_i is the vicinity (set of neighbors) of node v_i, $N(v_k)$ is the number of the nearest and the next nearest neighbors of node v_k, and $Q(v_j)$ is sum of $N(v_k)$ over each node in N_i. LC can considered more effective than degree centrality because it uses more information from the vicinity of distance 2, but has much lower computational complexity than betweenness and closeness centralities [39].

Another method considered a local ranking measure is ClusterRank (CR), proposed by Chen et al. [91]. CR quantifies the influence of a node v_i by taking into account not only its direct influence (out-degree k_i^{out}), and influences of its neighbors (like in the case of PageRank), but also its clustering coefficient c_i [10]. Formally, the ClusterRank score $CR(v_i)$ of a node v_i is defined as:

$$CR(v_i) = f(c_i) \sum_{v_j \in N_i} (k_i^{out} + 1) \tag{2.2}$$

where the term $f(c_i)$ represents the effect of v_i's local clustering, the term $+1$ results from the contribution of v_j itself, and N_i is the vicinity of node v_i [39]. Based on empirical analysis [91], the authors propose the exponential function $f(c_i) = 10^{-c_i}$.

The local centrality with a coefficient, denoted as CLC by Zhao et al. [90], is a combination of the previous CR and LC methods. The number of neighboring nodes is measured to identify cluster centers, and is combined with a decreasing function f for the local clustering coefficient of nodes, called the coefficient of local centrality $c(v_i)$, namely $f(c(v_i)) = e^{-c(v_i)}$ [39]. Mathematically, the influence of node v_i is measured as:

$$CLC(v_i) = f(c(v_i)) \cdot LC(v_i) \tag{2.3}$$

Considering the global information of the graph can give better insights, so we adopt the widely used betweenness Btw and closeness Cls centralities [10, 39]. Betweenness of a node v_i is expressed as the fraction of shortest paths between node pairs that pass through the node v_i, and is defined as [64]:

$$Btw(v_i) = \sum_{i \neq j \neq k \in G} \frac{\sigma_{jk}(v_i)}{\sigma_{jk}} \tag{2.4}$$

where σ_{jk} is the number of shortest paths between nodes v_j and v_k, and $\sigma_{jk}(v_i)$ denotes the number of shortest paths between v_j and v_k which pass through node v_i [39].

Closeness centrality of a node v_i is defined as the inverse of the sum of distances to all other nodes in G; it can be considered as a measure of how long it will take to spread information from a given node to other reachable nodes in the network [10, 39]:

$$Cls(v_i) = \left(\sum_{v_j \in G \setminus v_i} d(v_i, v_j) \right)^{-1} \tag{2.5}$$

Location-based measures

Location-based measures also require the structural information of the graph, but focus around the belief that the location of a node in a network is a more relevant. Driven by the limitations of simple graph metrics, such as degree centrality, Kitsak et al. propose *k-core decomposition* to quantify a node's influence based on the assumption that nodes in the same shell have similar influence, and nodes in higher-level shells are likely to infect more nodes [92]. To this end, the k-core decomposition method was validated by several studies [92, 93]. . While this method is often found in literature under both the names of k-core or k-shell decomposition, the two concepts differ. The k-core of a graph is the maximal sub-graph such that every vertex has degree at least k. A k-shell (KS), on the other hand, is the set of vertices that are part of the k-core but not part of the $(k+1)^{th}$-core [39].

Experiments show that by running a diffusion process on the network (*e.g.* SIR), the nodes with the same k_s values always have different number of infected nodes, namely spreading influence [94]. This phenomena suggests that the k-core decomposition method is not appropriate for ranking the global spreading influence of a network. Liu et al. [94] propose to solve this observed drawback by taking into account the shortest distance between a target node and the node set with the highest k-core value. In terms of the distance from a target node v_i to the network core G_c, the spreading influences of the nodes with the same k-core values can be distinguished using the following equation [39]:

$$\theta(v_i|k_s) = (k_s^{max} - k_s + 1) \sum_{v_j \in G_c} d_{ij}, \quad i \in G_{k_s} \tag{2.6}$$

In Equation 2.6, k_s^{max} is the largest k-core value of G, d_{ij} is the shortest distance from node v_i to node $v_j \in G_c$, G_c is the network core, and G_{k_s} is the node set whose k-core values equals k_s [39].

We will also refer to the Hirsch-index. The h-index HI [95] is a hybrid location-local-based centrality in which every node needs only a few pieces of information: the degrees of its neighbors. It was originally developed as a means to measure the scientific impact of scholars, but it now finds uses in quantifying the influence of users in social networks, or drugs in pharmacological interaction maps. The h-index of a node v_i is defined as the largest value h so that v_i has at least h neighbors with a degree $\geq h$ [39].

The algorithm is intuitive to apply, namely, for a node v_i with vicinity N_i, we order all its neighbors $v_j \in N_i$ in descending order of their degree k_{v_j}. The h-index $HI(v_i)$ is the position $h-1$ in the ordered list of nodes at which the degree of a neighbor becomes smaller than the position in the list. For example, given the list of degrees $L(v_i) = \{10, 8, 7, 6, 3, 1, 1\}$, we deduce $HI(v_i) = 4$, because $L(v_i)[4] > 4$, but $L(v_i)[5] < 5$ [39].

Diffusion-based measures

Diffusion-based measures are based on obtaining a state of balance in the network after applying a non-deterministic spreading processes, like a random walk. We make use of the fundamental Eigenvector centrality EC, which supposes that the influence of a node is not only determined by the number of its neighbors (*i.e.*, degree centrality), but also by the influence of each neighbor [96]. Inspired by EC, there are three additional algorithms we discuss in this thesis [39].

PageRank (PR) was first implemented as a random walk on the network of hyperlinks between web-pages [97]. A damping factor d is introduced as the probability for a user to jump to a random website, and $1 - d$ is the probability for the user to continue browsing through hyperlinks. The influence $s_t(v_i)$ of a node v_i at time t is given by:

$$PR(v_i) = \frac{1-d}{|V|} + d\left(\sum_{v_j \in G} \frac{PR(v_j)}{k_j^{out}}\right) \tag{2.7}$$

where $|V|$ is the number of nodes in G, k_j^{out} is the out-degree of node v_j, and $d = 0.85$, but d requires step-wise optimization based on the network [39].

$HITS$ is similar to PR, based on the concept that good hub nodes will point to good authority nodes, and good authorities will point by good hubs [98]. The hub score of all nodes at time $t = 0$ is initialized with 1; the authority score $Aut_t(v_i)$, at any moment in time t, is expressed as:

$$Aut_t(v_i) = \sum_{v_j \in G} a_{ji} \cdot Hub_{t-1}(v_j), Hub_t(v_i) = \sum_{v_j \in G} a_{ji} \cdot Aut_t(v_j) \tag{2.8}$$

Finally, the LeaderRank (LR) algorithm represents an improvement over PR, since the probability parameter is adaptive, leading to a parameter-free algorithm directly applicable on any type of complex network [99]. The method is applied by adding an additional ground node v_g that is connected to all other nodes, ensuring the graph is connected. A random walk then adds a score of $+1$ to each visited node v_i [39]. The ground node starts with $s_g(0) = 0$, and all other nodes in G have $s_i(0) = 1$. Using the notation $s_t(v_i)$ at time t for a node v_i, the evolving score can be expressed as:

$$s_{t+1}(v_i) = \sum_{v_j \in G} p_{ij} s_t(v_j) = \sum_{v_j \in G} \frac{a_{ij}}{k_i^{out}} s_t(v_j) \tag{2.9}$$

The score $s_t(v_i)$ is proven to converge towards a steady state at time t_c [99]; the score of the ground node is then evenly distributed to all other nodes $V \in G$ to conserve the scores on the nodes of interest [39]. The final, stable LR score is expressed as:

$$LR(v_i) = s_{t_c}(v_i) + \frac{s_{t_c}(v_g)}{|V|} \tag{2.10}$$

2.2 Network Growth using Betweenness Preferential Attachment

The dynamics of social networks is a complex process, as there are many factors that contribute to the formation and evolution of social links. Currently, there is no accurate model to provide a full understanding of social network dynamics, even if some real-world social network properties (e.g., the scale-free property) are captured by the degree-driven preferential attachment (DPA) model. Nevertheless, other important properties such as community formation, link weights, or degree saturation can not be completely explained [27].

Degree Preferential Attachment (DPA) is considered to be one of the key factors for complex network emergence and evolution [10, 3]. The scale-free topologies generated with the BA model can reproduce real-world social network properties such as low average path length APL and power-law degree distribution $P(k) = k^{-\beta}$, but DPA has its own limitations [27]. First, real-world social networks are typically weighted; the BA model does not work with weighted links [100]. Also, the BA model does not accurately describe how people connect and how their social ties evolve over time [100, 101, 102]; this is because:

- People are limited, form a psychological and physical point of view, to a maximum number of friendships in the real-world; this fact suggests a saturation on node degree [103, 104]. Conversely, in the BA model no such limit exists.
- People develop weighted relationships, meaning that not all ties are equally important. Studies show that a person knows, on average, about 350 persons, can maintain active contact with 150 people (Dunbar's number) [103], but only has very few strong ties [105]. The BA model does not account for such weights.

We start with a topological analysis on a variety of real-world network datasets and show that node betweenness (Btw) is power-law distributed and tightly correlated with both node degree (Deg) and link weight distributions. We note that our findings are supported by previous research on some particular cases of social networks [102, 106]. To further investigate the significance of betweenness, we experimentally test several alternative centralities as possible drivers for the preferential attachment models. We conjecture that: (i) Node betweenness is the main drive for new social ties, as opposed to degree or any other centrality metric, and (ii) considering the weight of social ties is paramount for an accurate description of social networks in the real-world [27].

Our main theoretical contribution is the introduction of the Weighted Betweenness Preferential Attachment (WBPA) model which is an intuitive and fundamental mechanism able to reproduce realistic social network topologies more accurately than state-of-the-art models based on DPA or other specific network parameter tuning. We explain WBPA's accuracy from a socio-psychological perspective which emphasizes node betweenness as the crucial factor behind the emergence of social networks [27].

In all datasets, node degree, node betweenness, link betweenness, and link weights are power-law distributed. Moreover, the power-law slope of degree distribution is steeper in comparison with node betweenness distribution. More precisely, as presented in Figure 2.1a, the average degree slope is $\beta_{deg} = 2.097$ (standard deviation $\sigma = 0.774$) and the average betweenness slope is $\beta_{btw} = 1.609$ ($\sigma = 0.431$), meaning that β_{deg} is typically 30.3% steeper than

β_{btw} across all datasets. For all considered datasets there is a significant non-linear (polynomial or exponential) correlation between node betweenness and node degree (see Figure 2.1b); this further suggests that node betweenness may be the source of imbalance in node degree distribution.

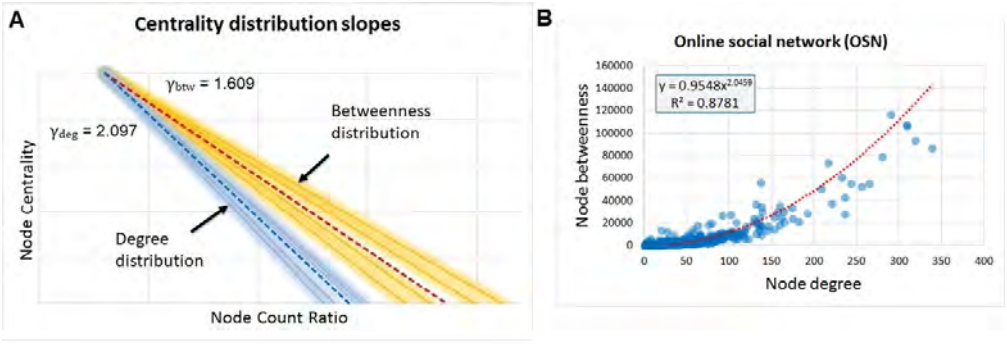

Figure 2.1: (a) Overview of centrality distribution slopes for all empirical datasets, highlighting the average slopes for degree (blue), and betweenness (red). (b) The representative non-linear correlation of Btw and Deg. These results show that, in social networks, Deg and Btw have a power-law distribution (with a steeper slope for degree), and that there is a non-linear correlation between the two centralities.

The fundamental difference between the degree-driven and betweenness-driven preferential attachment is illustrated in Figure 2.2a.; the upper panel shows that, under DPA, the nodes with high degree (colored in orange) will gain an even higher degree. The lower panel in Figure 2.2b shows that under BPA the nodes with high betweenness (orange) will attract more links and increase degree, which in turn will redistribute and decrease betweenness, thus limiting the number of new links per hub as a second order effect. This can explain why, in real-world networks, the number of new links is limited for high degree nodes. Figure 2.2b explains the proposed WBPA algorithm, step by step. As such, (a) all bidirectional links E in graph G are initialized with weights w_{ij}, respectively w_{ji}. Each outgoing link weight of node v_1 is proportional to the fitness (indicated as $w \sim f$) of the target neighbor nodes, and then normalized such that the sum of outgoing weights is 1; (b) A new node v_6 connects to existing ones v_1-v_5 based on probabilities proportional to the normalized fitness ($p \sim f$) of the target nodes. Say, v_6 connects only to v_1 based on fitness f_1; (c) Once v_6 and v_1 connect, node v_1 assigns a weight w_{1-6} on the new link that is proportional to fitness f_6. As such, a proportional weight ratio of $w_{1-6}/4$ is subtracted (indicated with a minus sign) from the four already existing links. If any of the newly resulting weights drop below 0, the corresponding link is removed from node v_1. According to the BPA principle, fitness f is represented by the betweenness centrality.

In order to measure the similarity in terms of network parameters and centralities, between the synthetic complex networks, generated according to each algorithm, and the real-world network reference, we introduced the network fidelity metric φ [55, 37] as:

$$\varphi^j = \begin{cases} \frac{1}{n}\sum_i \frac{\overline{m_i}}{2\overline{m_i}-\overline{m_i^j}} & if\ \overline{m_i^j} < \overline{m_i} \\ \frac{1}{n}\sum_i \frac{\overline{m_i}}{\overline{m_i^j}} & if\ \overline{m_i^j} \geq \overline{m_i} \end{cases} \tag{2.11}$$

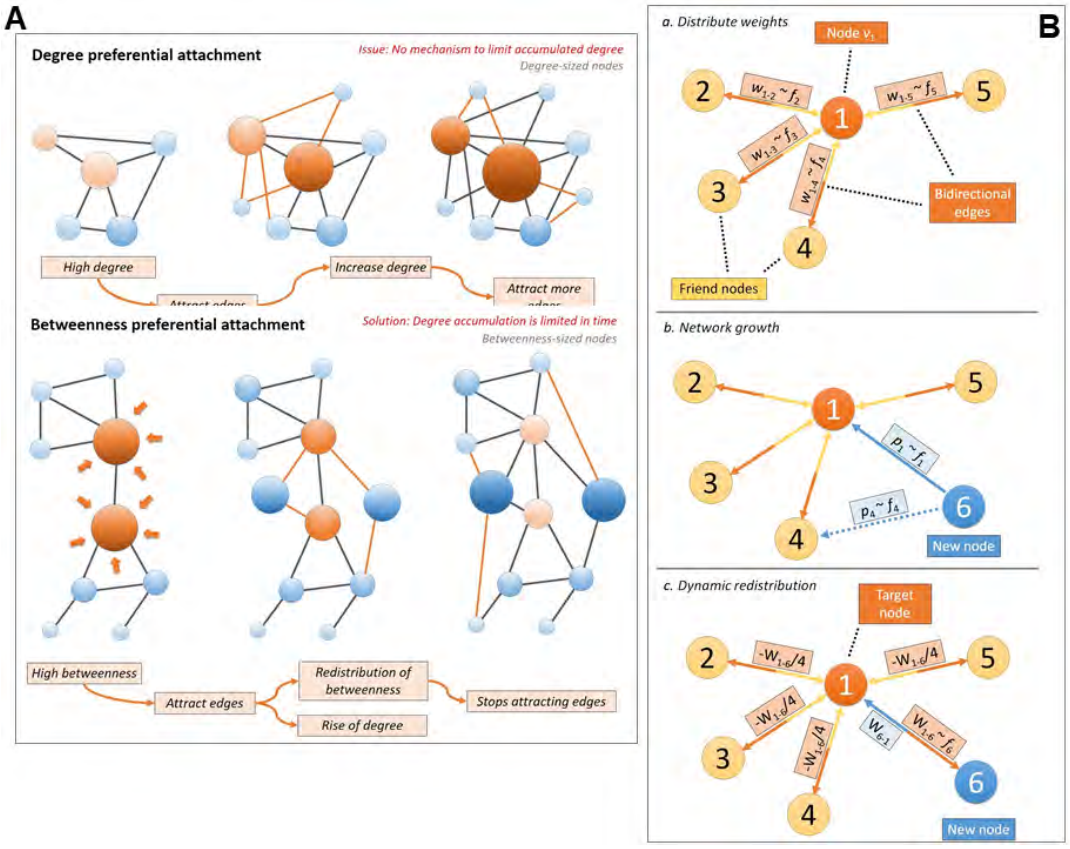

Figure 2.2: (a) The mechanisms of degree preferential attachment (DPA) versus betweenness preferential attachment (BPA) depicted in terms of acquiring new links and limiting the (excessive) accumulation of degree over time. In DPA, nodes with high degree attract even more links, and thus increase degree *ad infinitum*. Conversely, in BPA, nodes attracting new links because of their high betweenness will eventually lose betweenness to neighboring nodes, thus limiting the acquired degree. (b) Network evolution according to the Weighted BPA algorithm.

In equation 2.11, j represents the index of the network being compared to the reference network. The index of the network metric which describes the two compared models (e.g., average path length, average degree etc.) is denoted by $i = \{1, 2, ...n\}$, where n is the total number of common metrics taken into consideration. Fidelity takes values between 0 and 1 (or as percentiles), with 1 representing perfect similarity. The metric measurements on the reference model are m_i, respectively m_i^j on the model being compared [46].

A summary of the results is given in Table 2.2, where the upper half contains the average fidelity φ [37] of WBPA, DPA and the two null model networks, towards the real-world reference networks. The lower half of Table 2.2 contains the other state of the art synthetic networks. Our WBPA obtains the highest fidelity towards the empirical references, e.g., 13-68% higher φ_{FB}, 21-81% higher φ_{OSN}, 4-47% higher φ_{TK} than all other synthetic models. As such, we prove the increased realism of our model in comparison with some elaborated state-of-the-art models. Compared to DPA, our model produces networks with higher fidelity values; when averaged over all empirical networks we obtain: $\overline{\varphi}_{Btw} = 0.831$ and $\overline{\varphi}_{Deg} = 0.777$.

Table 2.1: Statistical fidelity φ of WPBA, DPA, two *Null* models (random and small-world), and four state of the art network (Cellular, Holme-Kim, Toivonen, Watts-Strogatz with degree distribution) models, obtained by comparing the topologies with multiple real-world datasets. Values in bold represent the highest fidelity on each column (*i.e.*, most realistic topology).

Datasets	φ_{FB}	φ_{GP}	φ_{CoAu}	φ_{OSN}	φ_{BTC}	φ_{MOvr}	φ_{HEP}	φ_{POK}	φ_{EmE}
WBPA	**0.835**	**0.842**	0.735	**0.801**	**0.897**	**0.814**	**0.845**	**0.771**	0.837
DPA	0.694	0.796	**0.778**	0.634	0.754	0.692	0.836	0.758	**0.851**
Rand	0.681	0.719	0.681	0.597	0.816	0.761	0.779	0.754	0.733
SW	0.737	0.718	0.705	0.554	0.644	0.579	0.603	0.669	0.769
Cell	0.543	0.707	0.637	0.52	0.566	0.559	0.503	0.508	0.792
HK	0.704	0.778	0.578	0.66	0.687	0.679	0.522	0.577	0.787
Tvn	0.638	0.676	0.711	0.55	0.571	0.561	0.558	0.601	0.831
WSDD	0.497	0.708	0.673	0.443	0.547	0.535	0.511	0.556	0.825

We consider that BPA transcends the mere topological perspective on social relationships evolution [27]. In the field of social psychology, individuals are perceived as *social creatures* who strive for social recognition, validation, approval and fame [107, 100, 22, 108]. As such, individuals tend to connect to either individuals who are popular in their communities (*i.e.*, typically they have high degree), or individuals who connect multiple communities (having high betweenness). The former type of connection is mostly related to the popularity of individuals within local communities, and appears to be an epiphenomenon of the latter [27].

Towards this end, we introduce the concept of *social evolution cycle* [27], which revolves around betweenness assortativity rather than degree assortativity [108, 109, 110]. According to our approach, individuals become more influential over time by increasing their own betweenness. Therefore, the exhibition of one individual's desire to increase his betweenness will (i) attract new ties (*i.e.*, increase in degree), and will create stronger ties (*i.e.*, increase in link weight); this process continues for the next generation of individuals who aspire to climb the social ladder. As shown, this conclusion is supported by the evolution of networks generated with WBPA [27].

We conclude that the WBPA model is quantitatively more robust than DPA, as it can reproduce more accurately a wide range of real-world social networks. Also, WBPA explains

the dynamic accumulation of degree and link weights, as well as the eventual degree saturation, as a second order effect. Consequently, we believe our work paves the way for a new and deeper understanding of the mechanisms that lie behind the dynamics of complex social networks [27].

2.3 Structural Antifragility Under Sustained Attack

Antifragility is a counter-intuitive property of systems, which makes them even stronger when being subjected to stressors such as attacks, volatility, or errors. The term was introduced by N.N. Taleb [111] to describe a system that actually benefits from being exposed to attacks, thus growing stronger (up to a point). Indeed, as opposed to conventional concepts such as robustness or resiliency, antifragility does not represent a mere resistance to attacks [112].

The exploration of antifragility is relatively new to Network Science, having been only recently addressed in [113, 38]. Similar to [114, 115], we interpret a *robust* network as being characterized by a high connectivity between nodes, whereas a *fragile* network as having a low connectivity. Accordingly, *antifragility* increases network connectivity when subjected to attacks. Quantitatively, we measure the robust, fragile, and antifragile behaviors using the largest connected component size (LCS) and the number of connected components (NCC), as these parameters are directly related to network's communication capacity [116, 117].

To uncover the topological features that foster antifragile behavior in complex networks, we simulate multiple attack–repair scenarios on some generic synthetic topologies (random, mesh, small-world and scale-free), as well as real-world network topologies. Accordingly, the *antifragile* network behavior is detected at macro-scale if, as simulation unfolds, the connectivity of the network (measured via LCS and NCC variations) does increase under sustained attacks. The *fragile* network behavior under attack corresponds to the network's reduced tolerance to incurred faults (i.e., destroyed links), leading to degraded LCS and NCC [15, 118, 119]. More precisely, a network is considered fragile if its LCS decreases rapidly during simulation, and antifragile if its LCS increases during the attack-repair process up to a specific stress point. To consider this scenario, we start our simulations with disconnected networks ($NCC > 1$); depending on the attack-repair ratios, the LCS may – counter-intuitively – increase, even though the network is losing edges overall (e.g., the attacks mostly remove edges in smaller components, while edge repairs connect the new nodes to the largest connected component). This is the *antifragile* effect that we intend to quantify [38].

The modeled edge attacks imply that, during each iteration, a fixed ratio α of edges (which we call the attack rate) is removed. In [113], four α values are considered ($\alpha \in \{1\%, 2\%, 5\%, 10\%\}$) to conclude that $\alpha = 0.05$ (5%) is the optimal trade-off between speed and amplitude of network destruction. State of the art attack strategies focus primarily on node centralities [10, 6], but random attacks are also used. The response after each attack is a set of edge repairs which happen with rate β. We analyze two different simulation settings: one in which $\beta < 1$ (i.e., fewer edges are repaired than destroyed), and another in which $\beta \geq 1$ (i.e., more edges are repaired than destroyed).

For each computer simulation, we obtain two time series, namely the evolution of $LCS(t)$ and $NCC(t)$ over 100 iterations, $t \in \{1, ..., 100\}$. To quantify the antifragile response, we use two intuitive measures:

- The *maximum improvement* (I) of LCS for repair rate β, based on the maximum of ensemble averages $I_e(t = k)$, is defined as

$$I_e(t = k) = \frac{\text{average}\{LCS(0 < t \leq k)\}}{LCS(t = 0)} \ , \ I_\beta = \max\{I_e(t)\}. \tag{2.12}$$

- For a repair rate β, when $I > 1$ (i.e., antifragility is present), the *duration of antifragility* D_β is the time interval when $LCS(t) \geq LCS(t = 0)$,

$$D_\beta = \{t_2 - t_1 \mid LCS(t) \geq LCS(t = 0),\ t_1 \leq t \leq t_2\}. \tag{2.13}$$

If the simulation exhibits any antifragile behavior, then $I \geq 1$ and $0 < D \leq 100$. If the edge repair rate is higher than the attack rate, we obtain a high duration, $D \approx 100$. However, to maintain a robust topology with minimal repair costs and limited resources for response, we are most interested in scenarios where $D > 0$ for a repair rate of $\beta < 1$.

Our attack-repair mechanism implies edge repairs at every iteration. In the real world, these edge repairs would incur corresponding costs. For instance, either we consider adding or repairing power lines, creating new physical links between routers, or establishing new social links, we need to minimize the cost of repairs [38].

We define the absolute cost at iteration t as the sum of degrees for all target nodes receiving new links in that iteration $costAbs\,(t) = \sum_j k_j\,(t)$ (where $k_j\,(t)$ is the degree of node v_j which receives a new link at iteration t). Further, for each iteration, we define the absolute repair efficiency as the gain/cost ratio $LCS(t)/costAbs(t)$ [38].

In Table 2.2, we provide the improvements I and duration D measured on all datasets in the context of random ($Rand$), degree (Deg), betweenness (Btw), and Eigenvector (Eig) paired attack-repairs with repair rates of $\beta = 0.7$ (*i.e.,* reduced repair rate) and $\beta = 1$ (*i.e.,* balanced repair rate).

By analyzing the data in Table 2.2, we find that antifragility does occur in our simulations, as LCS increases for a limited period although the network loses more edges than it regains ($\beta < 1$).

The emergence of antifragile responses in synthetic and real-world networks seems to follow a correlation with the complexity of the underlying topology. Namely, the real-world networks (especially the natural ones) show the highest antifragile improvement of $I \approx 1.0 - 78.8$, followed by WD ($I = 17.08$), then SW ($I = 1.24$), SF ($I = 1.03$), and finally ER and Me ($I < 1$).

In general, we conclude that the paired *random* repair-attacks are the best combination for triggering an antifragile behavior in both synthetic and real-world networks. Second, the betweenness attacks consistently rank as the most destructive strategy overall, regardless of the repair strategy. Third, we find that the random strategy offers the highest improvements I, on average, with the degree strategy providing slightly longer antifragile duration D [38].

When comparing paired and non-paired attack-repair strategies on synthetic networks we conclude that:

- In terms of efficiency of random repairs, the mesh (Me) topology has a different response than the other three topologies (ER, SW, SF).

Table 2.2: Topological improvement I and antifragile duration D (in parentheses) for paired centrality attack-repairs on each network, with $\beta = 0.7$ (upper half) and $\beta = 1$ (lower half). A higher I denotes a stronger antifragility, $I < 1$ means no antifragility. A higher D value indicates a longer response measured as the number of attack-repair rounds, a dash (–) means no antifragile response. The antifragile behaviors are shown in bold.

	Network	*Rand*	*Deg*	*Btw*	*Eig*
$\beta = 0.7$	*ER*	0.98 (-)	0.97 (-)	0.94 (-)	0.94 (-)
	Me	0.99 (-)	0.98 (-)	0.95 (-)	0.96 (-)
	SW	**1.19 (17)**	**1.24 (21)**	0.73 (-)	**1.07 (9)**
	SF	**1.03 (16)**	0.99 (-)	0.97 (-)	0.99 (-)
	WD	**17.08 (85)**	**14.76 (100)**	**12.88 (55)**	**11.83 (56)**
	UP	0.98 (-)	0.96 (-)	0.95 (-)	0.97 (-)
	Rt	**1.0 (10)**	0.99 (-)	0.99 (-)	1.0 (-)
	Em	**1.03 (1)**	**1.02 (1)**	**1.03 (1)**	**1.03 (2)**
	Mo	0.99 (-)	0.98 (-)	0.98 (-)	0.99 (-)
	Tw	**78.82 (68)**	**74.49 (96)**	**49.06 (20)**	**67.31 (55)**
$\beta = 1$	*ER*	**1.03 (100)**	**1 (1)**	0.94 (-)	0.96 (-)
	Me	**1.02 (100)**	0.99 (-)	0.95 (-)	0.95 (-)
	SW	**1.49 (100)**	**1.35 (100)**	0.85 (-)	**1.17 (100)**
	SF	**1.21 (100)**	**1.0 (5)**	0.98 (-)	0.99 (-)
	WD	**18.02 (100)**	**14.97 (100)**	**13.84 (100)**	**13.04 (100)**
	UP	**1.0 (6)**	0.98 (-)	0.98 (-)	0.98 (-)
	Rt	**1.10 (68)**	0.99 (-)	0.99 (-)	0.99 (-)
	Em	**1.00 (3)**	**1.0 (7)**	**1.0 (1)**	**1.0 (8)**
	Mo	**1.02 (91)**	0.99 (-)	0.98 (-)	0.98 (-)
	Tw	**81.73 (100)**	**75.22 (100)**	**57.03 (100)**	**70.19 (100)**

- There is a transition around $\beta = 1$ between the efficiency of *Deg* (paired) versus *Rand* (non-paired) repairs. On meshes, *Deg* is more efficient than the other repair strategies for $\beta > 1$ and less efficient than *Rand* for $\beta < 1$. The opposite is true for the other topologies.

Furthermore, we analyze the costs for topologies where antifragility is observed. In Figure 2.3 we depict the scaling of the proposed edge repair cost ratio: $LCS(t)/costAbs(t)$ on the *SW* and *SF* networks, and on the *Mo* and *Tw* networks respectively.

All the cost efficiency plots ($LCS/costAbs$) show that the random strategy is better at first but – as the simulation progresses and a significant number of edges is lost – the network becomes more fragile and, in this context, the degree-driven *Deg* strategy becomes more efficient [38]. We validate the empirical results for real-world networks using their corresponding rewired versions, which preserve the number of nodes, the number of edges, and the degree distribution (see the dashed random rewiring (RR) and preferential rewiring (PR) lines in Figure 2.3) according to [120, 121].

We have first shown that antifragility is more pronounced on more complex synthetic topologies such as *WD* and real-world networks. We then have found that the random tar-

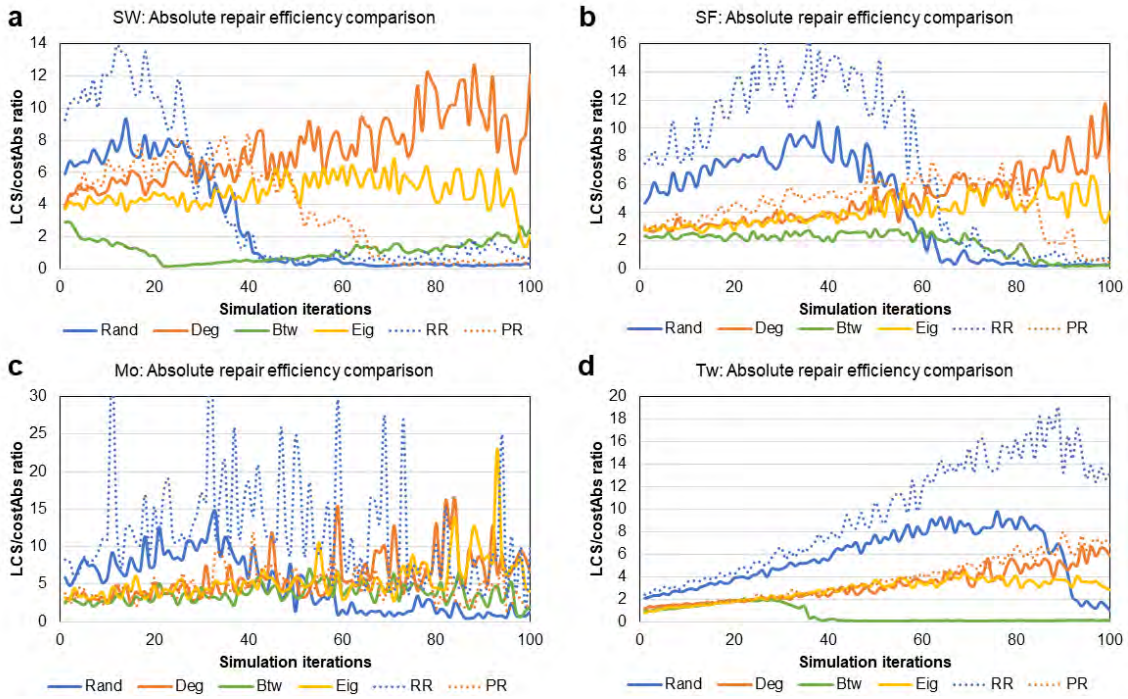

Figure 2.3: Scaling of $LCS(t)/costAbs(t)$ on the SW (a), SF (b), Mo criminal (c), and Tw Twitter (d) networks using four different paired attack-repair strategies. All plots indicate that the random *Rand* strategy is initially better, but as the network loses links, the centrality-driven strategies (especially *Deg*) become more efficient.

geting repair strategy provides the highest improvements at first, thus confirming the theory stating that antifragility appears in the context of random solution searches, rather than deterministic ones [111]. We have also found that betweenness-driven attacks are the most destructive on all tested datasets. Another important observation is that natural real-world topologies have a stronger drive towards antifragility than their technological counterparts. Finally, the efficiency analysis based on costs shows that the random strategy is initially better but, as the network becomes more damaged, the degree-driven HDF (high-degree first) strategy becomes more cost-effective. Taken together, these results suggest that, for network systems that require a high resilience, the evolutionary strategy of trying solutions at random and then letting the environment perform the selection is more efficient when the system is not too damaged and has enough time to react; otherwise, preferential attachment works best [38].

We hope that our findings will stimulate new research towards developing dynamic edge reconfiguration models based on the principle of antifragility. Further research will need to consider more sophisticated repair strategies based on different node centralities.

2.4 Network Centrality Analysis and Benchmarking Influence Rankings Methods

The development of new methods to identify influential spreaders in complex networks has been a significant challenge in network science over the last decade. Practical significance spans from graph theory to interdisciplinary fields like biology, sociology, economics and marketing. Despite rich literature in this direction, we find small notable effort to consistently compare and rank existing centralities considering both the topology and the opinion diffusion model, as well as considering the context of *simultaneous* spreading. To this end, our study introduces a new benchmarking framework targeting the scenario of *competitive opinion diffusion*; our method differs from classic SIR epidemic diffusion, by employing competition-based spreading supported by the realistic tolerance-based diffusion model. We review a wide range of state of the art node ranking methods, and apply our novel method on large synthetic and real-world datasets [39].

Novel approaches, combined with classic graph centrality measures have led to the emergence of the three main categories of influence ranking methods [39]. A first category of scientists argue that the location of a node is more important than its immediate ego-network, and thus proposed k-core decomposition [92, 93], along with improved variants, such as [122, 123, 94, 124]. A second category of scientists quantify the influence of a node based solely on its local surroundings [88, 91, 125]. Finally, a third category of scientists evaluate node influences according to various states of equilibrium for dynamical processes, such as random walks [99, 89], or step-wise refinements [126]. Examples of commonly used measures of node importance include node degree, node centralities (betweenness, closeness, PageRank, HITS authority, Eigenvector), or node vulnerability (in dynamic context) [10, 127, 5].

State of the art benchmarking methodologies for spreading processes on complex networks often rely on the SIR(SIS) model [128, 7]. With this approach, an initial subset of nodes is infected according to a centrality measure, then the simulation measures how fast surrounding susceptible nodes become recovered (*i.e.*, including dead). Indeed, if we take the example of an

epidemic, it spreads independently from other epidemics, and has its own temporal evolution. On the other hand, if we consider opinion between social agents, it is often exclusive (in regard to other contradicting opinions), and is also dependent on the timing with the spread of other ideas [39].

We set out to discover fundamental drivers in the underlying graph structure which shape and influence opinion spreading in complex networks. To this end, our experimental setup is focused on a comparative benchmark analysis involving the reviewed node centrality metrics defined in Section 2.1.2. For an objective comparison, we make use of two types of datasets: synthetic data (10,000 node random, mesh, small-world and scale-free networks [10]), and real-world data (consisting of large, representative complex networks sized between 1,900 and 29,000 nodes).

We let each of the $n = 10$ selected centrality measures compete in a one-to-one scenario over the 4 synthetic and 4 real-world datasets. Every dataset comprises a total of $n \times (n-1)/2 = 45$ pairs of simulations, translating into $2 \times 45 = 90$ individual simulations due to alternating the selection of spreaders. For statistical rigor, each experiment is repeated 10 times, consisting of a simulation batch of 20 simulations, leading to $45 \times 20 = 900$ simulations per dataset, amassing to an overall $8 \times 900 = 7,200$ unique experiments. Condensing the simulation results, we present in Table 2.3 the average performance of the 10 ranking methods on the 8 datasets. This performance is quantified as an average percentage of opinion coverage ρ obtained from the one-to-one competition benchmarks (*e.g.*, HITS obtains a coverage of 64.42% on the OSN dataset).

Table 2.3: Average performance of the 10 ranking methods on the 8 datasets. Performance is expressed as opinion coverage (%) obtained in the one-to-one opinion diffusion competitions with every other ranking method.

Centrality	Rand	Mesh	SW	SF	OSN	FB	Emails	POK
Deg	66.18	71.26	68.94	61.71	52.76	56.18	63.52	63.28
Cls	23.02	5.47	11.39	1.83	2.55	11.49	2.40	45.78
Btw	66.15	42.93	56.96	62.78	40.37	57.51	58.33	58.27
HITS	66.28	69.32	76.92	61.63	64.42	62.10	63.56	63.09
PR	77.16	65.35	71.93	55.74	41.08	55.99	63.55	63.94
HI	12.13	52.82	33.25	54.72	24.23	41.36	39.60	36.30
LR	76.95	67.57	66.72	61.53	64.39	68.06	63.97	66.87
KS	0.99	39.65	37.87	45.89	28.77	28.87	42.07	13.33
CLC	33.93	52.36	60.24	26.99	44.74	55.91	48.43	48.01
EC	23.12	32.96	39.43	43.09	62.83	44.54	51.49	32.27

Similar to the state of the art SIR epidemic benchmarking, our obtained results are easy to understand, and offer the possibility of direct comparison between ranking methods on the same dataset. On the other hand, we notice two improvements by applying our methodology:

1. There is much higher variation between measures on the same dataset. For example, on the FB dataset we obtain Deg= 56.18% and Cls= 11.49%, which suggest an obvious performance difference. On the other hand, using SIR as benchmark, the coverages are $\rho_{Deg} = 95.31\%$ and $\rho_{Cls} = 95.17\%$.

2. There is greater emergent granularity between measures on different datasets. For example, Cls turns out to be much less efficient on a SF topology (1.83%) than on a SW topology (11.39%) [39].

Assessing the results in Table 2.3, we find an objective comparison of state of the art ranking methods used in current social networks research. The top three ranking methods, according to our original proposed methodology, are LeaderRank (LR), HITS, and node degree (Deg) [39].

Additionally, we provide a suggestive visual example of the opinion coverages at the end of a simulation, after balancing is attained [59] with our used tolerance diffusion model [34]. The Mesh topology is exemplified here because it offers the most intuitive 2D spatial feedback after applying a Force-directed layout. To this end, Figure 2.4 shows the coverage of competing centrality measures in three different scenarios:

- Two ranking methods with high overlapping and balanced outcome: Deg (orange) 56.70%–LR (blue) 43.30% (Figure 2.4a).
- Two ranking methods with moderate overlapping, and inefficient seed selection for one method (Btw) : LR (orange) 74.26% – Btw (blue) 25.74% (Figure 2.4b).
- Two ranking methods with low overlapping and extreme outcome: Cls (orange) 5.24% – HI (blue) 94.76% (Figure 2.4c).

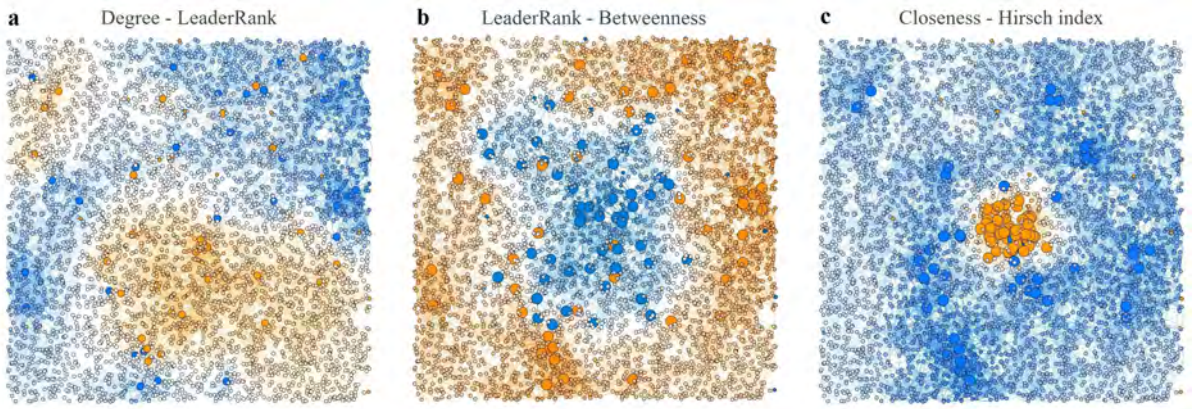

Figure 2.4: Three opinion diffusion benchmarks highlighting the final opinion coverage over the *Mesh* dataset ($N = 10,000$). Orange nodes are influenced more by the first ranking method, and blue nodes are influenced more by the second ranking method; whiter nodes are closer to indecision (50%); larger nodes represent seeders (1% of N).

Finally, to highlight the superior quantitative power of our competition-based benchmark we aggregate the results in Figure 2.5. When trying to discern between the top 2 ranking methods on a particular dataset, SIR manages to place them apart by only $\approx 0-1.07\%$ (0.31% on average), while our method manages to produce higher differences within $\approx 0.28 - 8.75\%$ (3.56% on average). Another advantage of our proposed method is the overall uniformity obtained for the performances of each centrality across the 8 selected datasets. For instance,

if LR and HITS result as the most efficient spreading methods on one topology, their performance is replicated with high confidence on the other topologies as well. When employing SIR benchmarking, the performances are not consistent across datasets. This aspect is suggested visually in Figure 2.5, where we highlight the most (LR) and least (Cls) efficient centralities, as they are ranked over the 8 datasets. It is easy to notice how LR is positioned in the top 3 and Cls in the last 2-3 methods overall. In the individual SIR benchmarking, there is no such uniformity.

Individual benchmarking

	Random	Mesh	SW	SF	OSN	FB	Emails	POK
1	HI	EC	HITS	Cls	Cls	Deg	Cls	EC
2	HITS	CLC	HI	CLC	EC	HITS	PR	HI
3	KS	Deg	PR	Btw	CLC	CLC	LR	Btw
4	Btw	HI	KS	LR	Deg	Btw	Deg	CLC
5	LR	PR	EC	PR	Btw	LR	Btw	KS
6	PR	HITS	Btw	Deg	KS	PR	HITS	PR
7	EC	Btw	Deg	HI	HITS	HI	EC	Deg
8	CLC	Cls	Cls	KS	LR	EC	CLC	Cls
9	Cls	KS	CLC	EC	PR	KS	HI	LR
10	Deg	LR	LR	HITS	HI	Cls	KS	HITS

Competition-based benchmarking

Random	Mesh	SW	SF	OSN	FB	Emails	POK	
PR	Deg	HITS	Btw	HITS	LR	LR	LR	1
LR	HITS	PR	Deg	LR	HITS	HITS	PR	2
HITS	LR	Deg	HITS	EC	Btw	PR	Deg	3
Deg	PR	LR	LR	Deg	Deg	Deg	HITS	4
Btw	HI	CLC	PR	CLC	PR	Btw	Btw	5
CLC	CLC	Btw	HI	PR	CLC	EC	CLC	6
EC	Btw	EC	KS	Btw	EC	CLC	Cls	7
Cls	KS	KS	EC	KS	HI	KS	HI	8
HI	EC	HI	CLC	HI	KS	HI	EC	9
KS	Cls	Cls	Cls	Cls	Cls	Cls	KS	10

Figure 2.5: Visual representation of the uniformity in benchmarking influence ranking methods across different networks. We highlight the positions obtained by LR (top centrality in terms of spreading) and Cls (least effective centrality) across our 8 datasets in the context of individual (left panel) and competition-based (right panel) benchmarks. The position of a centrality on the vertical corresponds to its obtained rank (1-10) after benchmarking. *E.g.*, LR is 5^{th} best on Random and 10^{th} best on Mesh.

In conclusion, our benchmarking methodology – which is specifically designed for the *competitive* social network context – provides significant quantitative separation between influence ranking methods on synthetic and real social network topologies. This numerical separation is over one order of magnitude greater than the one provided by classic SIR simulation – a standard methodology used in epidemic spreading, where the diffusion context is less competitive, and more ego-centered. Therefore, we encourage the use of our proposed method in specific real-world applications of dynamic social networks.

Chapter 3

Contributions in Computational Network Analysis

3.1 Introduction

An important challenge in Network Science is the study and understanding of social opinion dynamics and personal opinion fluctuations [13, 129, 130]. The benefit of understanding these complex processes is also a major concern for research fields like psychology, philosophy, politics, marketing, finances and even warfare [9, 131, 132]. The distribution of opinion in a community, at a certain time, is a reflection of the distribution of socially influential people in that particular community [133, 134]. Social influence is the ability of persons (agents) to influence others in either one-on-one or group settings. Influential people motivate others to participate in certain activities, agree with their ideas and eventually follow their lead. Without social influence, the society would have a non-deterministic, erratic behavior which would be hard to predict. Political, religious, and community leaders use social influence to shape their communities. Consumer groups and public opinion influencers use social influence to motivate others member of the society to act and to build a unified effort towards an envisioned economic or political goal [10].

Marketing, for example, uses many techniques to understand the needs, strengths and weaknesses of different social layers or groups. Current research focuses on understanding when, how, where and why a product may be bought by people, and how the psychological factors behind this process can be influenced [135]. Like in most market studies on opinion formation and influence propagation, the buying process is modeled by combining elements from psychology, sociology, anthropology and economics [136, 9].

Computer simulation brings Computer Science & Engineering and mathematical modeling together in order to better understand the theory and characteristics of complex systems. Indeed, computer simulations help observe, evaluate, refine and enhance the models behind simpler theories. There are a series of diverse tools for visualizing graph data, like Gephi, Cytoscape, GraphViz, Pajek, iGraph, Tulip, GUESS, Neo4J, yED, Walrus, and the R language [75]. However, these tools offer in-built visualization engines, alongside layout, coloring, filtering and other options; some offer development as they are open source. Nonetheless, there is no available tool that brings a built-in framework for network simulations, particularly in the area of opinion dynamics.

Our solutions in computational network analysis (CNA) come in the form of a comprehensive opinion dynamics simulation framework which allows the study of multiple topological models, as well as customizable social interactions. Furthermore, we present a series of discrete event simulations by defining a more refined interaction model, and propose the concept of agent tolerance, a personality metric which evolves over time based on interactions with other agents. Our simulations discover a previously unobserved phenomenon, namely a phase transition in opinion dynamics that depends on varying concentration of opinion sources. These results could only have been observed using simulation. Consequently, our two main contributions in CNA are:

- We provide a probabilistic evaluation of our original tolerance-based social network interaction model [34]. Namely, a Markov chain model is used in order to assess the asymptotic behavior of social agents. In other words, the main result of this evaluation is the probability of tolerant behavior as the time approaches infinity.
- We take inspiration from micro-scale temporal epidemic models and develop an original time-aware (TA) forecasting methodology which is able to improve the prediction of opinion distribution in an electoral context.

3.2 Modeling and Simulation of Opinion Dynamics

Existing studies on opinion formation and evolution [136, 32, 137, 130, 138, 12] revolve around 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 [139], or thresholds evolving according to simple probabilistic processes that are not driven by the internal state of the social agents [140, 141]. To mitigate these limitations, the dynamical features of opinion spreading have to be targeted by a mathematical model. Many recurring real-world observations can be explained using the tolerance model introduced as a new social interaction model which takes into account the evolution of individual's inner state [34]. The model was validated using empirical data from Yelp, Twitter and MemeTracker, and by using our opinion dynamics simulation framework - SocialSim, which includes multiple complex topological models, as well as customizable opinion interaction and influence models [59].

3.2.1 The Tolerance-Based Agent Interaction Model

The tolerance model [34] is based on the classic voter model [28], being a refinement of the stubborn agent model [32, 137], with the unique addition of a *dynamic* decision-making threshold, called tolerance θ_i, for each node.

We further introduce the specific network science notations to mathematically define our model. Given a social network $G = \{V, E\}$, the neighborhood of node $v_i \in V$ is defined as $N_i = \{ v_j \mid (e_{ij}) \in E\}$. Exemplifying for a context with two competing opinions, we introduce two disjoint sets of stubborn agents V_0, $V_1 \in V$ which act as opinion sources. Stubborn agents 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 more of their direct neighbors. We represent with $x_i(t)$ the

opinion of agent v_i at time t. Normal (regular) agents start with a random opinion value $x_i(0) \in [0, 1]$. We represent with $s_i(t)$ the state of an agent v_i at moment t having continuous opinion $x_i(t)$. In case of a discrete opinion representation $x_i(t) = s_i(t)$, and in case of a continuous opinion representation $s_i(t)$ is given b equation 3.1.

$$s_i(t) = \begin{cases} 0 & if\ \ 0 \leq x_i(t) < 0.5 \\ none & if\ \ x_i(t) = 0.5 \\ 1 & if\ \ 0.5 < x_i(t) \leq 1 \end{cases} \tag{3.1}$$

In the assumed social network, agents v_i and v_j are neighboring nodes if there is an edge e_{ij} that connects them. Some agents may not have an opinion, or may not participate in the diffusion process (*i.e.*, $s_i(t) = none$), so interacting with these agents will generate no opinion update. A regular node will periodically poll one random neighbor (simple-diffusion), or all its neighbors (complex-diffusion), average the surrounding opinion $\bar{x}_{N_i}(t)$ (*i.e.*, vicinity N_i of an arbitrary node v_i, at time point t), and update its opinion $x_i(t)$ using a weighted combination of the past opinion and that of its neighbor(s), as:

$$x_i(t) = \theta_i \cdot \bar{x}_{N_i}(t) + (1 - \theta_i) \cdot x_i(t-1) \tag{3.2}$$

The tolerance θ_i parameter is the amount of accepted external opinion, and changes after each interaction based on whether a node has faced competing opinion, or supporting opinion (in a binary context with opinions A and B). Once a node is in contact with the same opinion for a long enough time, it becomes intolerant ($\theta_i(t) = 0$), so that the network converges towards a state of balance [59]. Opinion fluctuates, and is transacted by all nodes, but stubborn agents are the only nodes which do not become influenced in turn, acting as perpetual sources for the same opinion [32].

The evolution towards both tolerance or intolerance varies in a non-linear fashion, as an agent under constant influence becomes indoctrinated at an increased rate over time. If that agent faces an opposing opinion, he will eventually start to progressively build confidence in that other opinion. As such, the tolerance model employs a non-linear fluctuation function, unlike most models in literature [35, 36]. Based on realistic socio-psychological considerations in the dynamical opinion interaction model, we model tolerance evolution as:

$$\theta_i(t) = \begin{cases} max\ (\theta_i(t-1) - \alpha_0\varepsilon_0,\ 0) & if\ \ s_i(t-1) = s_j(t) \\ min\ (\theta_i(t-1) + \alpha_1\varepsilon_1,\ 1) & otherwise \end{cases} \tag{3.3}$$

Tolerance is decreased by $-\alpha_0\varepsilon_0$ if the state of the agent before interaction, $s_i(t-1)$, is the same as the state of the randomly interacting neighbor $s_j(t)$. If the states are not identical (*i.e.*, opposite opinion), then the tolerance will be increased with the dynamic product of $+\alpha_1\varepsilon_1$. The two scaling factors, α_0 and α_1, both initialized with 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, scaling factor α_0 is increased by $+1$ as long as an agent interacts with another agents having the same state (*i.e.*, $s_i(t-1) = s_j(t)$), and is reset to 1 otherwise. Scaling factor α_1 is increased as long as the interacting state is always different from that of the agent, as is reset if the states are identical. We introduced the scaling factors to model bias,

and 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 $\varepsilon_0 = 0.002$ and $\varepsilon_1 = 0.01$. We have determined these values as explained in [34].

3.2.2 Probabilistic Modeling of Tolerance-Based Interaction

In order to provide a probabilistic model for the tolerance-based model of social network interaction, we consider that there are two pure states of the social agent (tolerance and intolerance) and at least one intermediary state. State transition occurs at each interaction with an agent that has an opinion; interaction with non-opinionated agents do not produce state transitions. Interaction with a different opinion will generate a transition to the closest more tolerant state of the agent. Also, interaction with the same opinion will cause a transition to the closest less tolerant state of the agent.

in order to apply the Markov analysis, we assume that any agent in a given social network topology, is characterized by a rate λ of encountering the same opinion and a rate μ of encountering a different opinion. Of course, $\lambda + \mu \leq 1$ with $\lambda + \mu = 1$ when there are no agents without opinion. When the sum is less than 1, we have $\lambda + \mu + \rho = 1$ with ρ being the rate of interacting with an agent that has no opinion. At the same time, we assume that λ and μ are exponentially distributed, therefore the probability of encountering the same opinion is $1 - e^{-\lambda t}$ and the probability of interacting with a different opinion is $1 - e^{-\mu t}$ [59].

If an agent encounters an opinion that it already holds, then the level of intolerance is incremented, therefore one of the following transitions occur: $S_0 \rightarrow S_1$ or $S_1 \rightarrow S_2$. If the agent is in S_2, then its intolerance cannot be incremented, even if it encounters the same opinion. The same rationale applies when the agent encounters a different opinion, so that its tolerance level is decremented, $S_2 \rightarrow S_1$, or $S_1 \rightarrow S_0$. Figure 3.1 presents the Markov diagram that corresponds to the entire process of tolerance evolution in social agents [59].

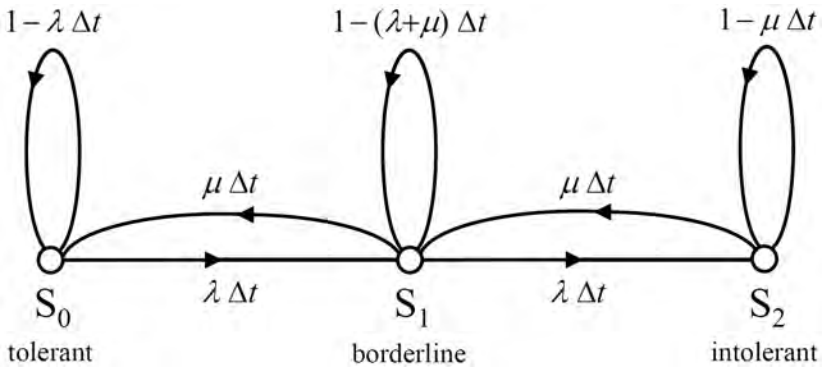

Figure 3.1: The Markov diagram for the 3-state model of tolerance evolution of a social agent. State S_0 represents total (or pure) tolerance, state S_2 represents total intolerance, while S_1 stands for borderline state (i.e., undecided between tolerance and intolerance

From Figure 3.1 we derive the state probability expressions at $t + \Delta t$, assuming that we know the current state at t:

$$\begin{cases} P_{S_0}(t+\Delta t) &= (1-\lambda\Delta t)\,P_{S_0}(t) + \mu\Delta t P_{S_1}(t) \\ P_{S_1}(t+\Delta t) &= \lambda\Delta t P_{S_0}(t) + \mu\Delta t P_{S_2}(t) + \\ & \quad + [1-(\mu+\lambda)\,\Delta t]\,P_{S_1}(t) \\ P_{S_2}(t+\Delta t) &= \lambda\Delta t P_{S_1}(t) + (1-\mu\Delta t)\,P_{S_2}(t) \end{cases} \tag{3.4}$$

Initially, at moment $t = 0$, we consider $P_{S_0}(0) = 1$ and $P_{S_1}(0) = P_{S_2}(0) = 0$. By applying the Laplace transformation, so that variable t is substituted by s, we obtain state expressions:

$$P_{S_0}(s) = \frac{s^2 + (2\mu+\lambda)\,s + \mu^2}{s^3 + 2\,(\mu+\lambda)\,s^2 + (\mu^2+\mu\lambda+\lambda^2)\,s} \tag{3.5}$$

and

$$P_{S_1}(s) = \frac{\lambda\,(s+\mu)}{s^3 + 2\,(\mu+\lambda)\,s^2 + (\mu^2+\mu\lambda+\lambda^2)\,s} \tag{3.6}$$

Therefore, the probability of not getting to the intolerance state, in the 3- state Markov model, is given by:

$$\begin{aligned} P_{tol-3}(s) &= P_{S_0}(s) + P_{S_1}(s) = \\ &= \frac{s^2+2(\mu+\lambda)s+\mu^2+\mu\lambda}{s^3+2(\mu+\lambda)s^2+(\mu^2+\mu\lambda+\lambda^2)s} \end{aligned} \tag{3.7}$$

From 3.7, we get the probability of tolerance state at infinity, which can be interpreted as the expected stable tolerance state of the social agent:

$$\lim_{t\to\infty} P_{tol-3}(t) = \lim_{s\to 0} sP_{tol-3}(s) = \frac{\mu^2+\mu\lambda}{\mu^2+\mu\lambda+\lambda^2} \tag{3.8}$$

If the number of agents without opinion is 0 or negligible, then $\mu+\lambda \simeq 1$ and consequently:

$$\lim_{t\to\infty} P_{tol-3}(t) = \frac{\mu^2+\mu\lambda}{\mu^2+\mu\lambda+\lambda^2} = \frac{\mu\,(\mu+\lambda)}{(\mu+\lambda)^2-\mu\lambda} \simeq \frac{\mu}{1-\mu\lambda} \tag{3.9}$$

Similarly, by repeating the same Markov model for four states ($S_0 - S_3$) we obtain the probability of tolerance state as:

$$\lim_{t\to\infty} P_{tol-4}(t) = \frac{\mu^3+\mu^2\lambda}{\mu^3+\mu^2\lambda+\mu\lambda^2+\lambda^3} = \frac{\mu^2}{1-2(\mu^2\lambda+\mu\lambda^2)} \tag{3.10}$$

The probability of tolerance for $t \to \infty$ (interpreted as corresponding a mature, stable society) can be represented as a function of λ (i.e. the rate of a social agent interacting with another agent with the same opinion). For a convenient graphical representation, ρ is fixed, so that the expression from equation 3.8 becomes function of λ: $P_{tol-3}(\lambda)$, respectively $P_{tol-4}(\lambda)$ as presented in Figure 3.2 [59].

We calculate λ throughout the computer simulations by polling each node in the following manner: starting from the opinion of a node n_i we can say that the ratio between direct neighbors of the same opinion (filtered vicinity N_i^*) and direct neighbors of both opinions (vicinity N_i) is equal to λ, i.e., the rate at which n_i can come in contact with the same opinion. This ratio is obtained as $\lambda_i = |N_i^*|/|N_i|$, for each node, and averaged to $\lambda = 1/n \sum \lambda_i$ for the whole network [59].

For each topology, we notice a variation in time for λ and run simulations until the real-time measured λ and the updated median value is less than 3%, but not more than 50,000

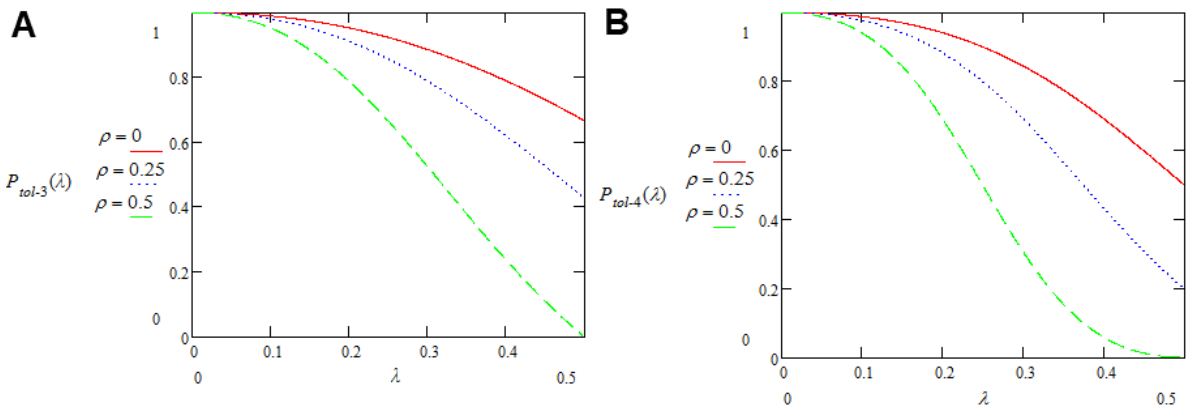

Figure 3.2: Representation for the probability of tolerance (a) $P_{tol-3}(\lambda)$ in the 3-state model, respectively (b) $P_{tol-4}(\lambda)$ in the 4-state model, for three different values of $\rho = \{0, 0.25, 0.5\}$.

iterations. Table 3.1 shows the minimum, maximum and median values registered for λ. These values strengthen the observations presented in [34], namely that:

- Regular and scale-free networks foster intolerance: the nature of the links between nodes is such that closed opinion clusters emerge, and nodes remain in contact only with adjacent (in meshes) or local hubs (in scale-free networks) that have the same opinion. This type of interconnection lowers node tolerance in time and creates a social network less prone to change and less dynamic. This observation is supported by our results, namely $\lambda_{mesh} = 0.82$, respectively $\lambda_{SF} = 0.81$.
- Random and small world networks foster tolerance: links are displayed between nodes such that even though opinion clusters may form, nodes will always have (random) long-range links to other communities with different opinion. This type of interconnection raises node tolerance in time and creates a social network open to change change and more dynamic in terms of opinion. This observation is supported by our results, namely $\lambda_{rand} = 0.62$, respectively $\lambda_{SW} = 0.63$.

Table 3.1: The minimum, median and maximum values of λ (rate of interacting with the same opinion) obtained through simulation on mesh, random, small-world and scale-free networks of 1000 nodes.

	Mesh	Rand	SW	SF
λ_{min}	0.72	0.59	0.57	0.67
λ_{med}	0.82	0.62	0.63	0.81
λ_{max}	0.96	0.65	0.91	0.98

3.2.3 Discussion and Conclusions

The main conclusion of the tolerance model assessment, using probabilistic Markov chain modeling, is that the probability of tolerance is much higher, for the average social agent,

when the number of agents without opinion is small (or ρ is small). For the model with three states, which assumes that the borderline state still represents some sort of tolerance, the probability of tolerance decreases almost linearly with the rate of interacting with the same opinion λ. However, when we consider the more realistic model with four states, we observe that the tolerance probability decreases exponentially with λ.

The dynamics of opinion formation is influenced by topology, which may direct the evolution of opinion towards social balance or intolerance. Overall, the topology has the strongest influence on opinion formation and spread, namely, random and small-world networks foster opinion diversity and social balancing, being representative for a decentralized and democratic society. On the other hand, mesh and scale-free networks act as a conservative, stratified and oligarchic type of society which has a numb reaction to new stimuli [59].

3.3 Prediction of Macro-scale Opinion Distribution

One of the ongoing challenges in computational network science is to better understand and reliably predict diffusion phenomena [7, 9]. Be it under the form of a rumor, a virus, a blog post, or a product, diffusion (or propagation) processes are receiving substantial attention from diverse fields of research, like epidemiology [142, 143], information propagation [144, 145, 39], social networks [9, 34, 146, 25], and even marketing [147, 148] or educational science [49, 149, 50].

The modeling of diffusion processes can be inferred by designing interactions at microscopic level (i.e., between individual social agents), and forecasting network evolution at macroscopic level [150]. Namely, we often try to understand the macroscopic behavior by: (i) monitoring when social agents become *indoctrinated* by their neighborhood (i.e., they adopt information, get infected, buy merchandise [7, 9, 5]), then, (ii) being able to predict how cascades of information flow, and eventually, how the diffusion process is percolated by individuals. Nevertheless, temporal aspects are shown to play an essential role in the diffusion of influence [151]. Many predictive assumptions are made only by observing when an agent gets indoctrinated, and not by considering the variable connectivity in the network, the dynamic trust that builds in ego-networks, or the sources of information [152].

Based on the available studies, we know for sure that social networks have a decisive role in the diffusion of information, and have proven to be very powerful in many situations involving macroscopic behavior [13, 153]. Examples include, but are not limited to, decisively influencing the Arab Spring in 2010 [154], and the U.S. presidential elections in 2008 [155], 2012 [156], and 2016 [157]. The popularity of such online frameworks permits (most) people to spread information in a way that we can consider as new layer of social life [151]. Analyzing the dynamics of this layer can offer substantial predictive power over the real-world social networks they model. Studies on prediction are found in marketing and public relations [9], epidemic spreading in a globalized world [158], hurricane forecasting [159], social media [157], or even using tweets to forecast box-office revenues of movies [160].

3.3.1 A Framework for Opinion Forecasting Based on Time-Aware Polling

Our study focuses on the predictability of election polls. This area of research was originally constructed by employing classic statistical models, applied on opinion polls prior to the election day [161, 162]. Ever since the late '70s, it became a scientifically proposed fact that correct *timing* of the election date can be crucial for the outcome [162]. Bridging over to social networks and media, there is a scientific debate on how the wide coverage of publicized opinion polls in media can affect voters before election [163]. We build upon the premises to extrapolate macroscopic behavior of a society during the pre-election period [60].

In this section we put several pieces of this puzzle together, as we consider the issue of predicting the temporal dynamics of the diffusion process starting off from the assumption that the macroscopic dynamics can be extrapolated solely from microscopic dynamics. Consequently, as we propose a prediction model which encapsulates the topological and behavioral properties of a network, all based on the properties of the microscopic level. As such, we present:

- An analytic methodology for modeling the macroscopic evolution of a multi-opinion system in a social network, targeting better election poll prediction.
- An experimental evaluation of the efficiency of our approach, and the validity of its underlying assumptions.
- Using public survey data from three presidential elections (between 2012-2019), we compare our proposed time-aware (TA) method with the best pollster predictions and two standard statistical approaches (cumulative counting and survey averaging).

First, we define the discrete temporal election axis $t = [0, e)$ as being relative to the date of first pre-election poll (i.e., $p(t = 0)$), and the election day $t = e$. The discrete observations we consider as opinion injection at any time $0 \leq t < e$ stem from all public opinion polls $p(t)$ preceding the election day e. We use as datasets, the following four: the US12 dataset contains 326 individual opinion polls, arching over the period 2010/4/19–2012/11/5; US16 contains 259 polls spanning over the period 2015/5/26–2016/11/8; RO19 contains 21 polls over the period 2019/6/26–2019/11/03; UK16 contains 51 polls over the period 2012/2/21–2016/05/24.

Second, we note that studies applied in epidemiology have proposed three popular parametric models for the likelihood of disease transmission rates, taking *time* into consideration [164, 165]: power-law, exponential, and Rayleigh. These functions model the damping (fading) in time of infectiousness after exposure; nonetheless, they can be used to trace the damping of opinion after each injection in the network.

Naturally, when opinion is injected in a network, we can measure an increase (like a spike) in activity (e.g., number of tweets on a a topic); this increase then slowly dampens, fading back towards a relaxed state (usually $\Omega(t \to \infty) = 0$, no opinion). Overall, the whole process resembles that of the electrical energy stored by a capacitor [60].

Here, we investigate the efficiency of the power-law time-aware (PA) model and the exponential time-aware model (EA). In their original form, these models express the transmission likelihood $\lambda_v(t)$ of a disease in time, after a relative time Δt since an individual $v \in V$ was infected, as expressed by the following two equations:

$$\lambda_v(t) = \alpha_v \cdot \Delta t^{-\beta_v} \qquad (PA) \tag{3.11}$$

$$\lambda_v(t) = \alpha_v \cdot e^{-\Delta t \beta_v} \qquad (EA) \tag{3.12}$$

Specific to the context of opinion contagion, we parameterize the models by adding an amplitude factor α_v and a damping factor β_v, specific for every individual $v \in V$ in the network G. The α_v factor determines the amplitude of the positive response when an opinion is injected in the network, and the β_v factor controls the damping speed towards the relaxed state ($\lambda(t \to \infty) = 0$) for an individual.

We further extrapolate the microscopic interactions models (Equations 3.11-3.12), from individual level, to the level of network G. Consequently, each of the two models (PA and EA) are used to estimate the dynamic weight w_i of each opinion i over the whole network G in time, expressed as:

$$w_i(t) = \alpha_i(t) \cdot t^{-\beta_i} \tag{3.13}$$

$$w_i(t) = \alpha_i(t) \cdot e^{-t\beta_i} \tag{3.14}$$

In both PA and EA models we define two parameters: the time-aware amplitude factor $\alpha_i(t)$, and the constant damping factor β_i. Here, index i represents the opinion i we are modeling in network G. We consider a system with a variable sized set of opinions Ω, where any opinion $i \in \Omega$ may evolve independently from any other opinion $j \in \Omega$ in the same network; As such, we can have different values of α_i and β_i for different evolving opinions [60].

Additionally, in the case of public opinion polls, it might be the case that either a certain percentage of the voters do not follow all the polls, or that polls have an overall higher or lower credibility. For instance, *FiveThirtyEight* currently ranks various polls based on their credibility, with possible effects in the voting population. These variations translate into a variation of opinion poll weight; as such, we introduce a credibility parameter $\xi(t)$ to each poll $p_i(t)$. Since credibility is applied to the entire poll, we have one single $\xi(t)$ parameter applied to each opinion i at time t. In this paper we will assign poll credibility as a uniform vector $\xi(t) = \{1 \mid \forall t\}$.

The amplitude factor evolves according to the following rules: if at moment t there is no opinion injection (i.e., $\nexists$ poll $p_i(t)$), then α_i remains unchanged, so that, as time increases $t \to t+1$, $w_i(t)$ will decrease; if there exists a poll $p_i(t) > 0$ at the current moment, then α_i is increased by an amplitude proportional to the number of votes (or normalized number of votes) times the credibility of that poll $\xi(t)$. An initial value is set $\alpha_i(0) = 1$; next, the evolution of α_i is given by the following equation:

$$\alpha_i(t) = \begin{cases} \alpha_i(t-1)t^{-\beta_i} + \xi(t)p_i^*(t), & \text{if } \exists\, p_i(t) > 0 \\ \alpha_i(t-1), & \text{if } p_i(t) = 0 \end{cases} \tag{3.15}$$

Here, $p_i^*(t)$ represents the *normalized* number of votes expressed in support for opinion i at time t. In order to handle variation in the amplitude of polls, we normalize each value in the range $[0, 100]\%$ of the total expressed opinion at time t. In the validation data we have polls ranging from about 400 to over 40,000 voters, so that a normalization of the amplitudes is imposed [60].

By measuring the evolution of each opinion weight in time, we can calculate the current opinion poll Ω_i by normalizing the weights of each opinion i in G by the sum of all weights $w_i(t)$, as:

$$\Omega_i(t) = w_i(t) / \sum_j w_j(t) \tag{3.16}$$

In addition to our proposed time-aware method, we use cumulative counting (CC) and survey averaging (SA) to serve as basic statistical methods for comparison, as well as multilevel regression and post-stratification (MRP) estimates from the best pollsters. CC is applied by summing up all votes expressed by the polls p_i for each opinion i over the total polling period $[0, e)$. Note that for CC we do not normalize p_i. SA is applied by averaging the current normalized poll results with the previously computed average, over each independent opinion. Here, we can express the opinion poll $s\Omega_i$ directly by using the normalized (*) number of votes for each poll p_i. The main distinction between CC and SA is that the first method uses the absolute number of votes for each opinion, whereas SA uses the same, but normalized values. MRP is a state of the art methodology that anticipates the choice of an individual voter using a statistical model for a sizeable national poll sample [166, 167]. Then, it applies local weights to the model predictions in order to generate forecasts of the result in each district. While we did not employ MRP ourselves, we know that many of the pollsters (e.g., *Real Clear Politics*, *FiveThirtyEight*) rely on (in-house) variations of the MRP method [168].

In terms of simulation results of our time-aware (TA) methods, Figure 3.3, we infer the evolution of poll results, in time, based on the measured weights $w(t)$ for each candidate. Here we focus only the last snapshot of the pre-election period (i.e., last 100 days). The snapshot in Figure 3.3a shows how the CC method prediction is monotonous, as it clearly suggests an advantage for the democratic (blue) candidate in US16 (i.e., Clinton). For CC, the variation in predicted polls converges within large threshold of $\approx 5\%$. Nevertheless, Figure 3.3b shows a much more dynamic prediction system where the advantage of the democratic candidate (blue) ranges between [-0.94–4.90]% over the last 100 days before election. In panels 3.3c,d we display an overview of the evolution of opinion over the full pre-election period. In case of the CC method, it is easy to delimit the advantage between democratic and republican candidates at any time (Figure 3.3c), but switching to the TA method (Figure 3.3d) suggests that there are periods when both candidates had equal chances of winning (e.g., around $t \approx$ 150 and 400).

The performances of all poll prediction estimation methods (PA, EA, CC, SA, best pollster) are displayed in Figure 3.4. We highlight in Figures 3.4a,b the results for US12, alongside the total poll estimations error. Similarly, Figures 3.4c-d, e-f, and g-h highlight the results for US16, RO19, and UK16. For the RO19 dataset we chose to display only the top 4 candidates (out of 8) in order to keep the figure readable.

The first column (gray) in each left hand-side panel represents the ground truth value, i.e., the real election results. With pink we display the two statistical prediction methods, with green the best pollster prediction, and with blue/red our two TA methods. The superior prediction power of the PA and EA methods becomes visible, both visually as well as numerically. For the US12 democratic candidate (BO) we measure offsets of 2.13–2.18% for the TA methods, while CC and SA are offset by $> 4\%$; for the Republican candidate (MR) we measure offsets of only 0.35–0.42% for the TA methods, and CC and SA are offset by $\approx 3\%$.

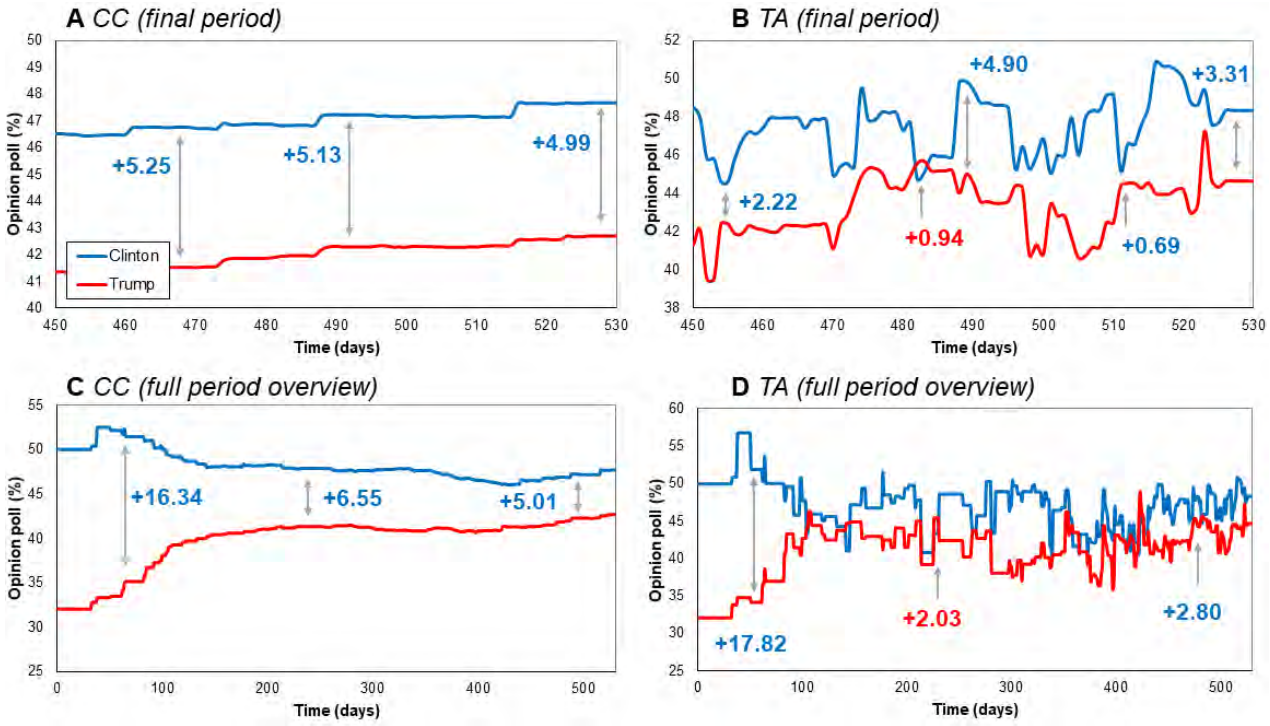

Figure 3.3: Overall polls $\Omega(t)$ evolution, calculated from weights $w(t)$ for the US16 pre-election period. We first provide snapshots of the final period before elections ($t = 450 - 529$) using cumulative counting (**a**), and the time-aware method (**b**) to estimate polls. Here, we exemplify the relative differences (Clinton–Trump) in polls at several time points. (**c**, **d**) We also provide an overview of the whole poll evolution using the same two estimation methods.

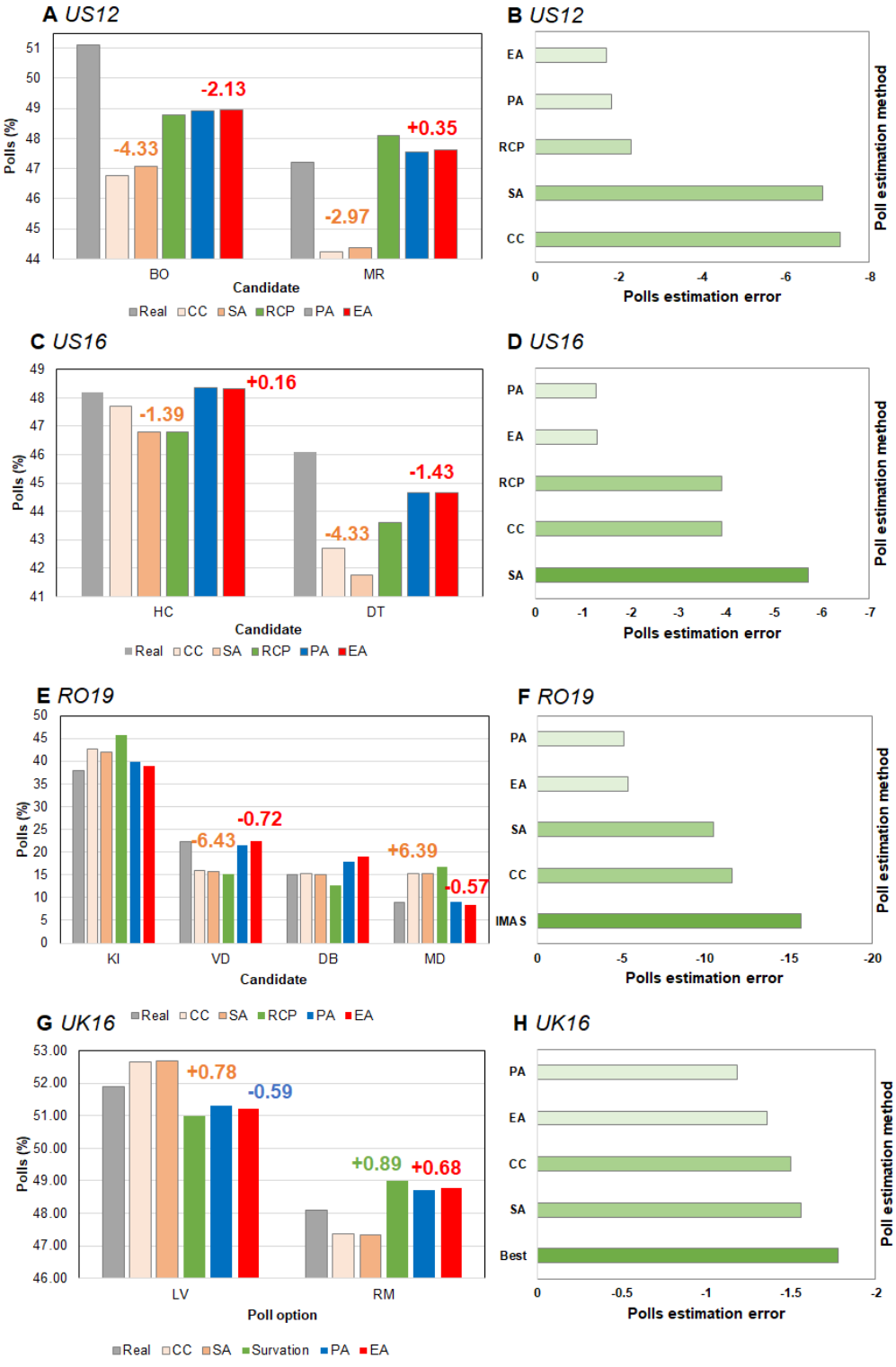

Figure 3.4: Predicted and real election results using the CC, PA, EA and SA methods and best pollsters (RCP, IMAS, Survation). The caption above each column represents the offset from the real election results. The total estimation error (for both candidates) is accumulated and displayed in the right panel (green).

For the US16 Democratic candidate (HC) we measure very small offsets of 0.14–0.16% for the TA methods, while CC and SA are offset by 0.51–1.39%; for the Republican candidate (DT) we measure offsets of 1.43-1.45% for the TA methods, and CC and SA are offset by 3.40–4.33%. For the RO19 top 4 candidates we measure offsets of -0.57–1.14% for the TA methods, while CC and SA are offset by -6.43–6.39%; the best pollster measures offsets between -2.68–2.93. Finally, for the UK16 "remaining" option we measure offsets of 0.59–0.68% for the TA methods, while CC and SA are offset by 0.75– 0.78%; the best pollster is offset by 1.64; for the "leaving" option we measure the same, but negative offsets.

3.3.2 Discussion and Conclusions

We worked towards improving the prediction of the popular vote using our time-aware (TA) forecasting methodology [53, 60]. Indeed, we find studies especially tailored to systems like the US, which are based on the college system [169, 167], and also tailored to systems employing the direct popular vote, like in France [161]. The work of [169, 167] manages to forecast presidential, senatorial, and gubernatorial elections at the state level by incorporating state level demographics to better predict the college vote. Nevertheless, we have developed the TA forecasting model to be usable outside any political context, as long as there is sufficient and reliable pre-election poll data. This choice may give it a theoretical disadvantage in the US system, but as our results show in practice, our model still yields superior performance. Moreover, where other models may need specific tuning to be used in other countries of the world, TA will work without the need for customization [60].

Indeed, our TA model also brings some limitations along, which we further discuss. For instance, we consider social media as an ubiquitous diffusion mechanism, but there are also non social media users. Even in this case, we argue that our model's simplification remains robust, as a study on political attitudes concludes that no statistically significant differences arise between social media users and non-users on political attention, values or political behavior [170]. We also consider that the opinion injected in the social network has a very high media coverage. Recent studies, on how US adults keep informed about political candidates and issues, show that TV (news) occupies the leading spot with 73%, followed by 45% for news websites/apps, 24% for newspapers, and 21% specifically for social media [171]. Another realistic simplification in our model allows us to consider the electoral system relatively hard to shape from outside, so that we do not have to account for data beyond our reach. The liberal democracy index was developed to measure the robustness of a political system, and, according to a study by the Swedish V-Dem institute, the USA scores 0.75 (out of 1) and lies within the top 20% liberal nations; the UK scores 0.80 and lies within the top 10% [172]. As such, we can consider the studied electoral systems as robust.

Despite our simple assumptions, that we can apply a microscopic model to predict macroscopic response, our results pinpoint to the fact that time-awareness is more significant in poll prediction performance than previously considered. For the 2012 US elections, we are able to approximate the final results within a 2% margin, while SA and CC produce offsets of about 7%. Similarly, for the 2016 US elections, our method manages to come within 1.5% of the real election results, while SA and CC stay outside the 4% margin; for the 2019 Romanian elections, our method comes within 6% of the real election results, while SA and CC stay close to the 16% margin; for the 2016 UK Brexit, our method manages to come within 1.2-1.3% of

the real election results, while SA and CC stay outside the 1.5% margin. In terms of quantifying the overall performance boost of our method, compared to the statistical benchmark methods, TA proves to be $\approx 75\%$ more accurate for the 2012 elections, $\approx 72\%$ for the 2016 elections, respectively $\approx 69\%$ for the 2019 elections. We also use non-presidential poll data for validation, and here TA proves to be $\approx 21\%$ more accurate for the Brexit Referendum. Thus, we hope to pave a new path of research targeting dynamic and temporal social network analysis, with immediate applicability in real-world systems where the needs for predictability and control are paramount [60].

Chapter 4

Contributions in Network Medicine

4.1 Background and Motivation

References to the new term "Network Medicine" started to surface as Albert-Laszlo Barabasi popularized the term, in his book entitled *Network Medicine – From Obesity to the Diseasome*, published in 2007, in the prestigious New England Journal of Medicine [21]. A-L. Barabasi suggests that the majority of biological systems can be represented by entities interconnected in complex relationships, similar to social and technological systems, and organized according to the simple network principles.

In light of these claims, complex systems, such as biological ones, can be accurately described by complex network models. Specifically, the networks used in "Network Medicine" use nodes to represent bio-specific entities, such as patients, genes, diseases, phenotypes; the edges results results from respective common risk factors, shared metabolic pathways, physical interactions, shared genes etc [173, 40].

Furthermore, in [21], A-L. Barabasi introduces three layers that need to be modeled in order to properly understand human disease: (i) the metabolic network, (ii) the disease network, and (iii) the social network in which an individual lives. It is believed that the root causes and mechanisms of diseases can be explained, through network medicine, if we model gene regulation networks, metabolic reaction networks, and protein-protein interactions networks with high enough accuracy. For instance, the work of Goh, Barabasi et al. [174] defines a bipartite graph of the connections between genes and diseases. By projecting the diseases node set, we obtain the human disease network, which represent diseases sharing common genes (interactions). Using these insights, we can detect larger communities (classification) of diseases, which are then analyzed based on the genetic relationships between nodes. As such, network medicine is a modern, cutting edge tool for analyzing biomedical Big Data [175].

This section summarizes our contributions in Network Medicine, divided in two distinct classes. First, we are collaborating with the "Victor Babes" University of Medicine and Pharmacy Timisoara (UMFT) since 2013, on a cross-disciplinary research path to improve the diagnosis and severity prediction of Obstructive Sleep Apnea (OSA). With 3 research projects (including an ongoing Horizon 2020 project), 4 Q1 journal publications [41, 42, 43, 44], and over 10 medical congress abstracts, we have established a powerful Network Medicine team with members from UPT and UMFT. Second, we have created collaborations with the Faculty of Pharmacy (UMFT) and University of Southern California (USC) for employing drug-drug

interactions and repurposing analysis using network medicine. Also, with 1 research project and 2 Q1 journal publications [46, 47], we are forming a team of experts in the field of pharmacology, by employing network medicine as our analysis tool.

4.2 Diagnosing Obstructive Sleep Apnea using a Network-Based Approach

Network medicine has received a lot of attention during the last decade [176, 174, 21, 63]; this trend if fueled by the fact that complex network science can bring significant advances in various medical fields like genomics [177, 178], drug-target interaction [179], or cell metabolism [180, 181]. Consequently, it has been recently suggested that network medicine can be also used for addressing important problems in respiratory medicine [182, 183].

Obstructive sleep apnea (OSA) is a serious sleep respiratory disorder, which has a prevalence that is considered by many authors as epidemic [184, 185, 186, 187, 188, 189]. OSA consists of abnormal breathing pauses that occur during sleep, resulting in sleep fragmentation and excessive daytime somnolence [190, 191]; it is considered as part of the wider category named SDB (sleep-disordered breathing). In general, SDB produces an impaired quality of life, including an increased risk of causing motor-vehicle accidents. SDB also increases the mortality rate [192], because it contributes to the development of cardiovascular diseases [193] such as hypertension [194], type 2 diabetes [195], cancer [196], and chronic kidney disease [197]. Because it is associated with many co-morbidities [198], SDB has several distinct clinical phenotypes. If not properly diagnosed and treated, SDB may increase morbidity and preoperative risks as well [199, 200, 201, 202, 203].

OSA severity is quantified with the Apnea-Hypopnea Index (*AHI*). Apneas are defined as a decrease of at least 90% of airflow from baseline, which lasts for ≥ 10 seconds, whereas hypopneas are defined as a $\geq 30\%$ decrease of airflow that lasts ≥ 10 seconds; both are associated with either an arousal or a $\geq 3\%$ O_2 saturation decrease [204]. The AHI represents the mean number of apneas and hypopnoeas per hour of sleep. Clinically significant OSA is characterized by AHI ≥ 30. However, some studies are adopting different AHI thresholds for OSA, such as 15 (considered as the lower limit for moderate risk) or 20 [205]. Nonetheless, the clinical relevance and consequences of mild obstructive sleep apnoea is still unclear [206]. Also, there is a variability in scoring the respiratory events across different countries [207]. In current practice, there are four major predictive models based on questionnaires, namely Berlin, STOP, STOP-BANG, and NoSAS [208, 209, 210, 211, 212, 205]. Published studies indicate STOP-BANG as the best available predictive score, due to its high sensitivity: 83.6% for $AHI > 5$, 92.9% for $AHI > 15$, and 100% for $AHI > 30$. However, STOP-BANG has a low specificity (56.4% for $AHI > 5$, 43% for $AHI > 15$, and 37% for $AHI > 30$) [211, 212, 213] which prevents the usage of this score for population screening. NoSAS score comes to improve the prediction specificity by a considerable margin (69%), while maintaining a sufficient sensitivity value (79%). Although there are notable attempts for improving scores' specificity [213], they are mainly targeting narrow-type cohorts such as perioperative patients.

Consequently, our research path is aimed at analyzing the general case, with all patient categories being taken into account for screening, and not just some specific cohorts. To this end, our research is underpinned by a complex network perspective on uncovering OSA

phenotypes. Indeed, network science is already successfully used in medicine at disease-level [40], including respiratory applications [182, 183]. Our network-based approach on OSA risk factors allows for better, more accurate OSA phenotype identification, which in turn leads to a new predictive score (SAS_{Score}). In comparison with the state-of-the-art, our OSA risk prediction score achieves significantly better specificity in predicting actual AHI categories, which makes our SAS_{Score} very appropriate for screening big populations as part of preventive medicine programs.

4.2.1 From Patient Cohort to Network Model

In order to use network medicine for OSA research, we need real-world OSA patient datasets. Unfortunately, OSA patients datasets are scarce and not public; such a situation is justified by multiple aspects: big data techniques were only recently considered as tools for respiratory medicine and OSA, all patients must undergo hospital polysomnography (which entails a complex, expensive and time-consuming process), while coordinated research efforts for gathering data were only recently introduced. For instance, the biggest such OSA database, namely European Sleep Apnea DAtabase - ESADA [214], is not public and it gathers data from 15,956 patients in 24 sleep centers from 16 countries, since 2007. Also, a recent OSA study [205] where the validation is similar to our approach, uses only one (private) validation database, comprising 1101 patients [215].

Given the current context, we build our own Apnea Patients Database (APD), consisting of consecutive patients with suspicion of sleep breathing disorders, which were evaluated at "Victor Babes" Regional Hospital from Timisoara (Western Romania) starting from March 2005, to the present day, under the supervision of the hospital's Ethics Committee (internal briefing note no. 10/12.10.2013). Each patient had respiratory polygraphy performed using both Philips Respironics' Stardust polygraph (2005) and MAP's POLY-MESAM IV (1998). PSG was carried out with Philips Respironics' Alice 5 Diagnostic Sleep System, according to the appropriate guidelines [216]. The polygraphy was performed both at home and at the hospital, whereas PSG measurements were performed at the hospital under medical supervision. To preserve the information accuracy, all collected data were carefully verified; throughout this process, we have ensured complete data confidentiality. Our observational, retrospective study employs only procedures that are standardized and non-invasive, by excluding all useless investigations. Moreover, visits did not entail additional effort for the patients or supplemental budget for the clinic. [41]

Our first significant study [41] uses a cohort of $N = 1371$ patients with completed sleep study protocol and signed informed consent are included in the APD, each with corresponding 108 breathing parameters and anthropometric measurements. In order to verify if there is any difference between apnea and non-apnea populations in terms of how risk factors associate and converge, we built a 611 people non-OSA database NAD (using the same procedure as for the APD). Also, to evaluate the prediction score derived from our study, we gathered a distinct test database TD (fall of 2013) consisting of 231 patients, by following the same procedure. Figure 4.1 presents the distinct roles of our 3 databases, as well as the relationship between them.

Next, we build the unweighted Apnea Patients Network (APN), by assigning vertices and edges: each node corresponds to a distinct patient in our OSA patients database APD,

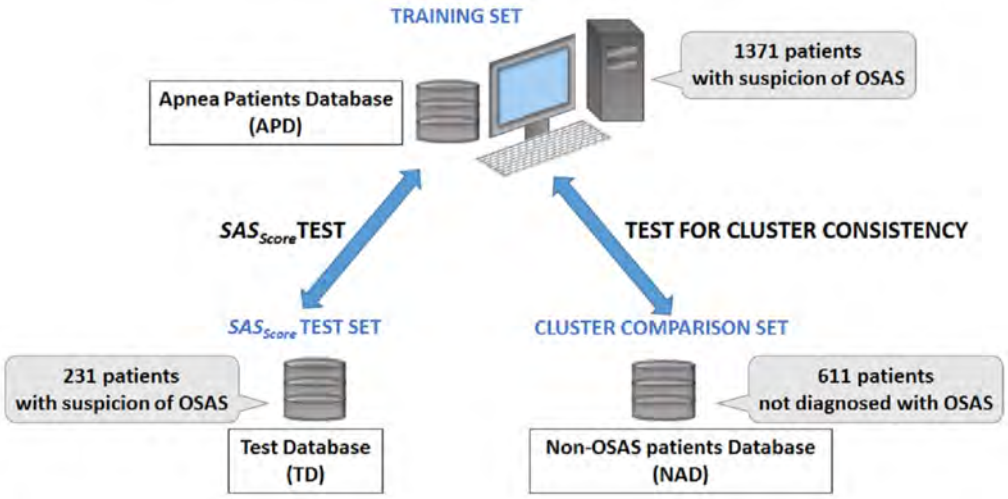

Figure 4.1: The main Apnea Patients Database (APD), comprising 1371 consecutive patients, is is used to build patient phenotypes and to render the SAS_{Score}. The distinct Test Database (TD), comprising 231 consecutive patients, is used to verify the sensitivity and specificity of predicting patient's AHI and OSA categories. The Non-OSA patients Database (NAD) uses consecutive assessed people which are not diagnosed with OSA in order to test for cluster consistency (i.e. compare how risk factors converge in clusters for OSA patients in comparison with people without OSAS).

while an edge (link) is created between two vertices if there is a risk factor compatibility between the patients represented by the two vertices (nodes). The risk factor compatibility is a binary function $f_{RFC} \in \{0, 1\}$ (0 means incompatibility and 1 means compatibility) based on six parameters with high relevance for OSAS: age, gender, BMI, neck circumference, blood pressure (systolic and diastolic), and Epworth Sleepiness Score. We build our APN by considering that $f_{RFC} = 1$ if at least 4 out of 6 parameters are identical; otherwise $f_{RFC} = 0$.

The six parameters are selected from the pool of all relevant risk factors, because they can be measured easily and objectively; such objective measurements can be performed anywhere, and are widely accepted in the medical literature [191]. In contrast, other scores consider snoring and witnessed apnea episodes as factors, but these are parameters which cannot be observed or measured objectively [41].

The reason for adopting the 4-out-of-6 criterion is that it assures the right amount of link density in the APN, meaning that there are enough links so that the APN is connected, but not too many links so that communities (i.e. clusters) can be rendered with energy model layouts [69]. Figure 4.2 shows that the 4 out of 6 link filtering represents the best alternative, we use this criterion to build the APN. To the best of our knowledge, this link filtering procedure is original and has not been used before in such network-based approaches [41].

The APN is clustered using our dual clustering methodology, i.e., a complementary use of energy-model layouts and modularity based partitioning. We have adopted similar approaches in [46, 43, 41, 44, 58, 57]. To this end, use the Force Atlas 2 algorithm [74] as network layout; this layout is very effective in clustering various types of complex networks, as it is based on previous theoretical foundation of force directed attraction-repulsion algorithms [217].

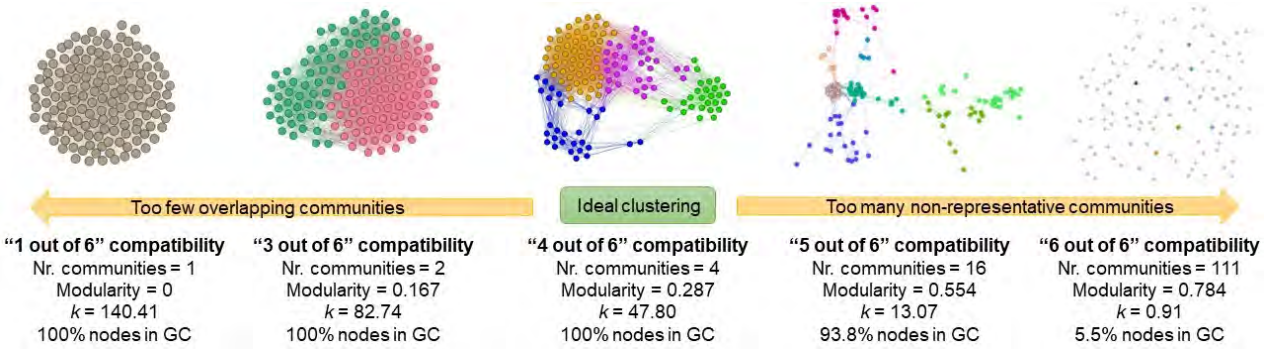

Figure 4.2: APN edge filtering, by considering different definitions for $f_{RFC} = 1$, when we adopt the x-out-of-6 criteria ($x = 1$, 2, 3, 4, 5, and 6). The visual result indicate $x = 4$ as the best solution, because the edge density is convenient for rendering topological clusters with energy model layouts. If a lower threshold is used (i.e., less strict) too few, dense and overlapping communities emerge. Conversely, if a higher threshold is used (i.e. more strict) to many, non-representative communities emerge and many nodes become completely disconnected from the giant component (GC) of the network.

4.2.2 Patient Phenotype Definition based on Our Dual Clustering Technique

The APN and NPN representation resulted from our clustering methodology is presented in Figure 4.3, where the distinct colors correspond to distinct modularity classes, and the well-defined topological clusters are explained accordingly. In Figure 4.3a, we interpret the 8 topological clusters as distinct phenotypes, and provide the risk factors prevalence as percentages (L, Mi, Mo, Se)% for each such cluster/phenotype. Upon visual inspection, Figure 4.3b suggests that in the non-OSA control population there are more patterns of risk factors association, which leads to a number of 12 topological clusters and modularity classes that are not correlated with OSA or AHI risk groups. As such, according to our network-based methodology, it occurs that the 6 considered risk factors consistently converge only for the individuals with OSAS.

classifying new patients in one of the phenotypes can be performed by adding the new patient to the APN and then running the modularity class and force-directed layout algorithms in Gephi [75] again. However, in clinical practice, physicians are often unable to perform these rather complex and time consuming computational steps. In order to deal with this problem, we propose a simplified solution for classifying *de novo* patients, using a computer algorithm that is implemented as a web-based/mobile application on Google Play: Morpheus: Sleep Apnea Syndrome [1]. Our mobile application employs a simplified algorithm for classifying new patients in one of the 8 phenotypes.

[1]URL: `https://play.google.com/store/apps/details?id=aerscore.topindustries.aerscore`

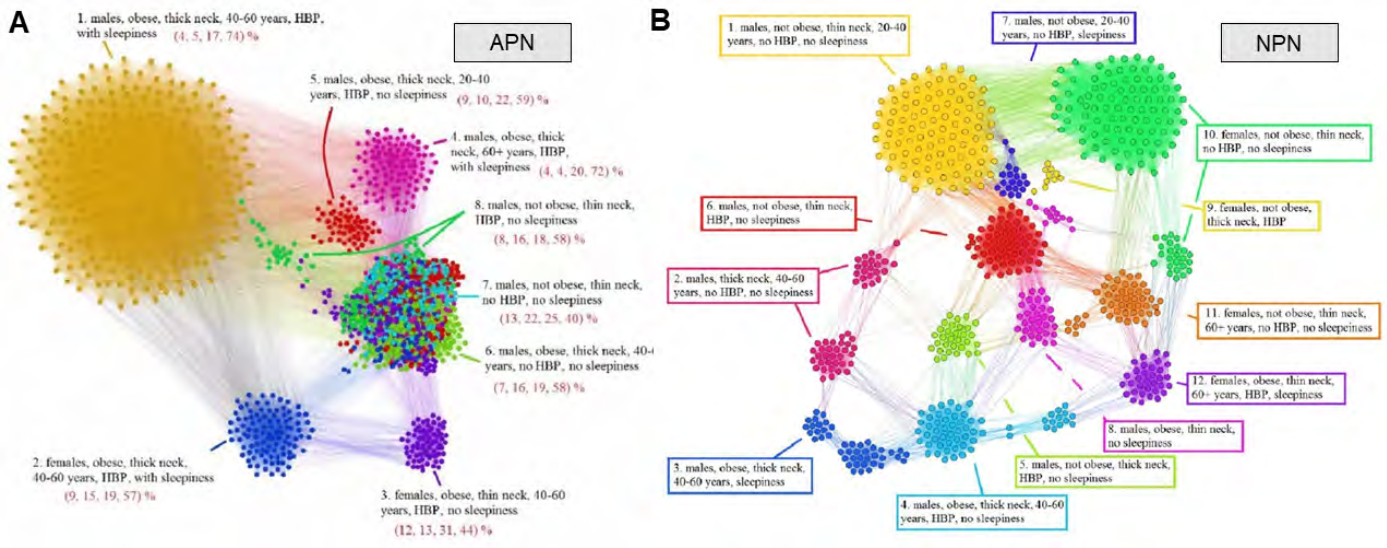

Figure 4.3: (a) Apnea Patients Network and (b) Non-Apnea Patients Network obtained with data from the APD, and NPD, according to the risk factor compatibility relationship, using our dual network clustering methodology. The assigned colors correspond to modularity classes, and the topological clusters are indicated. (a) For each topological cluster, statistics are provided in red (as percentages) for all AHI risk groups.

4.2.3 Gender-Based Differences in OSA Phenotyping

Our subsequent study [43] proposes the distinction between two networks, one for each patient gender. As such, we build the male patient network (MPN) and female patient network (FPN) based on our WestRo dataset with $N = 2796$ patients. The mapping of risk classes over the MPN and FPN is depicted in Figure 4.4. By visualizing the mappings, we can easily identify phenotypes' categorization into healthy (green) or sick (red).

We note that our dual clustering method renders relatively distinct phenotypes for both genders, meaning that different risk factors do associate in specific patterns. From a medical standpoint, this observation is consistent with several state-of-the-art studies that hold gender as an essential predictor of OSA severity [218, 219]. For each gender, we obtain 8 phenotypes that can be differentiated by the association of four main risk classes (obesity OB, thick neck TN, hypertension HT, and daytime sleepiness SL), as depicted in Figure 4.5a. Consequently, we uncover insightful patterns of OSA development for each gender through the differential comparison provided in Figure 4.5b. We find several identical phenotypes, e.g., the male $Ph5$ is the same as the female $Ph3$ (patients with SL). Conversely, we find multiple phenotypes that do not have a correspondence in the other gender; these are phenotypes $Ph2$, $Ph6$, $Ph7$, $Ph8$ for males and $Ph2$, $Ph7$, $Ph8$ for females, respectively.

By analyzing all enumerated male-only specific phenotypes, we find that all 4 of them have TN as a common risk class; furthermore, 3 out of 4 include SL, one includes HT, and another OB. This leads to the conclusion that TN is a major OSA predictor for male patients, and is associated—in order—with SL, and HT or OB. By analyzing all enumerated female-only specific phenotypes, we find that all 3 of them have HT as a common risk class. Moreover, 2 out of 3 include OB, and one includes SL; this leads us to the conclusion that HT is a major

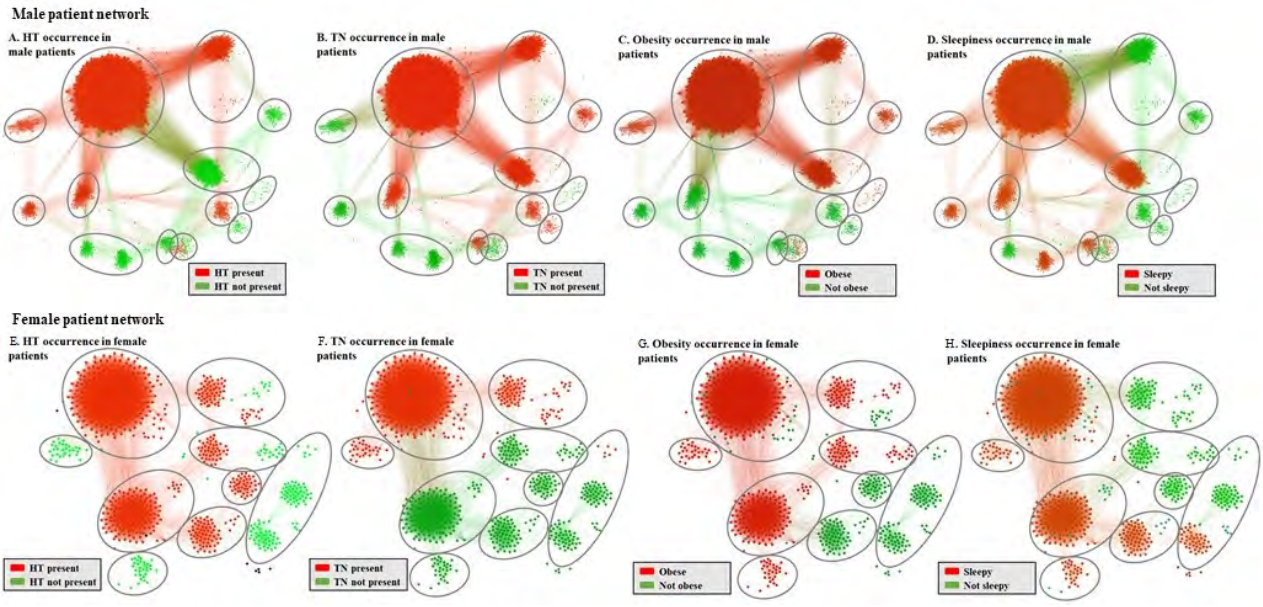

Figure 4.4: Mapping of OSA risk classes on the MPN (a-d) and FPN (e-h), where ellipses mark the eight identified phenotypes, for identifying: hypertension (HT), thick neck (TN), obesity, and daytime sleepiness.

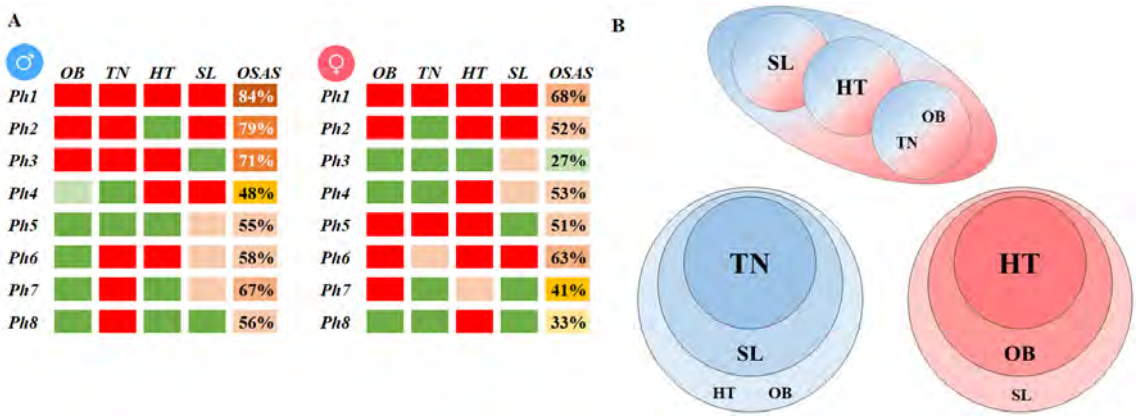

Figure 4.5: Gender phenotypes comparison. (a) Male and female phenotypes described by the four main associated risk classes (red – sicker, green – healthier). (b) Association of risk classes for the two genders (males – blue, females – red); the upper panel represents overlapping phenotypes, while the lower panel represents gender-specific phenotypes.

OSA predictor for female patients, and is associated—in order—with *OB*, and *SL*.

In conclusion, we applied a novel clustering method based on network medicine, which results in new gender-specific OSA phenotypes [43]. This innovative approach—based on assessing five objective patient parameters—results in identifying eight unique phenotypes for each gender. Some of the detected clusters match for both genders (e.g., the severe or mild OSA phenotypes), while others present a unique pattern of OSA risk factor association that is specific for each gender. As such, for males, we find that *large NC – sleepiness – hyper-*

tension/obesity represents a typical association pattern; for women, we find the association pattern *hypertension – obesity – sleepiness*. We believe that our work will stimulate future research in sleep medicine, help OSA prediction, and foster a personalized patient management process.

4.2.4 Developing a Tool for Population-Wide monitoring of OSA

Present practice shows that the existing screening tools (e.g., STOP-Bang, NoSAS [205]) have limited effectiveness when monitoring large populations (e.g. groups of more than $100,000$ people). In other words, current scores mainly focus on simplicity and high sensitivity, because these characteristics are paramount for clinical problems such as a rapid diagnosis of preoperative patients – the unfortunate consequence is a high rate of false positives.

Therefore, our follow-up study [42] aims at defining a score for OSA severity targeting high specificity. At the same time, to address the needs of practitioners in sleep laboratories, we simplify the computation of the score, so that it may easily be applied in daily scenarios. Altogether our contributions can be summarized as follows:

1. We redefine our computer-based algorithm that calculates SAS_{score} in a form that can also be used by practitioners in a much simpler way, without the need to employ dedicated *in silico* tools. To this end, we only marginally reduce the accuracy of the original SAS_{score}, while significantly increasing its usability.

2. We validate the simplified SAS_{score} on a cohort of $N = 2595$ patients diagnosed in several sleep centers from Western Romania.

3. We optimize the performance of our SAS_{score}, to maximize its specificity (using area-under-curve analysis AUC).

4. We compare SAS_{score} with state of the art monitoring tools (i.e. STOP-Bang, NoSAS) in terms of sensitivity, specificity, AUC, to conclude that SAS_{score} is indeed better suited for monitoring large populations.

Originally, SAS_{score} was created in such a way that, for every new patient, computer-based algorithmic processing is required to insert the patient into our curated apnea patient network [41]. Then, the patient is automatically assigned to one of the 8 graph clusters (phenotypes); after performing this assignment, the patient's SAS_{score} is computed with the following equation:

$$SAS_{score} = \frac{BMI}{BMI_{cluster}} + \frac{NC}{NC_{cluster}} + \frac{SysBP}{SysBP_{cluster}} + \frac{ESS}{ESS_{cluster}} \tag{4.1}$$

In Eq. 4.1 the index of the assigned cluster is $cluster \in \{1..8\}$. Each cluster has a set of precomputed average measures for body-mass index (BMI), neck circumference (NC), systolic blood pressure (SysBP), and Epworth Sleepiness Scale (ESS) [220]. Thus, the new patient's anthropometric parameters are normalized towards the cluster's average values, so that his/her SAS_{score} represents a relative risk as compared to the cluster average. Such an approach is owing to the normal/Gaussian distribution that was identified in all relevant parameters and anthropometrics [41].

However, the computational steps entailed by calculating the original SAS_{score} require specialized, computer-based software tools. Therefore, while maintaining our initial focus on building a high specificity and sensitivity OSAS monitoring tool, we simplify Eq. 4.1 according to the following principles: (i) all relevant patient parameters are considered, (ii) instead of performing the dual clustering technique, we use fixed average values for the considered parameters. In Eq. 4.2 the fixed average values for BMI, NC, $SysBP$, and ESS are standard values that can be found in literature and that are used in clinical practice.

$$SAS_{score} = \begin{cases} \frac{BMI}{30} + \frac{NC}{40} + \frac{SysBP}{140} + \frac{ESS}{11} & \text{, for women} \\ \frac{BMI}{30} + \frac{NC}{43} + \frac{SysBP}{140} + \frac{ESS}{11} & \text{, for men} \end{cases} \tag{4.2}$$

The resulted score is a rational number with no strict lower or upper bound. Nevertheless, due to specific limits of anthropometric and physiological measures, we found that scores mainly range within the [2, 7] interval. Because the score is consistently proportional with the patient's AHI, we also provide a direct risk classification which corresponds to the AHI-based risk groups:

$$SAS_{Risk} = \begin{cases} Low & \text{if } SAS_{score} < 3 \\ Mild & \text{if } 3 \leq SAS_{score} < 3.5 \\ Moderate & \text{if } 3.5 \leq SAS_{score} < 4 \\ High & \text{if } 4 \leq SAS_{score} < 5 \\ Very\ high & \text{if } SAS_{score} \geq 5 \end{cases} \tag{4.3}$$

The performance results of our score are presented in Table 4.1. The prevalence of OSAS in the cohort, as can be defined by adopting different AHI cut-off values are as follows: 2519 (97.1%) for $AHI \geq 5$, 2390 (92.1%) for $AHI \geq 10$, 2238 (86.2%) for $AHI \geq 15$, 2033 (78.3%) for $AHI \geq 20$, 1671 (64.4%) for $AHI \geq 30$, and 1093 (42.1%) for $AHI \geq 45$. Table 4.1 provides the performance comparisons for the $AHI = 30$ cut-off.

Table 4.1: Performance of STOP-Bang, NoSAS, and SAS_{score} in the WestRo cohort ($N = 2595$) when $AHI \geq 30$ events/h is considered the diagnosis criteria.

	Prevalence	**AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
STOP-Bang	2404 (92.6%)	0.69 (0.66-0.73)	0.968	0.149	0.673	0.723
NoSAS	2157 (83.1%)	0.66 (0.63-0.68)	0.901	0.294	0.698	0.621
SAS_{score}	1977 (76.2%)	0.73 (0.71-0.75)	0.829	0.359	0.701	0.537

The data within parentheses (from the 'AUC' column) represent 95% confidence intervals. AUC = area under the curve. PPV/NPV = positive/negative predictive value.

Overall, we notice that the prevalence according to the SAS_{score} (76.2%) is the closest to the real one (64.4%) – as obtained after rigorous polysomnography – and the AUC has the highest value (0.73) for SAS_{score}. In terms of sensitivity, SAS_{score} performs marginally weaker (0.829), yet it offers the best specificity among the three scores (0.359). These results mean that SAS_{score} obtains a specificity that is 140.9% higher than that of STOP-Bang.

4.2.5 Conclusions

Our results show that, using patient measurements that are easily available in primary care practice, the customizable SAS_{score} allows for reliable determination of clinically significant OSA, with a high and adjustable specificity, ranging from 0.359 to 0.607. Compared with existing state of the art screening scores, such as STOP-Bang (0.149 specificity) and NoSAS (0.294 specificity), SAS_{score} is indeed the most appropriate for monitoring large populations.

In conclusion, as suggested by the higher AUC and correct classification proportion (with respect to the other scores), our SAS_{score} has the potential of representing a better compromise between sensitivity and specificity, allowing clinically significant SDB to be reliably ruled out, without yielding too many unnecessary sleep investigations.

Our score can be a useful tool for OSAS/SDB screening in large population categories such as professional drivers, because, from January 2016, the new 2014/85/EU directive [221] targeting professional drivers is recommended across the entire European Union (Commission Directive 2014/85/EU of 1 July 2014 amending Directive 2006/126/EC – European Parliament and the Council on driving licenses). We are confident that our line of network medicine research [41, 42, 43, 44], with direct applicability in sleep research, represents a timely advancement in the field of OSAS monitoring and severity prediction.

4.3 Predicting Drug Interactions and Repurposing using Network Pharmacology

Conventional drug design has become expensive and cumbersome, as it requires large amounts of resources and faces serious challenges [222, 223]. Consequently, although the number of new FDA drug applications has significantly increased during the last decade, the number of approved drugs has only marginally grown [224, 225], calling for more robust alternative strategies [226].

One of the most effective alternative strategies is *drug repurposing* (or *drug repositioning*) [227, 228], namely finding new pharmaceutical functions for already used drugs. The extensive medical and pharmaceutical experience reveals a surprising propensity towards multiple indications for many drugs [229], and the examples of successful drug repositioning are steadily accumulating. Out of the 90 newly approved drugs in 2016 (a 10% decrease from 2015), 25% are repositionings in terms of formulations, combinations, and indications [225]. Furthermore, drug repositioning reduces the incurred research and development time and costs, as well as medication risks, which makes it particularly efficient for developing orphan/rare disease therapies [229, 230].

The recent developments confirm computational methods as powerful tools for drug repositioning [47]:

- The wide availability of omics (e.g., genomics, transcriptomics, proteomics, metabolomics) analytical approaches have generated significant volumes of useful Big Data [231, 232].
- Ubiquitous digital devices and social media, has tremendously expanded the amplitude of the process of gathering data on drug-drug interactions and drug side-effects [233, 234].

- The recent developments in physics, computer science, and computer engineering have created efficient methods and technologies for data exploration and mining, such as complex network analysis, machine learning, or deep learning [235, 232, 179, 236, 237, 238].

We developed a novel, network-based, computational approach to drug repositioning. To this end, we build a weighted *drug-drug network*, i.e., a complex network where the nodes are drugs, and the weighted links represent relationships between drugs, using information from the accurate DrugBank database [239]. In our drug-drug similarity network (DDSN), a link is placed between two drugs if their interaction with at least one target is of the same type (either agonistic/ activator or antagonistic/ inhibitor). The link weight represents the number of biological targets that interact in the same way with the two drugs. A target $t_k \in T$ (T is the set of targets) on which drug nodes v_i and v_j act in the same way, either both agonistically or both antagonistically. Within this framework, we build the DDSN graph G using drug-target interaction information from Drug Bank 4.2 [239]. We base our analysis on the largest connected component of the DDSN, consisting of $|V| = 1008$ drugs/nodes and $|E| = 17963$ links resulted from the analysis of the drug-target interactions with $|T| = 516$ targets [47].

To gain insights from the DDSN topological complexity, we identified specific drug clusters (or communities) using our dual clustering technique based on modularity-based graph partitioning [71], and the Force Atlas 2 layout algorithm [74]. In the case of DDSN, the clusters correspond to drug communities C_x, $x \in \mathbb{N}^*$, such that $V = \bigcup_{i=1}^{m} C_x$. Using the constructed DDSN from Drug Bank 4.2 and expert analysis, we label each cluster according to its dominant property (i.e., the property that better describes the majority of drugs in the cluster), which may represent a specific mechanism of pharmacologic action, a specifically targeted disease, or a targeted organ. Figure 4.6 illustrates the resulting DDSN, where the node colors identify the distinct modularity clusters [47]. We assess the ability of our method to uncover new repositionings by confronting our results with the latest (version 5.1.4) Drug Bank and with data compiled from interrogating scientific literature databases.

4.3.1 Network Centralities as Hints for Drug Repurposing

In our characterization of drug-drug similarity networks, a high degree node represents a drug with already documented multiple properties. Also, a high betweenness (i.e., the ability to connect network communities) indicates the drug's propensity for multiple pharmacological functions. By this logic, the high-betweenness, high-degree nodes may have reached their full repositioning potential, whereas the high betweenness, low degree nodes (characterized by high betweenness/degree value (b/d or simply *bpd*) may indicate a significant repositioning potential.

To explore the capability of *bpd* to predict the multiple drug properties, we exploit the community structure of DDSN by following a two-step approach.

First, we assign a dominant property to each community using expert analysis. Figure 4.6 illustrates the 26 DDSN communities as well as their dominant functionality. The dominant community property can be a pharmacological mechanism, a targeted disease, or a targeted organ. For instance, community C_1 consists of antineoplastic drugs which act as mitotic

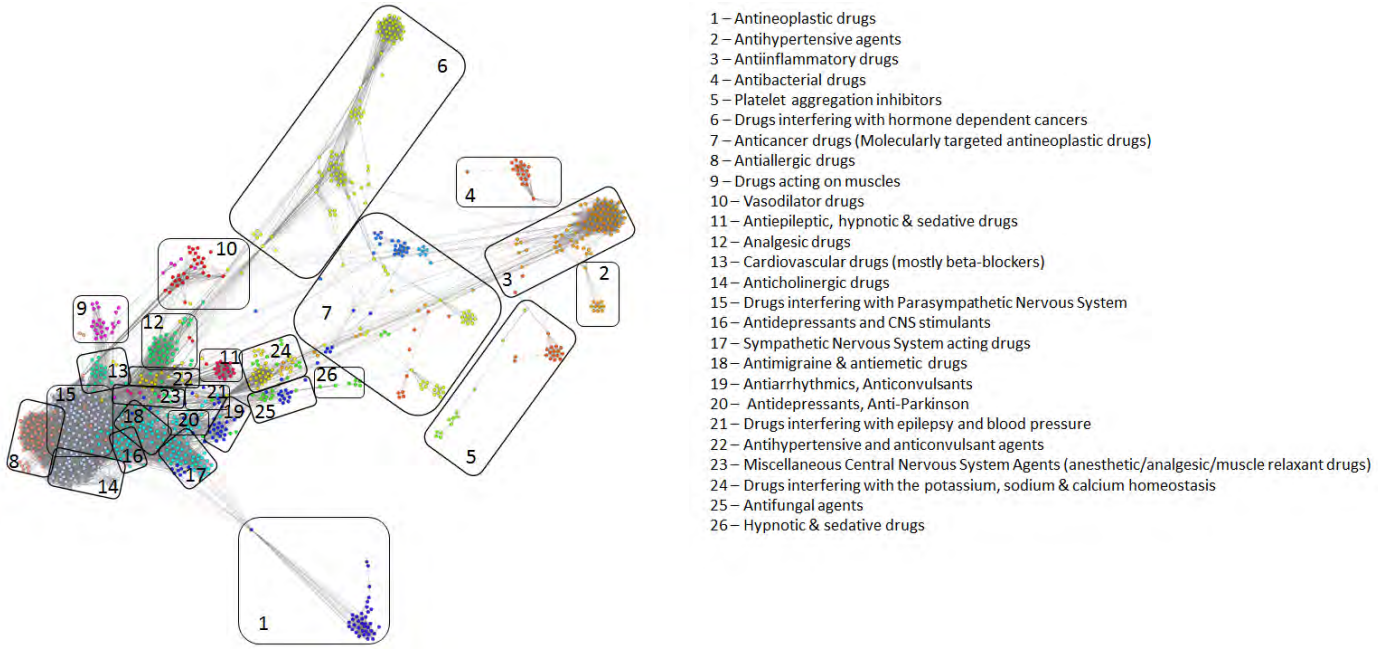

Figure 4.6: The drug-drug similarity network, where nodes represent drugs and links represent drug-drug similarity relationships based on drug-target interaction behavior. We identify 26 topological clusters with rounded rectangles and provide the functional descriptions for each of them.

inhibitors and DNA damaging agents; C_{13} consists of cardiovascular drugs, mostly beta-blockers. Second, in each cluster C_i, we identify the top t drugs according to their *bpd* values. From these selected drugs, $B_i^t \subset C_i$, some stand out by not sharing the community property or properties, and thus, can be repositioned as such. To this end, for $i = \overline{1,m}$ eliminated from B_i^t the drugs whose repurposings were already confirmed (i.e., performed by others and found in the recent literature), thus producing $m = 26$ lists of repurposing hints yet to be confirmed by *in silico*, *in vitro*, and *in vivo* experiments, $B_i^h = B_i^t \setminus B_i^c$.

The community Id depicted in Figure 4.7 identifies each top *bpd* node, excepting Meprobamate (in community C_{25}) and Acarbose (C_19), because these drugs do not seem to possess their community's main property; this indicates Meprobamate as antifungal (i.e., the property of community 25) and Acarbose as antiarrhythmic, anticonvulsant (i.e., the properties of community 19). As such, our clustering results indicate two top *bpd* drug repositionings, i.e., both repositionings refer to properties currently unaccounted in the DrugBank version 5.1.4 and the scientific literature we have screened. Meprobamate is a member of the (green) community of psychotropic drugs but is also well connected to the (dark blue) community of antifungal drugs. The placement of Meprobamate and the high *bpd* value suggest that it may also have an antifungal effect. A a last validation procedure, we use *molecular docking*, to further confirm the repurposing of Meprobamate.

Molecular docking represents an alternative, in silico simulation approach to drug discovery, which models the physical interaction between a drug molecule and a target (or a set of targets). With molecular docking, we estimate the free energy values of the molecular interactions to offer a good approximation for the conformation and orientation of the ligand into

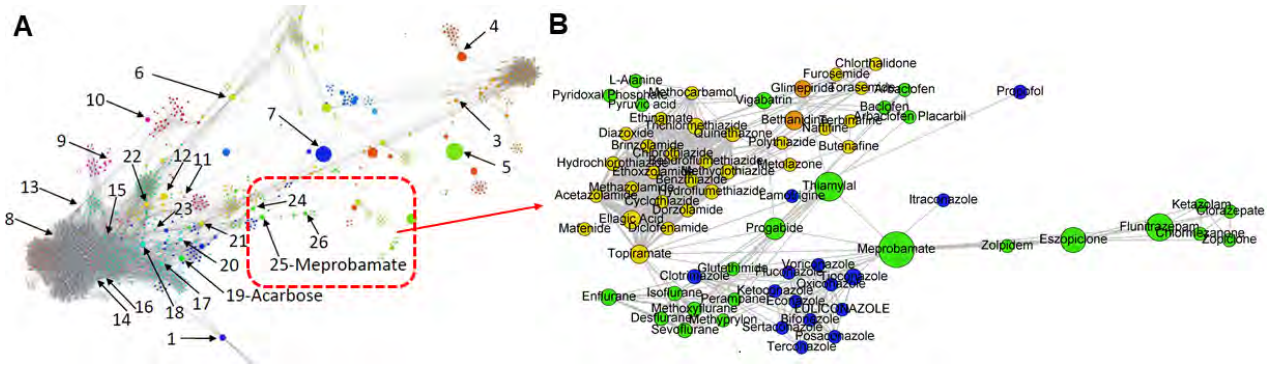

Figure 4.7: (a) The DDSN where node sizes represent their *bpd* values. The arrows indicate the top *bpd* node in each community. (b) Detail showing Meprobamate as the node with the biggest *bpd* in its community.

the protein cavity [240]. Along with many available molecular docking models, DOCK [241] is a dedicated software tool used in drug repurposing. Meprobamate is a known oral drug. However, when considering its potential antifungal activity, we cannot exclude the topical route of administration. To this end, we suggest that further investigations on biopharmaceutical properties (e.g., solubility, lipophilicity, octanol/water partition coefficient) are required [47].

4.3.2 Discussion

Our research in drug-drug interactions and repurposing using network medicine has led us to the definition of drug similarity based on drug-target interactions [46, 47]. As such, we built weighted Drug-Drug Similarity Networks (DDSN) according to the drug-drug similarity relationships. Using our dual clustering technique, we generate drug communities that are associated with specific, dominant drug properties. However, we find that 13.59% of the drugs in these communities seem not to match the dominant pharmacologic property. Thus, we consider them as drug repurposing hints. The resources required to test all these repurposing hints are considerable. Therefore we introduce a mechanism of prioritization based on the betweenness/degree *bpd* node centrality. By using *bpd* as an indicator of drug repurposing potential, we identify the drug Meprobamate as a possible antifungal. Finally, we use a robust test procedure, based on molecular docking, to further confirm the repurposing of Meprobamate.

Overall, our prediction of pharmacologic properties is validated for 85% of the drugs with functional information [46]. Hence, motivated by this high prediction accuracy, we argue that it is extremely likely that the predicted properties will also be confirmed for the remaining 15%. Theoretically, drug-drug interactions actually express the way that drug behaviors interfere, constructively or destructively. Consequently, we consider our work as a strong argument for the multi-level network approach to drug repurposing, an approach that integrates the behavioral and structural perspectives [21]. In this context, we identify Diseasome (the Human Disease Network), the Human Connectome, and Human Genome Projects as appropriate platforms for our dual clustering approach.

Our described methodology was also used successfully for clustering patients in medical databases, for instance in cardiovascular disorders [242], sleep apnea syndrome [41, 43].These studies prove that disease risk factors do not associate at random, they rater converge towards well-defining patient phenotypes, which in turn provide valuable information for network medicine and personalized medicine approaches.

Part II

Career Development and Future Research Directions

Chapter 5

Laboratories and Infrastructure

In this second part of the habitation thesis we discuss the candidate's career evolution plans, as well as his research and teaching directions to be considered. Furthermore, we describe the strategies and approaches to put these plans into practice. We underline the capability of leading research teams, implying support for PhD research opportunities. As such, we enumerate available resources (personnel and infrastructure) and future research thematic, infrastructure, and financial offers for prospective PhD students.

The second part of the thesis is structured into a detailed overview on laboratories and infrastructure (present and planned), research project results and opportunities, financial support, future research directions, and teaching perspectives.

5.1 The Advanced Computing Systems and Architectures Research Group

The ACSA group was founded by Prof.emerit. Mircea Vladutiu and has grown, in time, into a dynamic team of experts in the fields of Computer Architectures, Reliability of Computation, Bio-inspired Computation, Quantum Computing, Reconfigurable Hardware, Computer Engineering in general, as well as newer directions like Complex Network Analysis, Network Medicine, Internet of Things, Big Data, and Machine Learning.

ACSA was built on the legacy of digital computing architectures, as our research approached topics in unconventional computing in an attempt to further the borders of known principles and mechanisms of information processing. We develop algorithms that are more efficient and map them to novel architectures in order to accommodate more complex computational phenomena and to achieve new levels of reliability and fault-tolerance. We search to uncover and exploit new computing architectures by observing biological processes and importing them into digital silicon. We strive to understand the intricate science of the new paradigm of quantum computing as a cross-breed of computer science and engineering, physics, and mathematics in order to devise new algorithms and deliver fast and accurate performance assessments. Finally, we approach real-life issues, aiming at improving and expanding their current solutions. Digital devices and algorithms can be of assistance in sleep apnea disorders and cognitive disabilities, to name a few. And they can help to better understand the complex interactions that make up the common urban traffic for the purpose of

maximizing its flow and providing improved management.

Given our experience in coordinating PhD students over a broad range of scientific topics, we consider our group as an attractive opportunity for PhD programs. With the newest inclusion of Network Science in our research portfolio – since 2011 – we are able to offer a full-fledged PhD program using Network Science in the field of Computers and Information Technology. To the best of our knowledge, we are also the first Computers Department to offer this opportunity at national level.

5.2 Available Infrastructures

In terms of available computing infrastructure for potential PhD students, our ACSA group disposes of two laboratories for student activities. As such, B520a is equipped with 18 desktop PC stations for individuals use of students during laboratory hours. The lab also provides an Epson projector and projecting screen, bought from the department's budget. The PCs were donated by the company Nokia. Laboratory B521 is a newer addition to our group, and is also equipped with newer and decently powerful desktop computers. Both rooms serve as teaching spaces for subjects like Computer Architecture, Computer Organization, Computer Engineering, Hardware-Software Codesign, Reconfigurable Hardware, Fault Diagnosis and Design for Testability, Big Data Visualization, Emergent Systems, Big Data in Healthcare etc,, used for both the Bachelor and Master programs. We also offer the B520b office for PhD meetings and group member meetings. The room is equipped with AC, WiFi, a multitude of computers and monitors, and extensive materials for development.

In addition to equipment obtained though donations and internal Department strategies, we have successfully added important computing hardware to our group's infrastructure. As such, we have bought a server (running a Linux distribution), financed by UEFISCDI project PN-III-P2-2.1-PED-2016-1145 "Inception: Internet of things meets complex networks or early prediction and management of chronic obstructive pulmonary disease", currently hosted by the Faculty of Pharmacy at UMFT. A second server (running Windows) was bought through the UEFISCDI project PN-III-P1-1.1-PD-2016-0193 "IMPRESS: Improving the prediction of opinion dynamics in temporal social networks: mathematical modeling and simulation framework", currently hosted in our B520b office.

Our Computers and Information Technology Department (DCTI) provides a set of cloud services, maintained and kept up-to-date by our System Administrator. We enumerate the following services made available by the DCTI network:

1. Personal department email service (@cs.upt.ro) with access to webmail, and email client.
2. VPN access available both for staff members as well as for students (after an explicit request), backed up by user-personalized access rules.
3. Web hosting and hosting of personal profile pages.
4. Access to a Linux system by staff.
5. Restricting student access by staff per laboratory.

6. Control and clean-up of computers in the laboratory by the teaching personnel (available only for laboratories with Linux).

7. Department GIT versioning system. The service is valid for both staff and students (at the explicit request of a coordinating professor).

8. Hosting virtual machines on VirtualBox infrastructure or own Cluster with Proxmox infrastructure (96 CPUs, 128 GB RAM, 8 TB storage).

9. Automatic installation of laboratory computers.

10. Active Directory infrastructure for managing laboratories with Windows Operating System.

11. Multiple GPU video card processing cluster. Currently we provide: NVIDIA Tesla K40m, NVIDIA GeForce GTX 1050 Ti, NVIDIA GeForce RTX 2080 Ti.

Another, newer solution offered by our department is the *Vision* cloud system which provides staff (including external collaborators) with cloud services similar to those offered by Google and Microsoft. The main purpose of the provided service is to offer department staff with cloud facilities while maintaining privacy (confidentiality). As such, the Vision cloud is hosted in the department's own data center, using its own infrastructure, and is protected by UPS and incremental backup systems.The platform used is Nextcloud, an open source platform that wants to help those who want facilities similar to those offered by major cloud providers, but, at the same time, want to keep confidentiality and total control over their data; as such, NextCloud is a private, open-source alternative to existing public cloud infrastructures.

The main functionalities of our cloud service are: file sync and storage (similar to Google Drive), file sharing with both registered users of the platform and with external users through temporary links, collaborative Document Web Editor (via OnlyOffice, and similar to Google Docs), Forms (similar to Google Forms), Contacts Manager (similar to Google Contacts), Calendar (similar to Google Calendar), Password Manager (installed as a plugin that offers an encrypted storage of users passwords with a strong facility to organize by categories and offers protection with an additional master password; the plugin easily integrates with Mozilla Firefox or Google Chrome), Maps (via OpenStreetMaps, similar to Google Maps), Instant messaging Chat (similar to Google Talk). The current storage space was upgraded from 5GB to 200Gb per user, which is far more than any free alternative on the market.

Also, in the current pandemic context, our ACSA team has successfully launched a virtual teaching laboratory equipped with a professional video camera, microphone, lighting conditions, computer, monitor, WiFi and a glass board for writing. The special glass board, combined with white markers, works best in dim lighting conditions as it is lit by several LED bands, placed along its top and bottom margins, which facilitate a unique (3D-like) visualization for students. Furthermore, the department has also equipped and opened an online teaching area. The special dedicated space is equipped with two computers, monitors, microphone, camera, graphical tablet etc.

The university also offers support with its Virtual Campus (CV) (`www.cv.upt.ro`), which is an online educational environment of academic support for all UPT faculties and for distance

learning. As such, in terms of teaching perspectives for prospective PhD students, we are confident that teaching will be an efficient process.

The available research infrastructure, used throughout the directed research projects, and available to all department personnel (including PhD students) is the ERRIS infrastructure "Centrul de Cercetări în Calculatoare şi Tehnologia Informaţiei" located in Politehnica University Timișoara (http://erris.gov.ro/UPT-CCCTI). As published on the ERRIS platform, the current infrastructure supports research on modeling and simulations, network science, big data, graph algorithms, and data mining, all of which are directly contributing research fields to our group.

In May 2020, Politehnica University Timisoara (UPT) started the implementation of the CloudPUTing project, a "High Performance Cloud Platform at Politehnica University of Timișoara", which has as main result the creation of a heterogeneous HPC cloud node dedicated to research projects.

The aim of the project is to increase the research and innovation capacity of UPT by creating an energy efficient, private cloud based on open technologies, attached to national and international networks of research cloud infrastructures, with applicability in collection, storage, analysis, distribution and the protection of the heterogeneous data masses, produced within the research and innovation initiatives carried out in the western region of Romania.

The main and direct target group of the project consists of all UPT researchers and PhD students, regardless of the research field in which they operate, who can benefit from the services provided by tools specific to the Computers & Information Technology field, provided in the form of centralized computing services, storage and productivity.

The project has an implementation period of 2 years, benefiting from a budget of over 4 Million Lei, of which 3.8M LEI are non-reimbursable funds financed by the program POC / 398/1-Development of networks of R&D centers, coordinated at national level and connected to European and international networks and ensuring researchers' access to European and international scientific publications and databases.

Overall, we are confident that the infrastructures offered by the ACSA laboratory, the Vision NextCloud platform of the Department, the Virtual Campus, and the future available CloudPUTIng HPC platform will offer any PhD student enough support for a large diversity of teaching and research tasks.

Chapter 6

Research Project Results and Opportunities

This chapter is dedicated to describing the main research tracks and results obtained from the two directed projects: IMPRESS (period 2018–2019) and PollStream (period 2020–2021). The first project was aimed at bringing contributions to Social Network Analysis, and the second project opens new possibilities with contributions in Computational Network Science.

6.1 Improving the Prediction of Opinion Dynamics in Temporal Social Networks: Mathematical Modeling and Simulation Framework

This project aims to create a modeling and simulation framework for better predictability on the dissemination of public opinion, on a predefined population, over time, by offering an improvement over the classical statistical approach based on opinion polls. The most common classical prediction methodology is based exclusively on data collected from a small subset of the target population, from which statistics are extracted. This method has a limited perspective due to its static presentation [243]; on the other hand, the approach we propose, using complex networks, involves simulating the dynamics of opinion using diffusion models, which can provide a better prediction of the distribution of opinions. Accuracy is further enhanced by examining the temporal and spatial distribution of opinion sources [244, 245].

The results of this project can be applied in the context of industry, such as marketing research. For example, web marketing and referral systems are becoming more popular for spreading the influence, scientific support is needed for revenue maximization strategies, applicable on social platforms like Facebook or Twitter. As such, the results of the project could explain which are the optimal opinion propagators to target, at what time and for how long to maintain the injection of opinion in a social context, thus minimizing marketing investments and maximizing the impact of a campaign. To achieve the goal of the project we formulated the following objectives (illustrated in Figure 6.1):

(O1) Topological analysis based on graph metrics and similarity analysis on empirical data sets to determine statistically relevant communication patterns, and how they relate to

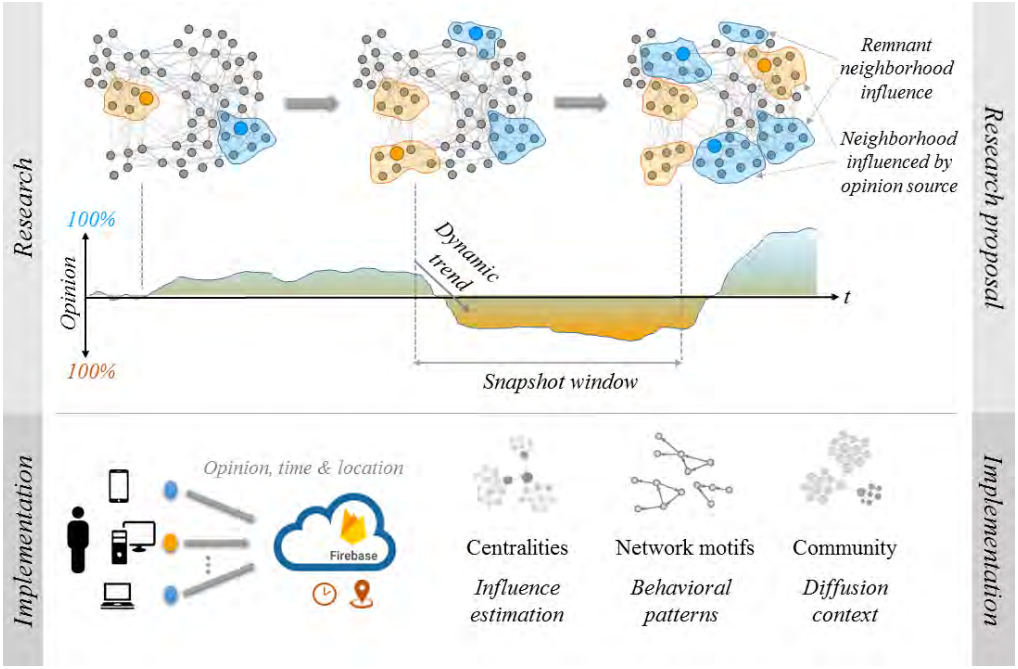

Figure 6.1: The main scientific objective of creating a dynamic model for injecting opinion, able to better predict the distribution of opinion at a given time t, taking into account the position, time and latent influence of each opinion expressed in the social network (upper panel). Implementation of the scientific model on the online platform, based on the data collected and storage on a cloud service. Time and location information is used to estimate influences and interaction patterns in the vicinity of the node (bottom panel).

opinion sources. (O2) The development of an innovative model of social interaction, considering the temporal aspect of opinion sources. (O3) Definition of a robust methodology for selecting opinion sources by analyzing the distribution of node and edge centrality for a better understanding of the emergence and growth of social networks. (O4) Synergy of results (O1-O3) with direct applicable socio-economic impact by developing a crowd-sourcing web platform for collecting anonymous empirical data from users (e.g., opinion, time, location). Each vote will represent an opinion injected into the topology of the social network, active for a predefined time interval, as defined by the results of the O2 research. Based on this information, the social simulation will run based on the results from O1 and O3, providing an improved predictability of opinion.

The activities carried out by the team members led to the achievement of all the planned results. As such, we worked on defining a framework for the comparison (benchmark test type) of the existing node centralities using the newly introduced concept of competitive diffusion.

Specifically, our method differs from the classical spread of SI, SIR, SEIR epidemics [7], by using the competition-based spread supported by our realistic tolerance-based diffusion model [34]. We studied a wide range of methods for estimating the influence of nodes [10, 64, 90, 99], and applied our new method to large, real-world synthetic datasets. To highlight the limitations of using a SIR simulation, we illustrate an example in Figure 6.2. In the first two panels, we apply two distinct classification methods (orange and blue), one at a time,

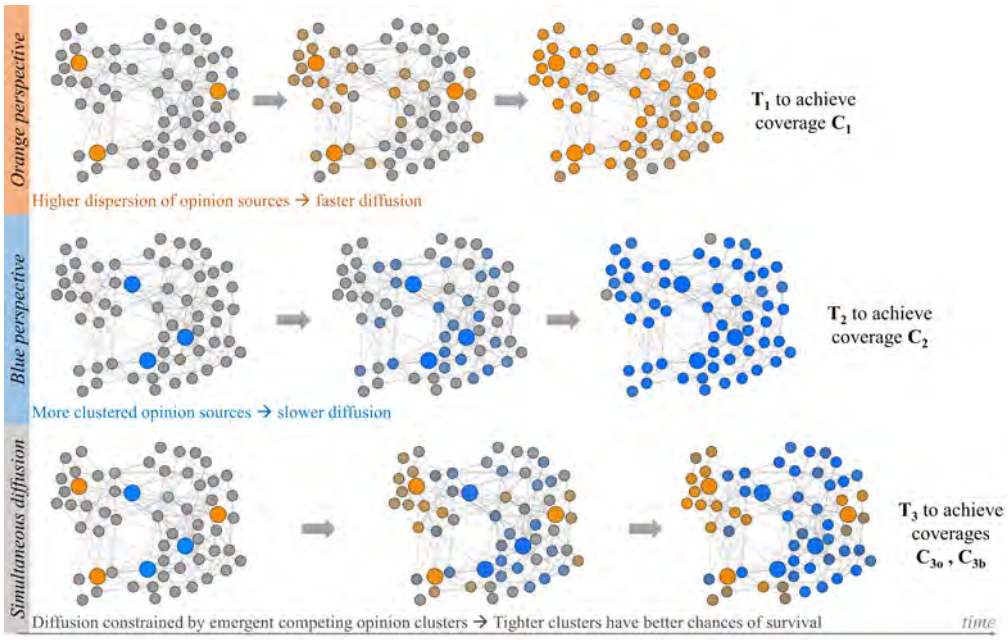

Figure 6.2: Limitations of comparing a dissemination process only from the point of view of a single opinion, although the real world context involves simultaneous dissemination and competition between several opinions.

and suggest that the diffusion process is unrestricted; we also suggest that the orange opinion manages to cover the network in time T_1, faster than the blue opinion in T_2, due to the greater dispersion of three original sources of orange opinion. In a SIR context, the two simulations may lead to the conclusion that the orange classification method is better (i.e., more efficient) than the blue one. In reality, we consider the scenario in the third panel to be the most likely. Opinions will be broadcast simultaneously and will face constraints due to competition on each node, i.e., orange and blue are mutually exclusive. In this case, we intuitively suggest that blue could gain in terms of network coverage because it has a closer initial community.

The computer simulations we employ show that our methodology offers a much larger quantitative differentiation between classification methods on the same data set and, in particular, the high granularity for a classification method on different data sets. We are able to identify - consistently - which classification method offers better performance than another, on a certain complex network topology. Our testing framework can provide a leap forward when analyzing real-time competition between agents. These results can bring great benefits to combating social unrest, spreading rumors, political manipulation and other vital and challenging applications in the analysis of social networks.

In terms of project results, we planned 4 deliverables: 1 impact journal publication, 2 conference proceedings papers, and 1 crow-sourcing web platform. Following the four actions provided for stages 1 and 2 of the IMPRESS research project, we exceeded the initial estimates as follows: 1 Q1 journal paper [39], 1 Q4 journal paper [246], 1 WoS journal paper [247], 4 conference proceedings papers [113, 248, 53, 249], and the web platform.

Indeed, further developments of our method are possible. For example, we can increase the number of opinion sources that operate simultaneously in a network. Consequently, the allocation of alternative opinions needs to be changed to suit all sources of opinion. A recent

study discusses the importance of targeting specific localized objectives, rather than achieving high network coverage [250]. Our method can be easily implemented to measure target coverage during or at the end of a spread simulation. Another study finds that each complex network can have a small "control set" of nodes, which, when triggered, will influence the entire network [251]. These control sets are considered to be surprisingly small (5-10% of nodes) and can also be associated with our benchmarking methodology. Overall, we believe that our work addresses a significant challenge in the study of opinion dissemination phenomena, and is a good starting point for many unresolved issues and new ideas found in the literature.

6.2 Agent-Based Interaction Models with Temporal Attenuation for Opinion Poll Prediction

This project proposes to corroborate its theoretical research results and apply them into an applicative context, namely that of electoral poll forecasting. To this end, we build upon the premises that we can extrapolate the macroscopic opinion dynamics of a society by inferring microscopic temporal dynamic models during the pre-election period. Our hypothesis is that the timing of publicizing opinion polls (i.e., opinion injection) plays a significant role in how opinion oscillates.

Current state of the art in electoral forecasting employs multilevel regression and post-stratification (MRP) [167]. However, MRP method is often cumbersome to apply, needing economic indices and detailed demographics to be accurate. Alternatively, we propose to elaborate on the concept of temporal attenuation (TA) [53, 60], which models the timed oscillation of poll data as opinion momentum. For this, we propose a research methodology based on computer simulation of information diffusion, on large datasets, using novel agent-based models [34], and integrating them with TA in order to improve the forecasting performance of opinion polls. We strongly believe that the contributions highlighted in this project proposal answer important and timely scientific and social issues, and will constitute an incentive for further research and collaboration.

The idea of influence maximization (IM) in networks is a nonlinear problem, which represents an element of difficulty for modeling and predicting dynamical systems represented as complex networks [252, 253]. Identifying an optimal, namely minimal set of spreader nodes, remains unsolved despite the vast use of heuristic strategies [7, 254]. Answering these questions can lead to developing a set of ubiquitous strategies for efficient control of information diffusion with direct impact in marketing, sociology and business applications.

In [255], we develop a new framework of computational intelligence, called "Optimal Genetic Selection of Opinion Distributors" (GenOSOS), to optimize IM, compared to state-of-the-art methodologies in selecting distributor nodes based on node centralities (degree, betweenness, PageRank and k-shell). Overall, (i) we propose the GenOSOS model, a genetic algorithmic approach to the IM problem, which is an original attempt to address the trade-off between spacing between distribution nodes and diffusion coverage; (ii) we propose a modeling specific to the problem of population and chromosome representation. Moreover, we implement the fitness function based on a graph coloring algorithm, which can accelerate the convergence of the spreading process. (iii) we define an individual (chromosome) as a unique set of diffusion, bringing customized implementations of crossover and mutation operators;

(iv) we estimate the effectiveness of GenOSOS on synthetic and real-world networks. The experimental results show that our algorithm has competitive performances compared to similar selection methods based on centralities. The GenOSOS genetic algorithm, shown in Figure 6.3, is based on three customized genetic operators – elitism, crossover and mutation.

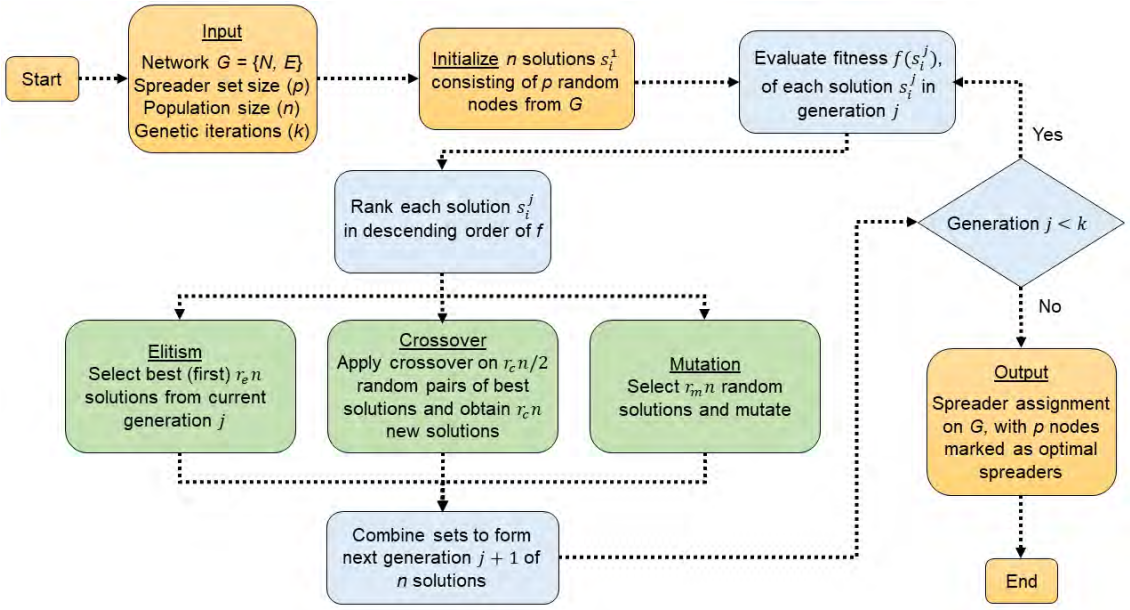

Figure 6.3: Flowchart of GenOSOS emphasizing the main algorithmic steps: input/output (orange), generation control (blue), and genetic operators (green). According to the flowchart, the algorithm finds an optimal solution s_i^j for placing p spreader nodes in a graph G, and runs k genetic iterations consisting of three operators that are used to generate n new solutions, from generation j, for the next generation $j + 1$. The output consists of a set of p nodes marked as spreaders in graph G.

The detailed analysis on three categories of network datasets show that the potential of our proposed solution is not only viable, but offers superior results compared to the state of the art centrality approach. Specifically, GenOSOS obtains a 11.45% higher coverage, averaged over all (12) used datasets. In essence, our solution is superior to the state of the art on 7 out of 12 datasets (58.3%) in terms of diffusion coverage.

In addition, we have formalized the temporal attenuation (TA) framework for electoral forecasting. From a scientific point of view, this work builds upon the premises that we are able to extrapolate the macroscopic behavior of a society (here, in the context of elections) by inferring microscopic temporal dynamic models during the pre-election period. TA uses solely pre-election poll data to improve electoral forecasting accuracy. The novelty of TA relies on characterizing the dynamic momentum of opinion, which builds up and dampens in the general population, according to the injection of pre-election polls data.

The efficiency of our forecasting model is measured using the Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) criteria. Our results are benchmarked against ARIMA models, fitted to our pre-election data, and the best pollster predictions for the US Presidential elections ranging from 1968–2016, including more recent pollsters using

MRP (2012, 2016).

In this project, (i) we formulate an analytic methodology for modeling the macro-scale evolution of a multi-opinion system, targeting better election poll forecasting; (ii) we define an experimental setup, based on pre-election data, to validate the underlying assumptions of our approach; (iii) we present a comprehensive case study on US Presidential Elections, ranging between 1968–2016, to measure the efficiency of our approach, using MAPE and RMSE, against state of the art forecasting estimates, including ARIMA and MRP, as recently used by the best pollsters in the USA; (iv) we explore the feasibility of applying TA in real time, during an ongoing pre-election period, and benchmark its performance against MRP at several points in time, relative to the election day. An overview of our TA framework is exemplified in Figure 6.4.

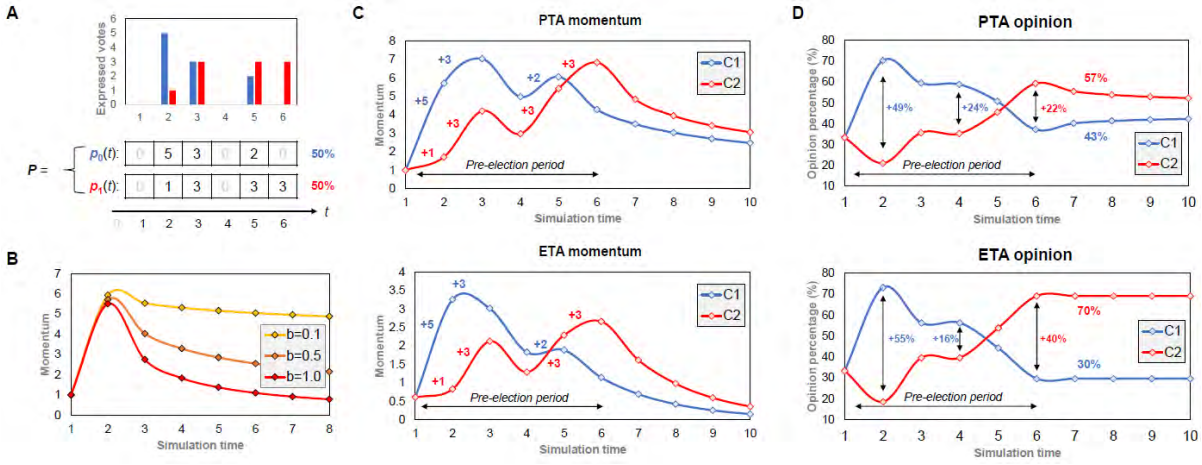

Figure 6.4: Overview of the temporal attenuation (TA) model applied on two candidates (C1-blue and C2-red) with 6 days before election. *(A)* Surveys are collected for the two candidates at $t = \{2, 3, 5, 6\}$. From these, set P is assembled consisting of poll vectors $p_0(t)$ and $p_1 t$. *(B)* A higher β translates into a more abrupt damping of the momentum. *(C)* Momentum $M_i(t)$ for PTA and ETA corresponding to the poll vectors $p_0(t)$ and $p_1(t)$. Individual votes are displayed in absolute value on the graphs. The simulation using dataset P corresponds to the pre-election period ($0 < t \leq 6$). *(D)* Opinion $\Omega_i(t)$ evolution for PTA and ETA corresponding to the momemtum in panel (c). Several poll differences are displayed at $t = \{2, 4, 6\}$ using the color of the virtual winner at that moment.

The planned results for the PollStream project are: 2 journal publications, 2 conference proceedings papers, and one mobile/web simulation platform for diffusion processes on large social networks. By April 2021, we have published: 2 conference proceedings papers [255, 52]. In addition, we have 2 journal papers (Scientific Reports–Q1, Expert Systems with Applications–Q1), and 1 conference paper (KES-2021) under review.

The importance of this project consists in defining a complementary electoral forecasting method, which does not need any demographic, economic, or political information related to the context of the election. This distinction represents a significant advantage for TA over alternatives, like MRP, since our method may be applied, given enough reliable public polls, in any political region of the world. In this sense, the results further presented in this paper did not need consideration of any additional information about the USA during the period

1968–2016. Since we address an ongoing socio-political challenge, we believe our investigations will have a high societal relevance, especially in industry and governments, as they provide insights into better understanding the dynamics of electoral systems.

6.3 Career Opportunities

In this section we discuss some of the financial, research and professional (didactic) career opportunities that can open for prospective PhD students joining our ACSA research group.

To provide a quantifiable example, we detail the financial opportunities that have been available during the during the 10-year (2011-2021) research activity of the candidate. As such, in addition to the academic-didactic activities in the department, we managed (as director) 2 national research projects (38K euro & 51K euro), financed by UEFISDCI. we also joined as project members 2 international project teams and 5 national project teams. The international projects were financed by Linde (75K euro), respectively by the Horizon 2020 framework (131K euro). The national projects were financed by ARUT (10K euro) and UEFISCDI (120K euro, 120K euro, 100K euro, 120K euro).

On behalf of these projects a number of technical equipment was bought to serve each team member. For instance, we bought two servers and two sets of personal laptops (in 2013, 2019). Also, we were able to publish several open-access journal papers, e.g., [46, 27, 39, 43, 47, 44], and participate at several conferences in Europe and North America.

The two projects managed by the thesis candidate generated additional indirect costs (i.e., *regie*) which can be used for upgrading technical equipment, conference participation, open-access journal publications, and other motivated professional development. Obviously, these project savings can be used to boost the early career of new doctoral students. Given the current and planned project savings, we have the possibility of supporting 2-3 years of conference participation for students. Also, given the financial support of the University for high impact journal publications, we are confident that students will be able to publish their results in a timely manner.

Also, in terms of open research ideas, we mention contributions in educational science using network science, and contributions in sleep research using machine learning and big data analysis. To this end, based on data gathered during the NOVAMOOC project (2015–2017), we started a follow-up study on student archetyping. This collaborations with the West University is open for research.

Our work focuses on placing a cornerstone for a framework of personalized MOOCs (massive open online courses) in the future. While some students will always be more prepared than others to embrace online education, we consider that the mass education of the future needs to be personalized [256, 257]. Even though this concept sounds difficult to apply, we suggest a step wise refinement through which we define increasingly reliable and specific profiles for online students. This work sets such an example, though which one could classify newly joining students into a specific profile, and thus personalize their way of being taught, graded, involved in social activity and projects, given responsibility etc. We believe that our study will enhance the understanding of how students relate to MOOCs, and thus open a new path of personalized online education.

The current drafted work aims to make a leap further in educational science, and combine

complex network analysis and sociology to model and analyze the emerging profiles of the new digital student. As such, we curate data through a large-scale online questionnaire, and analyze the opinion of 637 students from Romania regarding the advantages, disadvantages and reasons to choose MOOCs. Based on their expressed opinion, we create two graph models of compatibility based on key individual traits, and find six distinct student profiles in terms of engagement in MOOCs, and seven profiles for non-participants. Furthermore, we discuss these profiles and explain the implications, limitations and perspectives of this study. We consider our findings an important milestone both in understanding the needs of future modern students, and in optimizing the way online courses are developed to serve the challenges in personalized education.

In parallel, we aim to submit new project proposals, such as a TE (Tinere Echipe) or PED (Proiect Experimental Demonstrativ) project. Given the current ACSA members, we find both alternatives feasible. The TE can be targeted in the field of computational epidemics (a current impactful hot topic sub-field of computational network analysis). The PED may be targeted at the implementation of a sleep research related tool (for OSA).

Finally, the current Horizon 2020 project (2021–2025), in which several members of the ACSA and UMFT team are involved (same partners as for the projects on OSA and COPD diagnosis), open up endless possibilities for current Master's students, as well as current and future PhD students. Given the large number of research centers involved in the project, most research publications will have a solid scientific impact, this opening opportunities of visibility for all participants.

6.4 Conclusions

All the mentioned scientific achievements illustrate the high potential of the thesis candidate as an independent researcher and his ability to manage research teams and future PhD students. Furthermore, he is actively collaborating with several research groups, with some notable results: Carnegie Mellon University (CMU) / University of Texas (Prof. Radu Marculescu) [34, 27, 38], University of Southern California (USC) (Prof. Paul Bogdan) [46, 47], and Central European University (research visit supported by the IMPRESS project). In the near future, and based on contact at conferences, we propose establishing collaborations with the DSG group at Wroclaw University of Science and Technology (Prof. Przemysław Kazienko), the Collide group at Universitaet Duisburg-Essen (Prof. Ulrich Hoppe), and the DNDS department at CEU (Prof. János Kertész). These scientific collaborations should positively strengthen the international visibility, high originality of research, and relevancy of the work of ACSA group, as all mentioned teams work within the field of network science.

Furthermore, the high number of projects in which we participated over the last 10 years (2 as director, 7 as member) illustrate the accumulated experience and potential of attracting new funds for future ACSA members, within the UEFISCDI TE and PED programs. Moreover, the personal and University financing frameworks can support the open-access publication and conference participation of doctoral students in the near-term future.

Chapter 7

Future Research Directions

In this section we enumerate current research development plans in the areas of computational epidemics, network analysis in educational science, and network medicine. The diversity of these areas ensures a high attractiveness for potential new doctoral students. Furthermore, our experience, visibility and project partnerships can further boost the impact of future research for potential members in the ACSA group.

7.1 Computational Epidemics using Network Science

Understanding the dynamics of large, resurgent epidemics is an ongoing scientific effort aimed at controlling and preventing the spread of infectious diseases. Disease epidemiology, computational epidemics, network science, and computer science are some of the major scientific fields involved in this high impact social challenge. Notable research has been conducted over the past 30 years, answering important questions on the processes driving epidemics, and proposing strategies for prediction and control [258, 259, 260, 261]. The heavy socio-economical burden of epidemics has been demonstrated repeatedly during crises like SARS[262], Ebola [263] or recent COVID-19 [264]. To this end, we need to be able to predict long-term epidemic evolution, and the impact of governmental interventions, like isolation, travel restrictions, and vaccination/immunization of the population [265, 51, 266, 267, 268].

Along this new research direction, fueled by the novel Coronavirus pandemic, we propose two possible research directions. First, we need to understand the effectiveness of isolation strategies, as adopted by many countries starting with March 2020. In particular, we aim to model centralized and decentralized strategies of isolation, and compare them, based on a novel epidemic model (SIRCAS), using large heterogeneous network topologies. Second, we study heterogeneous population structures and mobility models (multi-scale, hierarchical) which increase the realism of epidemic simulations.

7.1.1 Centralized and Decentralized Isolation Strategies and Their Impact on Epidemic Dynamics

In the absence of an approved pharmaceutical treatment (or vaccine) and in-depth knowledge of the spreading mechanism, the best strategies against COVID-19 consist of reducing the

interactions between susceptible and infected individuals, *e.g.*, through early detection and social distancing [265]. Indeed, such non-pharmaceutical interventions (NPIs) turned out to be very effective during previous pandemics [269, 270].

The potential effect of social distancing interventions on the COVID-19 has already been studied in Singapore [271]. Indeed, Singapore was among the first regions to report imported cases and has so far succeeded in preventing community spread [51]. However, the scale and severity of the Singapore interventions are small in comparison with the measures implemented in China in response to COVID-19. The core Chinese interventions include shutting down schools and workplaces, closing roads and transit systems, canceling public gatherings, and imposing a mandatory quarantine on uninfected people (even those without known exposure to the virus) [272]. Although these actions seem to be working so far, imposing similar restrictions in other countries represents an ongoing challenge. To convince people, governments, and public authorities around the world that such extreme limitations are necessary, we need to back them up with scientific evidence. Indeed, in the absence of clear evidence, some countries will hesitate to adopt the strong social distancing actions, and this may have dire consequences. For instance, Sweden took only mild restrictions, with restaurants and bars still being open, playgrounds and schools too, and the government relying on voluntary action to stem the spread of COVID-19 [266].

Given the dynamics of COVID-19 spreading, we can assess the efficiency of the control measures for this novel pathogen by using mathematical modeling coupled with computer simulations of infectious spread under various scenarios. To this end, we propose [51] a new agent-based outbreak model called SICARS (*Susceptible - Incubating - Contagious - Aware - Removed — Susceptible*), which allows us to assess the impact of the centralized and decentralized isolation strategies on COVID-19 spreading across complex heterogeneous networks. Consequently, we run simulations of SICARS and test two fundamentally different strategies, as well as their combined effects:

1. Centralized (C) strategy, such as the government-imposed lockdown or quarantine; this means social distancing by the *synchronized* removal of a specified ratio of node social ties from the entire social network.

2. Decentralized (D) strategy, such as aware-isolation (DA) and auto-isolation (DI); this means an individual-level social distancing by *asynchronously* removing a specified ratio of personal social ties. More precisely, in DA, the individuals who become aware of their sickness cut the social links in their ego-network. In contrast, in DI, the healthy neighbors of sick individuals isolate themselves from the infected. In both scenarios, the social ties are removed repeatedly (*e.g.*, daily) based on a probability parameter.

3. Hybrid (C+D) strategy, whereby both policies are combined, hence the removal of the social ties involves both centralized and individual-level decision mechanisms. To this end, a fraction of social links are synchronously removed from all nodes in the network, then followed by repeated asynchronous distancing through self-aware isolation of sick nodes (DA), as well as auto-isolation of healthy nodes from their sick neighbors (DI).

The SICARS pandemic simulation is based on the model depicted in Figure 7.1a and exemplified in Figure 7.1b, with a small example network of five connected nodes, which become

infected [265]. By splitting the infectious stage into two sub-states, *contagious* and *aware*, the model becomes unique in its capability of implementing centralized and decentralized isolation strategies. In Figure 7.1b, the outbreak seed is node A which needs a period of $\delta_{Incubation}$ days to become contagious. After becoming contagious, node A spreads the disease to its neighbors B and C; these, in turn, become incubating after several days of contact with A. Node A is still not aware that it became a disease carrier, nor that it has infected his neighbors. After a delay of δ_{Aware} days, node A becomes aware of its infectious condition (or state). At this point, nodes B and C know that A is a threat, but are unaware if they have contacted the disease. In the same manner, nodes B and C become contagious, then aware. Nodes D and E are further infected, and the process continues similarly. After $\delta_{Removal}$ days (measured individually for each infected node), every node changes to one of three states: recovered (B, E), dead (C, D), or susceptible (A). The susceptible node A may start the same process all over again, going through all the SICARS states.

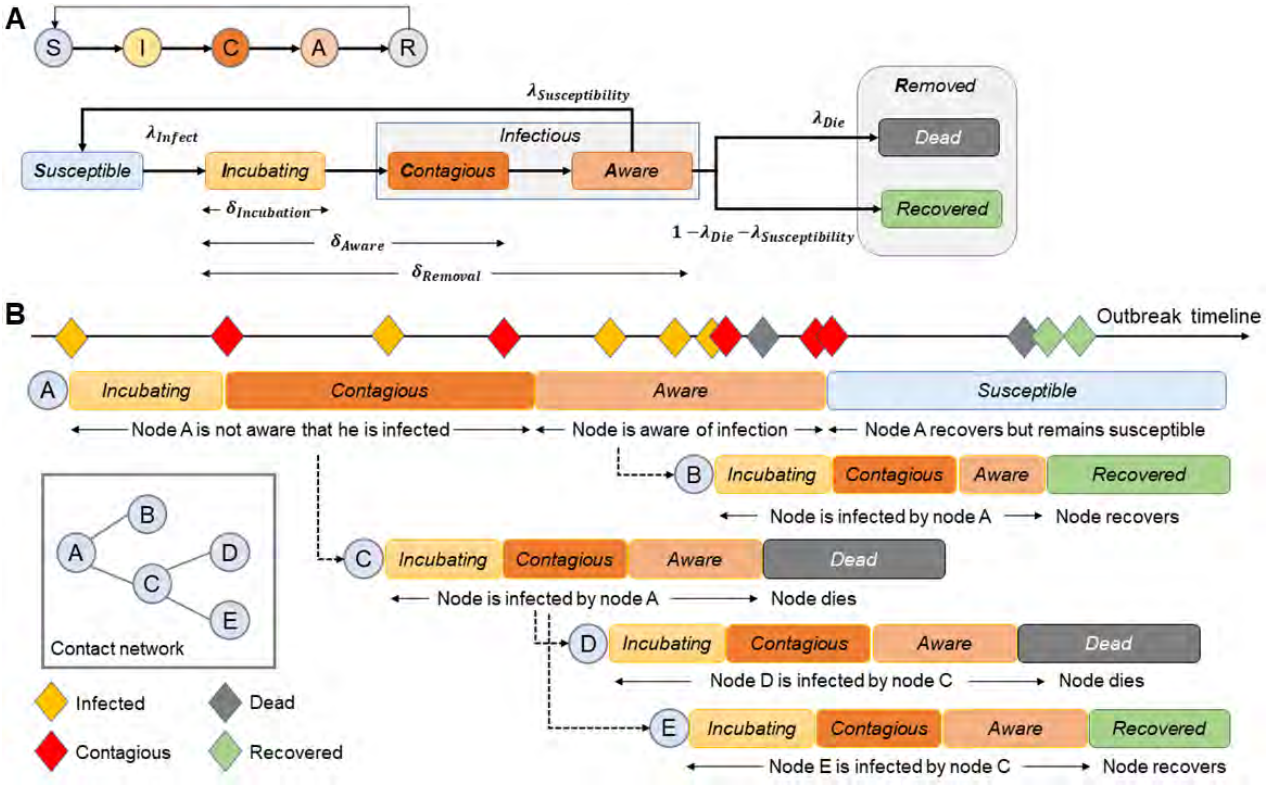

Figure 7.1: (a) The states and parameters defining the SICARS model. (b) Example of an outbreak process according to our SICARS model, when used over a hypothetical contact network with five nodes (A–E), starting with infected node A.

Taken together, we aim to improve the state-of-the-art with the following contributions:

- In contrast with the COVID-19 epidemic modeling proposed in [273, 268, 265], where isolation is modeled by reducing the size of the susceptible compartment, our network-centric approach targets isolation strategies as (local and global) edge removal mechanisms, hence an emergent and more realistic transmission dynamics.

- As the differential equations of SEIR do not apply to COVID-19, we use distributed simulation on well-known complex topologies instead of the compartmental models based on random uniform contact networks that are typically used to study epidemics spreading [274, 275, 276, 277, 268].
- Instead of focusing on assessing a specific isolation strategy (*e.g.*, Singapore), our study aims at differentiating the efficiency of centralized (global, government-imposed), decentralized (local, self-imposed), and hybrid isolation while using the SARS-CoV-2 specific biological parameters.
- We focus on providing a more accurate quantification of the *impact* of different levels of social distancing, and explore the realistic scenarios of delayed and progressive application of isolation, in the context of the current pandemic.

7.1.2 Geo-Hierarchical Population Mobility Model for Spatial Spreading of Resurgent Epidemics

We find recent studies that are predominantly augmenting mass-action models into tools suitable for analyzing large scale epidemics [278, 279, 280, 265, 268, 281]. However, in most cases, we notice that their underlying epidemic models (e.g., SI, SIS, SIR, SEIR, SIRS) adopt *homogeneous* mixing of the population (i.e., all individuals are fully connected inside single scale compartments or stochastic blocks) [282, 283, 284, 280, 281]. This over-simplification of social organization lacks the complexity of global scale population organization, which is dictated by geographical, historical, demographic and economic factors. Consequently, numerical simulation of such simplified models can lead to over- or under-estimations in terms of epidemic size [266, 273, 281] or duration [280, 282, 268, 265].

Conversely, we find some important studies which developed more robust and realistic models for epidemic dynamics and contagion, for heterogeneous population organization and human mobility. Without a doubt, the structure of networks is found to be paramount in explaining infectious spreading patterns [261] seen in empirical data for transmissible diseases; also, community structure is a known key factor influencing the speed of epidemics [285]. This research direction aims to prove the importance of incorporating accurate population modeling and human mobility, which represent ongoing challenges due to their theoretical complexity as well as limitation in available data for validation. Thus, we propose the novel geo-hierarchical population mobility model (GHPM) which lies at the crossroads of population organization and mobility, both of which are key aspects to consider when targeting realistic large-scale resurgent epidemic outbreaks [52]. We propose the novel idea of distributing a population into spatially organized communities (i.e., human settlements), which are then organized into a hierarchy of administrative divisions (i.e., district, neighborhood, street, block, household). Thus, the population is partitioned with very high granularity all the way down to household-/family-level, containing just a few individuals, but where the transmission risks are highest [286]. Embedded into our population model, we further propose a novel mobility algorithm based on the geographical distance between settlements and their size, which determines the complexity of the underlying hierarchical structure. Altogether, to test the complex interplay between the population mobility model of GHPM and the dynamics of a custom SIRV epidemic model with relapse, we use detailed computer simulations.

In contrast to other computational models like GLEaMviz [287], RAPIDD Ebola forecasting [288], or [7], our GHPM model is, to the best of our knowledge, the first of its kind to combine a geo-spatial and a hierarchical model to structure population. Using available empirical data on influenza and COVID-19, we show how GHPM reproduces similar epidemic dynamics (e.g., size, waves). The main focus of this study is to determine how the population organization, travel distance and travel frequency affect the spread of diseases on large scales (country-level), and how restriction and immunization strategies can be applied efficiently to control epidemics.

Figure 7.2 represents both a conceptual example of computing the GHPM mobility probabilities based on position and populations size, as well as a real-world mapping over the Kingdom of Spain. In Figure 7.2b, the modeled population is 33M inhabitants (70% of real size) placed in 735 settlements, all within a bounding area of 1000km × 850km (the Canary Islands have been omitted from the figure, but are included in the data model).

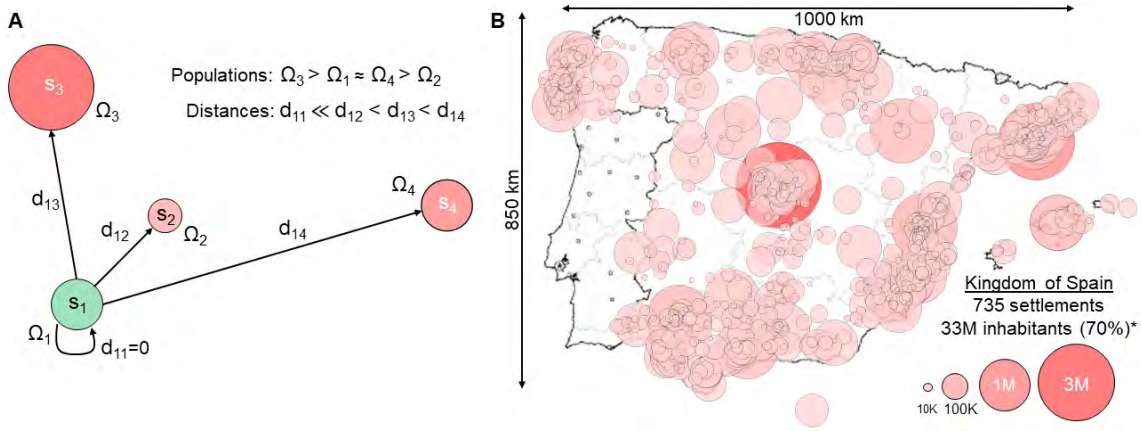

Figure 7.2: (a) Conceptual representation of the inter-settlement mobility on an example GHPM with 4 settlements $s_1 - s_4$. Any individual from s_1 (green) has an associated probability to remain within the same settlement or move to $s_2 - s_4$. The parameters affecting the probabilities are: target settlement population Ω_j, and distance d_{ij} to settlement. (b) Example of GHPM mapping of 735 settlements over Spain.

In this research track, we address the issue of modeling mobile heterogeneous population systems, where the community structure is defined by actual real-world geo-spatial data (i.e., position and size of human settlements). We summarize the research plan as follows:

- We introduce the geo-spatial hierarchical population model (GHPM) to investigate how the duration δ, size ξ and dynamics of an epidemic are quantified, comparing to a similar homogeneous mixing model and to real COVID-19 data [289]. Our research focus on the community structure and individual mobility [52], as well as introducing the original SIRV transmission model into computer simulations.

- We define the population system (e.g., a country) as a stochastic block model (SBM) where blocks (or communities) are modeled by real-world settlements from a chosen country. Their size and spatial positioning (latitude, longitude) are set by real-world data.

- We further define original individual mobility patterns based on the population (size) and distance between any pair of communities. Intuitively, individuals are more likely to move to a larger and/or closer settlement, than to a smaller and/or distant one [290].

- We show that the number of settlements in the population system, as well as altering the settlements' density (leading to more compact, or more sparse geo-spatial organization of communities) can directly impact the outbreak duration and size.

7.2 Network Analysis in Educational Science

The current learning environment is characterized by openness and dynamism, so that a significant proportion of students have a declared preference for flexible learning [291, 292], through which they can fulfill their academic pursuits, as well as job responsibilities and family chores [48]. The amount of data generated by virtual learning systems sometimes overwhelms educators, who are unable to process the information without the support of special business intelligence tools and techniques specific for large data and Big Data analysis and visualization [293].

The performance of students enrolled in both offline and online education is important for many institutions, because their strategic programs can be planned to improve this performance [48]. There are studies in this sense, taking into consideration the average grades upon graduation, or track completion [294, 295], based specifically on data mining techniques in order to predict the drop out rate of students. In particular, decision tree techniques [296, 48] are applied to create surveys that predict the likelihood to drop out from college, then these are turned over to management for direct or indirect intervention [295].

Following our study on decision tree learning used for the classification of student archetypes in MOOCs (massive open online courses) [48], we plan further follow-up studies based on our results from the NOVAMOOC project. As such, as a novel analytical approach, we rely on network science to go beyond the perspectives made possible by means of a classic statistics [5, 9]. More specifically, the methodological novelty for this study consists of clustering students based on their expressed reasons to participate or avoid online courses, by modeling students in a complex network where edges between them are formed by overlapping compatibility.

Our dataset is based on an online survey and consists of 69 questions from which we extract relevant data, such as *Demographics* (Gender, age, university, faculty, specialization, study year), *Participation* in past MOOCs (duration, language, finalization and certificate attainment), *Reasons for not participating* in MOOCs, or *Advantages* and *disadvantages* of participating in MOOCs, *Interests* in a future MOOC. At the time the study was conducted, we gathered $N = 637$ unique answers, out of which 472 students did not participate in a MOOC in the last 3 years or at all, and 165 students who participated in MOOCs before.

After collecting the data, we can create a compatibility graph of students, similar to our previous state-of-the-art methodology [57, 146, 58, 41]. What differs in the current approach is that the bipartite graph we start from is not based on social collaboration or physical resemblance between modeled nodes, but on common educational and individual aspects of each student. The reason for creating such an innovative graph representation is that

individual personality patterns are more relevant than physical or social personal features in the context of academic participation.

We represent students as nodes S in a graph $G = \{S, E\}$, and add links E based on student *compatibility*. Particularly in this study, compatibility is defined as the number of common individual traits two students s_i and s_j share in common. The more traits two nodes have in common, the greater is the weight w_{ij} of the edge e_{ij} connecting them. We focus on analyzing how common individual traits affect the emergence of clustering of students in the context of MOOC participation and awareness, though other educational, physical or psychological insights may be considered for future work.

Out of large number of questionnaire answers (69), we must select the criteria used as input for building the graph. The input parameters are used for classification/clustering, leaving all other answers as output /descriptive parameters. The dataset can be divided in two major, non-overlapping sets: students which have participated in MOOCs, from which we can analyze the advantages and disadvantages from their point of view; and students which have not participated in MOOCs, from which we can analyze the reasons for not doing so. As such, we use the following input/output parameters for classification for the two datasets:

1. **G1** (165 students which have participated in MOOCs)
    - Input has 7 parameters based on course elements such as participation, costs, finalization, certification, knowledge, gender.
    - Output consists of 10 advantages and 10 disadvantages of MOOCs, plus basic information.
2. **G2** (472 students which have not participated in MOOCs)
    - Input has 10 parameters based on 9 reasons for not participating in MOOCs, and gender.
    - Output consists of 10 advantages and 10 disadvantages of MOOCs, plus basic information.

By analyzing the layout of the 6 resulting communities in G1, we can support the claims through the visual observation presented in Figure 7.3. Namely, homophily plays an important role such that each node is placed in the vicinity of other nodes with the same seven chosen traits. By overlapping the measurable output properties we can describe each of the 6 communities in G1 in a unique way. The same approach, can be applied for network G2.

Our future work focuses on placing a cornerstone for a framework of personalized MOOCs in the future. While some students will always be more prepared than others to embrace online education, we consider that the mass education of the future needs to be personalized [297, 256, 257]. Even though this concept sounds difficult to apply, we suggest a step wise refinement through which we define increasingly reliable and specific profiles for online students. This work sets such an example, though which one could classify newly joining students into a specific profile, and thus personalize their way of being taught, graded, involved in social activity and projects, given responsibility etc. We believe that our study will enhance the understanding of how students relate to MOOCs, and thus open a new path of personalized online education.

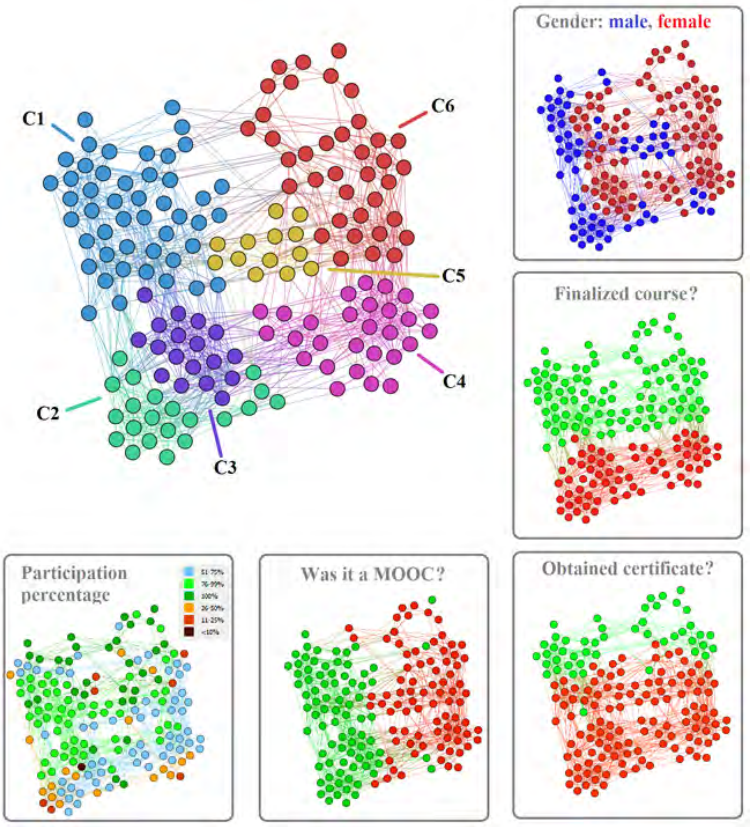

Figure 7.3: Representation of G1 consisting of the 165 students which participated in MOOCs. The large panel shows the six emerging communities of students (profiles), and the smaller panels show nodes colored by different binary metrics regarding the online course. Green nodes are students who positively answered questions, and red nodes represent negative answers

Another, well-known and persisting problem in modern education is academic dishonesty. There are various forms of such dishonesty, like plagiarism, which is often debated in the media, but cheating during examination perpetuates, and remains one of the oldest and most impactful forms of altering one's educational outcome and diminishing an institution's reputation. The applied prevention of this phenomenon is the subject of scientific attention, but the existing methods are most of the time insufficient or poorly applied. By analyzing the types of problems that occur during written exams, we have developed and implemented an innovative solution to decrease the amount of unwanted collaboration among students, by using their underlying friendship topology to the students' disadvantage [50].

Consequently, we have introduced an original student placement strategy inspired by the interdisciplinary field of social networks analysis, and compared it to no placement strategy at all, and to the state-of-the-art random method [50]. Our method is based on acquiring the social network of students participating in the exam, and using genetic algorithms to rearrange them in seats, such that there is minimal overlapping between real-world friendships and seated neighbors. The three methods have been applied independently on six different pools of students over the period 2013–2016, resulting in an extensive case study on $N = 586$ students in the Romanian higher education system. In [50] we discuss the meaning of the

results, as well as the applicability and limitations of our method. The analysis is presented both through empirical measurement of interaction between students during exam, as well as statistically, by introducing a metric for the placement effectiveness ϵ. Our proposed solution offers average improvements of $\times 2.8$ in terms of breaking up real-world friendships, and a $\times 3.3$ reduction in terms of empirically measured student interaction. On the other hand, we showcase that the easier to implement random placement brings about lower improvements of $\times 1.7$ (statistical) and $\times 2.3$ (empirically measured), over no seating strategy. Considering that many educational systems are unaware how vital the customization of student rearrangement is, we consider this case study to beacon an important institutional problem all around the world.

Given the increased availability of digital educational data, we foresee an improvement of our placement strategy [50] using datasets corroborated with individual information from a platform such as our Universities Campus Virtual. Hence, a bipartite graph between students and their online achievements (weekly attendance, assignments, activity, quiz grades) can be defined. Arching over our numerical analysis, the meaning of our case study analysis is that there is room for improvement in present-day cheating prevention systems. Apart from the reduced unwanted communication, we are able to change the subjectively perceived attitude of students, who are surprised, to some extent, by the unusual examination procedures. This in turn makes them focus more on the exam at hand, knowing that they are actively discouraged to communicate and cheat.

7.3 Network Medicine, Machine Learning and Engineering for OSA Diagnosis

The outline of the Horizon 2020 project, entitled "Sleep Revolution", is that obstructive sleep apnea (OSA) is associated with a high economic burden, and is currently rising. Almost 1 billion people worldwide are estimated to have OSA. The current diagnostic metric, however, relates poorly to these symptoms and comorbidities. It merely measures the frequency of breathing cessations without assessing OSA severity in any other physiologically relevant way. Furthermore, the clinical methods for analyzing PSG signals are outdated, expensive and laborious. Due to this, the majority of OSA patients remain without diagnosis or have an inaccurate diagnosis leading to sub-optimal treatment. Thus, it is evident that more personalized diagnostics are required including predictive and preventive health care and patient participation. The Sleep Revolution project aims to develop machine learning techniques to better estimate OSA severity and treatment needs to improve health outcomes and quality of life. These techniques are implemented to high-end wearables developed in this project to alleviate the costs and increase the availability of PSGs. Finally, we aim to design a digital platform that functions as a bridge between researchers, patients and healthcare professionals.

These goals are to be achieved through extensive collaboration between sleep specialists, computer scientists and industry partners. The collaboration network consists of over 30 sleep centers working together to provide the needed retrospective data (over 10000 sleep studies). The multi-center prospective trials involve experts and end-users to assess and validate the new sleep revolution diagnostic algorithms, wearables and platforms. With the commitment of the European Sleep Research Society and Assembly of National Sleep Societies (over 8000

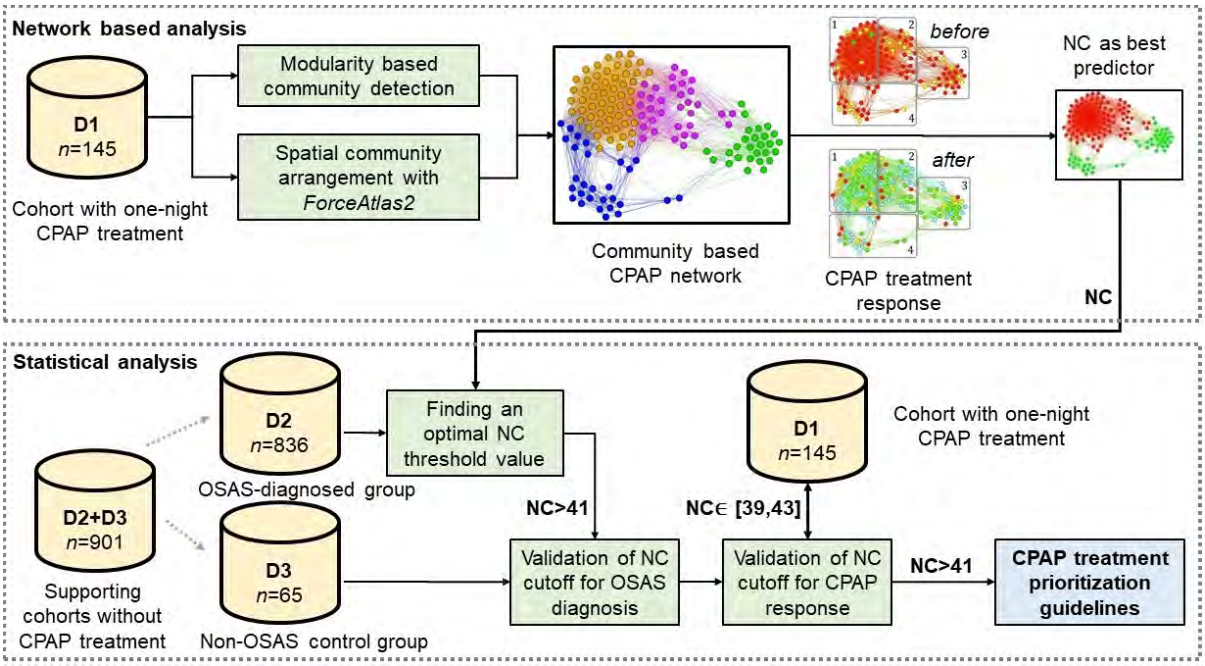

Figure 7.4: Overview of datasets and study design. The network based analysis stage uses cohort D_1 to model a CPAP patient network. By corroborating the community structure of this network with the CPAP treatment response of each patient (i.e., measured as AHI improvement), we extract NC as the most significant indicator of AHI improvement. We further use this information in the statistical analysis stage, which uses a larger $D_2 + D_3$ supporting cohort to find an optimal NC threshold value for OSA-diagnosed patients. Cohort D_3 is the non-OSA control group. The study results in the definition of a rule of thumb guideline for CPAP treatment prioritization of patients with OSA (blue).

members), we have the unique possibility to create new standardized guidelines for sleep medicine in the EU.

This project provides our ACSA team with access to the ESADA database, i.e., the largest and most complete database of apnea patients [214]. ESADA allows for the development of complex retrospective studies on apnea patients, like kidney diseases [197] or cPAP treatment response. Our recent study [44] aimed at analyzing neck circumference as an indicator of CPAP treatment response in OSA can benefit from access to the ESADA database.

In [44] we explore the relationship between OSA, patients' anthropometric measures, and the CPAP treatment response. We use a cohort including 145 subjects with a one-night CPAP therapy, and create a CPAP-response network of patients to find neck circumference NC as the most significant qualitative indicator for apnea-hypopnea index (AHI) improvement. We confirm the correlation between NC and AHI ($\rho = 0.35$, $p < 0.001$) and show that 71% of diagnosed male subjects have a bigger NC than subjects with no OSA (area under the curve is 0.71, with 95% CI 0.63–0.79, $p < 0.001$); the optimal NC cutoff is 41 cm, with a sensitivity of 0.8099, a specificity of 0.5185. Our NC = 41 cm threshold classifies patients' CPAP responses—measured as the difference in AHI prior and after the one-night use of CPAP—with a sensitivity of 0.913 and a specificity of 0.859.

Figure 7.4 offers a flowchart representation comprising the usage of all three patient datasets $D_1 - D_3$, and the design of our retrospective study.

The CPAP patient network ($n = 145$ patients) is depicted in the center of Fig 7.5, where we show the network with its distinctly colored communities (i.e., purple, olive, orange, cyan), and around them, we present how each of the six measured criteria is associated with each cluster. Excepting the age group, all the other five measurements consistently associate with specific communities. We recorded AHI for all the patients in D_1 before and after the CPAP treatment to uncover a possible correlation between the four obtained communities and the effect of the CPAP treatment. We measured AHI before and after the one-night treatment. The sizes of each community are: $C_1 = 55$ patients, $C_2 = 32$ patients, $C_3 = 29$ patients and $C_4 = 29$ patients.

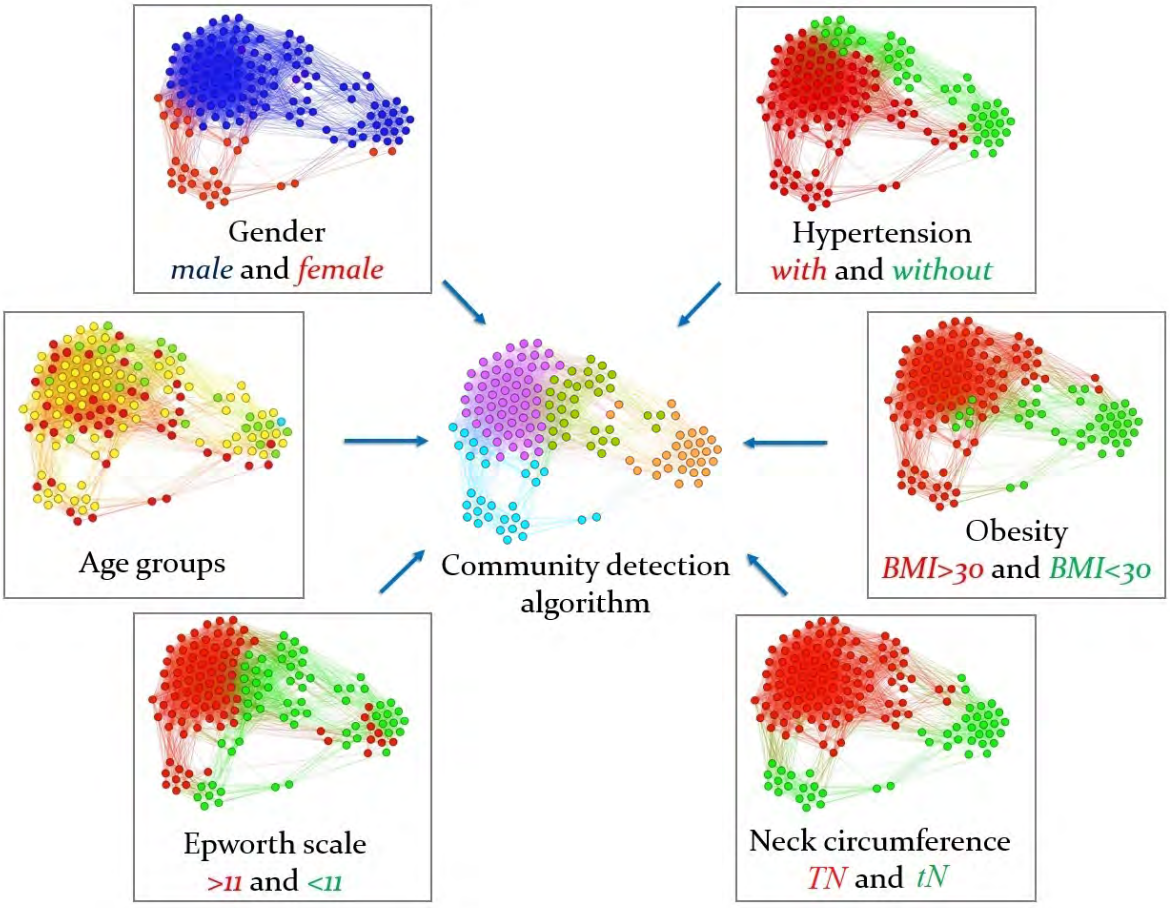

Figure 7.5: The network of 145 patients with overnight CPAP treatment shows the mapping of the six measurements (age, gender, blood pressure, BMI, Epworth scale, and neck circumference) that are relevant for the four patient communities detected for OSA (central panel). The four communities (purple, olive, orange, cyan) emerge from the modeled risk compatibility between patients and are used to study the association between patient risk factors and CPAP treatment response.

Our research group reports previous studies using network science to identify subgroups

(phenotypes) of patients with OSA [42, 41, 298]. However, in this retrospective study, we focus on using network analysis explicitly to analyze patients' response to CPAP treatment. To this end, D_1 ($n = 145$ patients) is the core dataset of our analysis, as it is the cohort of patients with one night CPAP treatment.

Our study emphasized the classification ability of neck circumference for CPAP responsiveness, in a population cohort of people referred to sleep labs for OSA evaluation and treatment [44]. The network analysis discovers NC as the best marker correlating with CPAP treatment, and our statistical analysis confirms a certain NC threshold for reliable treatment and prioritization. Moreover, we find that male patients with NC ≥ 41 cm should have a higher priority for the overnight sleep study and treatment. Measures of OSA severity, such as AHI alone, appear more weakly associated with CPAP adherence [299].

In line with this study, a future development for our statistical analysis could be to describe the optimal OSA risk thresholds that optimize trade-offs between true positives, true negatives, false positives and false negatives, through the use of a total cost function [300]. Also, we could define a complementary patient network leading to new insights, based on an alternative inference method which consists in the identification of a significant maximum mutual information (MI) network [301]; in this case, two patients are connected with each other if their shared MI value is maximal with respect to all other patients for at least for one of the two patients. Finally, replication of thus study on the much larger ESADA cohort remains open for prospective PhD students.

Chapter 8

Teaching Perspectives

In this section we discuss current and future courses and practical applications (laboratories, projects) being taught by the thesis candidate, and available for the career development of future doctoral students. We also mention several related courses from other ACSA members that may fall in line with the research topic of Master's and PhD students enrolled in our department. Finally, we detail an original gamification platform used for the in-class motivation of students.

8.1 Courses and Practical Applications

The thesis candidate currently teaches several distinct courses on varied topics in the field of Computers and Information Technology, such as Computer Engineering (3rd year, CTI Romanian section), Mobile Systems and Applications (4th year, CTI English section), and Big Data Visualization (1st year, Master of Machine Learning and Artificial Intelligence).

The Computer Engineering Course focuses on the idea of large scale communication. As such, we outline current trends in CPU and GPU development, which, limited by Moore's Law, have turned towards increasing the number cores and optimizing inter-process communication. To this end, we mention Networks-on-Chip and Systems-on-Chip as a motivation to study large scale, heterogeneous complex systems. The course is a mixture of engineering and network science aimed at offering Bachelor students an objective overview of the inter-dependency between computation, data and communication. Paradigms such as complex networks, topologies, computational social networks, Big Data, processes on networks are introduced. The laboratory of Computer Engineering is very bidding for PhD students who would like to start their teaching career in parallel to their research path. Over the past years, we had a constant number of four student sub-groups which could be managed wholly by one student.

Mobile Systems and Applications introduces students to the world of mobile operating systems and programming on mobile platforms. Specifically, we detail the Android OS architecture, and programming paradigms, such as activities, intents, activity lifecycle, task stack, services, broadcast receivers, and content providers. The laboratory has three student sub-groups and is dedicated towards developing tasks and projects of medium complexity. For proficient and motivated students, we are organizing the annual mobile applications de-

velopment student competition in UPT [1]. After nine organized editions (2013–2021), we are confident that students in the 4th years are capable of producing impressive results, worth presenting to local industry specialists (forming the panel of juries), to local newspapers (Renasterea Banateana), local Radio, and television (TV Politehnica).

Furthermore, the candidate has published two books in the area of mobile development. The first is entitled "Introducere in Programarea Android", authored together with Prof. Marius Marcu, and presents an overview of Android OS programming using Eclipse. The second book is entitled "Hands-On Android Application Development with Google Firebase", and combines Android OS specific programming with the Google Firebase platform for creating robust applications with a backend-as-a-service. This book is a welcome source of knowledge and practical guiding on specific concepts, application design aspects and programming examples for mobile platforms. It is a necessary and very useful material which addresses mainly the students in Computer and Information Technology area, but also all the professionals and enthusiasts interested in mobile application development and programming. Through its detailed content, with a lot of information and practical code examples, the book is addressed to all those who are passionate about programming mobile applications in Android and the Google Firebase platform, which have steadily become dominant players in the market of mobile apps. For future engineers, the book is important in that it offers them new skills, baked up by code examples, for integrating a Backend as a Service into their solutions.

The Big Data Visualization (BDV) course was newly introduced at our new Master on Machine Learning and Artificial Intelligence. This course is well suited for prospective doctoral students as well, since it introduces several key aspects of network science, as well as creating meaningful statistics, all supported by powerful visualizations. The BDV course was entered in the prestigious ANIS Scholarships Program for 2020, and has won a scholarship for the best new course on Big Data [2].

We consider BDV a timely course in the current data-driven and computationally-driven engineering context. As such, data visualization is a useful tool for analyzing both small-scale and large-scale data. One of the main skills of a data scientist (current or future) is the ability to create a story from available data. This process often involves viewing data and discoveries in an affordable and stimulating way. The current course starts from a series of new specializations that have appeared at major universities / companies in the world. Examples include Data Science: Visualization at Harvard, Data Visualization at the University of Illinois at Urbana-Champaign, Fundamentals of Visualization with Tableau at the University of California, Introduction to Data Science in Python at the University of Michigan, Applied Plotting, Charting & Data Representation in Python at the University of Michigan, Data Visualization with Python held by IBM. The BDV course at the Polytechnic University of Timișoara is addressed to students who want to discover what data visualization means, how it can be used to better understand data, and what are the steps for applying techniques on large data sets. Using state-of-the-art technologies such as Gephi, Cytospace, R studio, Excel, Plotly, Matplotlib, Folium, Jupyter Notebook, the basic concepts of data visualization will be examined and different tools will be applied to large data, with a strong emphasis on examples. from Network Science and Network Medicine. A graduate of this course will be able to prepare, import, process, view and analyze data, as well as explain the depen-

[1] SCMUPT 2021 - `https://sites.google.com/site/alexandrutopirceanu/projects/scmupt2021`

[2] Success story of the BDV course: `https://anis.ro/povesti_succes/alexandru-topirceanu/`

dence between data analysis and data visualization. This course provides the opportunity to learn skills and get involved in methods for discovering patterns on Big Data, and modeling them in the form of complex network models. The competencies acquired in the BDV course can be listed as follows: (i) assimilation of the necessary principles for translating Big Data into meaningful visualizations; (ii) practical experience in using several state-of-the-art tools for data modeling; (iii) ability to prepare, import, process and analyze large data sets; (iv) pattern discovery skills and modeling based on visual analysis.

While the BDV course will get feedback from its first generations of students, it is planned to integrate several elements such as Open Educational Resources (OER) and MOOC courses, such as some offered by Coursera or edX. In addition, if the Big Data specialization from the MLAI master becomes very appreciated, it can be considered the creation of a MOOC type course together with the other two holders of the Big Data series courses. Also, over time, the candidate can integrate his innovative learning motivation platform using gamification, suggestively titled - Gamified. The "experienced (XP) approach gained by students, instead of grades" has yielded notable results in undergraduate subjects [49]. At the same time, the candidate can apply, from his scientific activity, an original methodology of classification (archetyping) [48] of the students enrolled in the course in order to offer them a personalized way of learning.

Other courses in the ACSA group from which doctoral students can benefit while undergoing a PhD in network science, are: Big Data in Health and Bioinformatics, Big Data in Cloud and IoT.

Two additional courses are planned in the upcoming 2–3 years, namely Complex Networks and Applications and Design and Analysis (RCA) of Mobile Applications (PAAM), both during the 4th year, at the CTI Romanian section. With RCA, Bachelor students will have hands-on contact with the field of network science, thus leading to increased popularity of our master's program on MLAI and, possibly, an increased desire to follow a PhD program within the ACSA group.

In conclusion, the development of the academic process through blended-learning creates challenges for teachers, such as increasing the complexity of teaching and motivation, the difficulty of selecting appropriate MOOCs for the discipline in question, and the evaluation of student activity. In the long run, however, these efforts bring several benefits and satisfactions, which can be demonstrated by increasing students' interest in the discipline and their appreciation of education.

8.2 Gamification for Student Motivation

A modern and successful tackle on education is represented by new teaching techniques which imply online courses, collaborative assignments, dynamic grading systems, real-time feedback and motivational inserts into the process of learning. E-learning together with massive open online courses (MOOCs) have seen a recent rise in popularity and integrate many of the aspects that enable distant students to take part in higher levels of education. While the perspective of migrating towards a pure online environment is in line with the trend of the younger generations, most professional and intellectual skills can only be effectively learned through physical attendance and practical guided work.

The past 5 years, alongside the current pandemic situation, have underlined the weak point of most classic educational systems: the constant decreasing motivation it gives students - individuals who have grown and are embedded in many virtual realities form where they draw the needed intrinsic motivation and energy. To overcome this limitation, we introduced an educational platform named *Gamified*, which simplifies the educational and grading systems in modern schools and universities [49]. It relies on the fundamental aspects of the theory of Gamification, namely bringing motivational elements from (video) games into non-game contexts. It does this through the abandonment of grades (seen by us as negative feedback, a demotivator), and integration of heroes, accumulated experience, levels, level-ups, achievements, quests, guilds, and other representative elements taken from role-playing games (positive feedback at different levels of motivation). Not only do these elements sound familiar to a majority of today's students, but they also trigger interest in the new approach to learning. We validated this technique (over a period of 3 years) on different generations of college students [49], compared the results with control groups, and obtained consistent feedback - both in terms of grades and participation, as well as in student attitude towards learning.

This project was started back in 2013 with the belief that gamification can foster the appearance of a new avant-garde teaching system which could rise the intrinsic drive to learn, so that students and educators may benefit from it. In [49] we provide a detailed snapshot of the Gamified project, and the obtained results prove the impact of our proposal, which is further backed up by feedback offered at the end of each semester by participants. Nevertheless, technology, in this context, is presented not as the indispensable drive of education, but merely as facilitator for the necessary visual cues and automated computation; the educator remains in our view the true drive of meaningful education. He only has to enrich his techniques with the use of custom motivators with whom young people emphasize, namely game elements in an educational context, without sacrificing any of the academic context. A comparison to the student control groups, which relied on classic grading schemes, shows that, in the gamified groups, all metrics are in favor of the more modern approach. For instance, we obtain an overall attendance boost from 50-72% to 77-93%, the percentage of students with full attendance rises from 6-12% to $> 50\%$, and the amount of high marks is increased by a factor of roughly 4-8 times [49].

With the introduction of our original gamified platform, we hope to foster both research in the areas of educational science, and data mining from student social networks [48], as well as motivate other readers to adopt game elements in their educational practice. By adopting our platform, we believe that the goal of educators will shift from making the young just realize they *have to learn*, or accept they *must learn*, and transcend to making the young incapable of quenching their thirst for knowledge, and so, making them teach others in their turn.

Chapter 9

General Conclusions

This thesis discusses the potential of Computer Science and Engineering interleaved with Network Science to solve relevant open research questions of the 21^{st} century. Network Science is interdisciplinary by definition, as it stems from Computer Science and Engineering, Mathematics and Physics, with vast applicability in Technological, Biological, Social, Political, and Economic sciences altogether.

One of the most important aspect underlined by this thesis is that thethe methodologies presented here, involving large amounts of data, modeled as complex systems, are supported by Computer Engineering & Information Technology. To this end, we presented approaches where computer algorithms, genetic algorithms, simulation tools and databases are developed for the processing and understanding of social network, medical, epidemiological, pharmacological, political, and educational data. Advancements in Communication technologies support today's large online social networks, and further motivate research in social physics and computational epidemics with global impact. Computer-based technologies, such as Machine Learning, Big Data Analytics, and Complex Network Analysis are used in recent developments of personalized decision making systems, e.g., in education and medicine.

This present thesis shows how Network Science improves our understanding of: (i) network growth using the novel concept of Betweenness Preferential Attachment, (ii) the antifragile response of large network-based complex systems under structural attacks, (iii) benchmarking centrality measures in a competitive context, (iv) micro- an macro-scale opinion dynamics and improvement of opinion distribution forecasting, (v) diagnosis of obstructive sleep apnea, (vi) drug-drug interactions and repurposing, (vii) dynamics of epidemics using heterogeneous population and mobility models, and (viii) student archetyping in online education. Overall, the thesis is divided in two parts: contributions and career development. The first part addresses the most important research challenges tackled over the 2011-2021 period. These include contributions in social networks analysis, computational network analysis and network medicine. The second part enumerates the career evolution and future research plans, the capability of attracting and leading new research projects, and underlines the infrastructure, financial, research and professional opportunities of prospective PhD students.

Given our experience in coordinating PhD students over a broad range of scientific topics, we consider our ACSA (Advanced Computing Systems and Architectures) research group as an attractive opportunity for PhD programs. With the newest inclusion of Network Science in our research portfolio (since 2011), we are able to offer a completely developed PhD program

focusing on Network Science in the field of Computers and Information Technology.

The current computing infrastructure, as published on the ERRIS platform, supports research on modeling and simulations, network science, big data, graph algorithms, and data mining, all of which are directly contributing research fields to our group. Furthermore, the infrastructures offered by the ACSA laboratory, the Department's Vision NextCloud platform, the University Virtual Campus, and the future available CloudPUTIng high performance computing platform will offer PhD students more than enough support for a diverse teaching and research career.

Our scientific and academic results are summarized by the management of 2 national research projects (financed by UEFISCDI), membership in an additional 2 international projects (financed by Linde and Horizon 2020), and 5 national projects (financed by UEFISCDI and ARUT), publication of 2 books, over 50 Web of Science indexed papers, out of which 16 journals (12 indexed in Q1/Q2), a cumulative impact factor over 45, a WoS h-index of 9 and 171 citations (330 citation in Google Scholar), review in diverse multidisciplinary and IEEE journals, organization of the 9 editions (2013–2021) of the SCMUPT student competition for mobile development, coordination as member in one PhD committee, and member of multiple PhD report committees, and coordination of over 90 Bachelor and Masters theses.

In conclusion, this thesis serves as a strong proof for the high impact research that can be achieved by employing Computer Science and Engineering in cross-disciplinary fields. As such, we intend to further narrow the gap between Computers and Network Science by further tackling challenging research topics from diverse fields of science, by participating at major conference venues in our field, by establishing long-lasting international collaboration, by creating project partnerships, by initiating new specialized undergraduate courses in our department, and by integrating new doctoral students in our ACSA research group.

Part III

Relevant Bibliography

Bibliography

[1] P. Erd6s and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hungar. Acad. Sci*, vol. 5, pp. 17–61, 1960.

[2] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[3] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[4] A.-L. Barabási, "Network science," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1987, p. 20120375, 2013.

[5] A.-L. Barabási and M. Pósfai, *Network science*. Cam. Univ. press, 2016.

[6] A. Vespignani, "Complex networks: The fragility of interdependency," *Nature*, vol. 464, no. 7291, p. 984, 2010.

[7] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Reviews of modern physics*, vol. 87, no. 3, p. 925, 2015.

[8] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

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

[10] X. F. Wang and G. Chen, "Complex networks: small-world, scale-free and beyond," *IEEE circuits and systems*, vol. 3, pp. 6–20, 2003.

[11] 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.

[12] Z. Ruan, G. Iniguez, M. Karsai, and J. Kertész, "Kinetics of social contagion," *Physical review letters*, vol. 115, no. 21, p. 218702, 2015.

[13] J. Golbeck, *Analyzing the Social Web*. Access Online via Elsevier, 2013.

[14] S. González-Bailón, J. Borge-Holthoefer, and Y. Moreno, "Broadcasters and hidden influentials in online protest diffusion," *American Behavioral Scientist*, vol. 57, no. 7, pp. 943–965, 2013.

[15] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *nature*, vol. 406, no. 6794, p. 378, 2000.

[16] G. A. Pagani and M. Aiello, "The power grid as a complex network: a survey," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 11, pp. 2688–2700, 2013.

[17] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[18] 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*, pp. 298–307, ACM, 2012.

[19] W. Pan and C. Chai, "Measuring software stability based on complex networks in software," *Cluster Computing*, vol. 22, no. 2, pp. 2589–2598, 2019.

[20] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Social science. computational social science.," *Science (New York, NY)*, vol. 323, no. 5915, pp. 721–723, 2009.

[21] A.-L. Barabási, "Network medicine—from obesity to the "diseasome"," *New England Journal of Medicine*, vol. 357, no. 4, pp. 404–407, 2007.

[22] 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)*, vol. 323, no. 5915, p. 721, 2009.

[23] H. Hexmoor, *Computational network science: an algorithmic approach*. Morgan Kaufmann, 2014.

[24] P. Kazienko, "Computational network science: From data to social models," in *2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*, pp. 1–1, IEEE, 2017.

[25] A. Topirceanu, M. Udrescu, and M. Vladutiu, "Genetically optimized realistic social network topology inspired by facebook," in *Online Social Media Analysis and Visualization*, pp. 163–179, Springer, 2014.

[26] A.-L. Barabási, *Linked: The New Science Of Networks*. Basic Books, 2002.

[27] A. Topirceanu, M. Udrescu, and R. Marculescu, "Weighted betweenness preferential attachment: A new mechanism explaining social network formation and evolution," *Scientific reports*, vol. 8, no. 1, pp. 1–14, 2018.

[28] R. A. Holley and T. M. Liggett, "Ergodic theorems for weakly interacting infinite systems and the voter model," *The annals of probability*, pp. 643–663, 1975.

[29] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[30] 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*, vol. 12, no. 3, pp. 211–223, 2001.

[31] D. Acemoglu and A. Ozdaglar, "Opinion dynamics and learning in social networks," *Dynamic Games and Applications*, vol. 1, no. 1, pp. 3–49, 2011.

[32] D. Acemoğlu, G. Como, F. Fagnani, and A. Ozdaglar, "Opinion fluctuations and disagreement in social networks," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 1–27, 2013.

[33] 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*, vol. 4, 2014.

[34] 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*, vol. 2, p. e42, 2016.

[35] R. Hegselmann and U. Krause, "Opinion dynamics and bounded confidence models, analysis, and simulation," *Journal of Artificial Societies and Social Simulation*, vol. 5, no. 3, 2002.

[36] W. Weidlich, "Sociodynamics—-a systematic approach to mathematical modelling in the social sciences," *Chaos, Solitons & Fractals*, vol. 18, no. 3, pp. 431–437, 2003.

[37] 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*, vol. 94, no. 9, pp. 1787–1805, 2017.

[38] A. Topîrceanu, M. Udrescu, and R. Mărculescu, "Complex networks antifragility under sustained edge attack-repair mechanisms," in *International Conference on Network Science*, pp. 185–199, Springer, 2020.

[39] A. Topîrceanu, "Competition-based benchmarking of influence ranking methods in social networks," *Complexity*, vol. 2018, 2018.

[40] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature reviews genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[41] S. Mihaicuta, M. Udrescu, A. Topirceanu, and L. Udrescu, "Network science meets respiratory medicine for osas phenotyping and severity prediction," *PeerJ*, vol. 5, p. e3289, 2017.

[42] A. Topîrceanu, M. Udrescu, L. Udrescu, C. Ardelean, R. Dan, D. Reisz, and S. Mihaicuta, "Sas score: Targeting high-specificity for efficient population-wide monitoring of obstructive sleep apnea," *PloS one*, vol. 13, no. 9, p. e0202042, 2018.

[43] A. Topîrceanu, L. Udrescu, M. Udrescu, and S. Mihaicuta, "Gender phenotyping of patients with obstructive sleep apnea syndrome using a network science approach," *Journal of Clinical Medicine*, vol. 9, no. 12, p. 4025, 2020.

[44] S. Mihaicuta, L. Udrescu, M. Udrescu, I.-A. Toth, A. Topîrceanu, R. Pleavă, and C. Ardelean, "Analyzing neck circumference as an indicator of cpap treatment response in obstructive sleep apnea with network medicine," *Diagnostics*, vol. 11, no. 1, p. 86, 2021.

[45] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D668–D672, 2006.

[46] L. Udrescu, L. Sbârcea, A. Topîrceanu, A. Iovanovici, L. Kurunczi, P. Bogdan, and M. Udrescu, "Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing," *Scientific Reports*, vol. 6, 2016.

[47] L. Udrescu, P. Bogdan, A. Chiş, I. O. Sîrbu, A. Topîrceanu, R.-M. Văruţ, and M. Udrescu, "Uncovering new drug properties in target-based drug–drug similarity networks," *Pharmaceutics*, vol. 12, no. 9, p. 879, 2020.

[48] A. Topîrceanu and G. Grosseck, "Decision tree learning used for the classification of student archetypes in online courses," *Procedia Computer Science*, vol. 112, pp. 51–60, 2017.

[49] A. Topîrceanu, "Gamified learning: A role-playing approach to increase student in-class motivation," *Procedia computer science*, vol. 112, pp. 41–50, 2017.

[50] A. Topîrceanu, "Breaking up friendships in exams: A case study for minimizing student cheating in higher education using social network analysis," *Computers & Education*, vol. 115, pp. 171–187, 2017.

[51] A. Topirceanu, M. Udrescu, and R. Marculescu, "Centralized and decentralized isolation strategies and their impact on the covid-19 pandemic dynamics," *arXiv preprint arXiv:2004.04222*, 2020.

[52] A. Topîrceanu, "Analyzing the impact of geo-spatial organization of real-world communities on epidemic spreading dynamics," in *International Conference on Complex Networks and Their Applications*, pp. 345–356, Springer, 2020.

[53] A. Topîrceanu and R.-E. Precup, "A novel methodology for improving election poll prediction using time-aware polling," in *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pp. 282–285, 2019.

[54] A. Topirceanu, "Electoral forecasting using a novel temporal attenuation model: Predicting the us presidential elections," *arXiv preprint arXiv:2005.01799*, 2020.

[55] 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*, pp. 289–296, IEEE, 2013.

[56] A. Topirceanu, A. Duma, and M. Udrescu, "Uncovering the fingerprint of online social networks using a network motif based approach," *Computer Communications*, vol. 73, pp. 167–175, 2016.

[57] A. Topirceanu, G. Barina, and M. Udrescu, "Musenet: Collaboration in the music artists industry," in *Network Intelligence Conference (ENIC), 2014 European*, pp. 89–94, IEEE, 2014.

[58] A. Topirceanu and M. Udrescu, "Fmnet: Physical trait patterns in the fashion world," in *2015 Second European Network Intelligence Conference*, pp. 25–32, IEEE, 2015.

[59] M. Udrescu and A. Topirceanu, "Probabilistic modeling of tolerance-based social network interaction," in *Network Intelligence Conference (ENIC), 2016 Third European*, pp. 48–54, IEEE, 2016.

[60] A. Topîrceanu and R.-E. Precup, "A framework for improving electoral forecasting based on time-aware polling," *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 1–14, 2020.

[61] E. Estrada, *The structure of complex networks: theory and applications*. Oxford University Press, 2012.

[62] M. E. Newman, A.-L. E. Barabási, and D. J. Watts, *The structure and dynamics of networks*. Princeton university press, 2006.

[63] M. Vidal, M. E. Cusick, and A.-L. Barabasi, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986–998, 2011.

[64] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[65] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[66] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[67] M. E. Newman, "A measure of betweenness centrality based on random walks," *Social networks*, vol. 27, no. 1, pp. 39–54, 2005.

[68] M. E. Newman, "The mathematics of networks," *The new palgrave encyclopedia of economics*, vol. 2, no. 2008, pp. 1–12, 2008.

[69] A. Noack, "Modularity clustering is force-directed layout," *Physical Review E*, vol. 79, no. 2, p. 026102, 2009.

[70] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the national academy of sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[71] 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*, vol. 2008, no. 10, p. P10008, 2008.

[72] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[73] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On modularity clustering," *IEEE transactions on knowledge and data engineering*, vol. 20, no. 2, pp. 172–188, 2007.

[74] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," *PloS one*, vol. 9, no. 6, p. e98679, 2014.

[75] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks.," in *ICWSM*, 2009.

[76] A. D'Andrea, F. Ferri, and P. Grifoni, "An overview of methods for virtual social networks analysis," *Computational social network analysis*, pp. 3–25, 2010.

[77] S. Wasserman, *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.

[78] A. Topîrceanu and M. Udrescu, "Strength of nations: A case study on estimating the influence of leading countries using social media analysis," in *European Network Intelligence Conference*, pp. 219–229, Springer, 2017.

[79] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, no. 1, pp. 1–122, 2014.

[80] M. E. Dickison, M. Magnani, and L. Rossi, *Multilayer social networks*. Cambridge University Press, 2016.

[81] M. Newman, *Networks: an introduction*. Oxford University Press, 2010.

[82] C. Alt, O. Astrachan, J. Forbes, R. Lucic, and S. Rodger, "Social networks generate interest in computer science," *ACM SIGCSE Bulletin*, vol. 38, no. 1, pp. 438–442, 2006.

[83] D. Lusseau, "The emergent properties of a dolphin social network," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. Suppl 2, pp. S186–S188, 2003.

[84] G. Csányi and B. Szendrői, "Structure of a large social network," *Physical Review E*, vol. 69, no. 3, p. 036131, 2004.

[85] G. Pruessner and H. J. Jensen, "Broken scaling in the forest-fire model," *Physical Review E*, vol. 65, no. 5, p. 056707, 2002.

[86] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[87] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," *Physics Reports*, vol. 650, pp. 1–63, 2016.

[88] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica a: Statistical mechanics and its applications*, vol. 391, no. 4, pp. 1777–1787, 2012.

[89] Q. Li, T. Zhou, L. Lü, and D. Chen, "Identifying influential spreaders by weighted leaderrank," *Physica A: Statistical Mechanics and its Applications*, vol. 404, pp. 47–55, 2014.

[90] X. Zhao, F. Liu, J. Wang, T. Li, *et al.*, "Evaluating influential nodes in social networks by local centrality with a coefficient," *ISPRS International Journal of Geo-Information*, vol. 6, no. 2, p. 35, 2017.

[91] D.-B. Chen, H. Gao, L. Lü, and T. Zhou, "Identifying influential nodes in large-scale directed networks: the role of clustering," *PloS one*, vol. 8, no. 10, p. e77455, 2013.

[92] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature physics*, vol. 6, no. 11, p. 888, 2010.

[93] C. Castellano and R. Pastor-Satorras, "Competing activation mechanisms in epidemics on networks," *Scientific reports*, vol. 2, p. 371, 2012.

[94] J.-G. Liu, Z.-M. Ren, and Q. Guo, "Ranking the spreading influence in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 18, pp. 4154–4159, 2013.

[95] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences of the United States of America*, pp. 16569–16572, 2005.

[96] P. Bonacich, "Some unique properties of eigenvector centrality," *Social networks*, vol. 29, no. 4, pp. 555–564, 2007.

[97] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[98] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[99] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PloS one*, vol. 6, no. 6, p. e21202, 2011.

[100] L. Adamic, O. Buyukkokten, and E. Adar, "A social network caught in the web," *First monday*, vol. 8, no. 6, 2003.

[101] R. S. Burt, "Attachment, decay, and social network," *Journal of Organizational Behavior*, vol. 22, no. 6, pp. 619–643, 2001.

[102] 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*, vol. 6, no. 3, pp. 403–412, 2012.

[103] R. I. Dunbar, "Neocortex size as a constraint on group size in primates," *Journal of Human Evolution*, vol. 22, no. 6, pp. 469–493, 1992.

[104] M. E. Brashears, "Humans use compression heuristics to improve the recall of social networks," *Scientific reports*, vol. 3, 2013.

[105] D. Krackhardt, "The strength of strong ties: The importance of philos in organizations," *Networks and organizations: Structure, form, and action*, vol. 216, p. 239, 1992.

[106] L. Leydesdorff, "Betweenness centrality as an indicator of the interdisciplinarity of scientific journals," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 9, pp. 1303–1319, 2007.

[107] S. Plous, *The psychology of judgment and decision making.* Mcgraw-Hill Book Company, 1993.

[108] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444, 2001.

[109] S. Johnson, J. J. Torres, J. Marro, and M. A. Munoz, "Entropic origin of disassortativity in complex networks," *Physical review letters*, vol. 104, no. 10, p. 108702, 2010.

[110] D. Zhou, H. E. Stanley, G. DAgostino, and A. Scala, "Assortativity decreases the robustness of interdependent networks," *Physical Review E*, vol. 86, no. 6, p. 066103, 2012.

[111] N. N. Taleb, *Antifragile: how to live in a world we don't understand*, vol. 3. Allen Lane London, 2012.

[112] N. N. Taleb and R. Douady, "Mathematical definition, mapping, and detection of (anti) fragility," *Quantitative Finance*, vol. 13, no. 11, pp. 1677–1689, 2013.

[113] A. Topirceanu and M. Udrescu, "Topological fragility versus antifragility: understanding the impact of real-time repairs in networks under targeted attacks," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1215–1222, IEEE, 2018.

[114] M. Lichtman, M. T. Vondal, T. C. Clancy, and J. H. Reed, "Antifragile communications," *IEEE Systems Journal*, vol. 12, no. 1, pp. 659–670, 2016.

[115] S. W. Duxbury and D. L. Haynie, "Criminal network security: An agent-based approach to evaluating network resilience," *Criminology*, vol. 57, no. 2, pp. 314–342, 2019.

[116] S. He, S. Li, and H. Ma, "Effect of edge removal on topological and functional robustness of complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 11, pp. 2243–2253, 2009.

[117] W. Sun and A. Zeng, "Target recovery in complex networks," *The European Physical Journal B*, vol. 90, no. 1, p. 10, 2017.

[118] P. Crucitti, V. Latora, M. Marchiori, and A. Rapisarda, "Error and attack tolerance of complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 340, no. 1-3, pp. 388–394, 2004.

[119] S. Iyer, T. Killingback, B. Sundaram, and Z. Wang, "Attack robustness and centrality of complex networks," *PloS one*, vol. 8, p. e59613, 2013.

[120] A. Dekker, "Realistic social networks for simulation using network rewiring," in *International Congress on Modelling and Simulation*, pp. 677–683, 2007.

[121] M. Pósfai, Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Effect of correlations on network controllability," *Scientific reports*, vol. 3, no. 1, pp. 1–7, 2013.

[122] A. Zeng and C.-J. Zhang, "Ranking spreaders by decomposing complex networks," *Physics Letters A*, vol. 377, no. 14, pp. 1031–1035, 2013.

[123] J.-H. Lin, Q. Guo, W.-Z. Dong, L.-Y. Tang, and J.-G. Liu, "Identifying the node spreading influence with largest k-core values," *Physics Letters A*, vol. 378, no. 45, pp. 3279–3284, 2014.

[124] Y. Liu, M. Tang, T. Zhou, and Y. Do, "Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics," *Scientific Reports*, vol. 5, p. 13172, 2015.

[125] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, and H. A. Makse, "Searching for superspreaders of information in real-world social media," *Scientific Reports*, vol. 4, p. 5547, 2014.

[126] Z.-M. Ren, A. Zeng, D.-B. Chen, H. Liao, and J.-G. Liu, "Iterative resource allocation for ranking spreaders in complex networks," *EPL (Europhysics Letters)*, vol. 106, no. 4, p. 48005, 2014.

[127] E. Estrada and N. Hatano, "A vibrational approach to node centrality and vulnerability in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 17, pp. 3648–3660, 2010.

[128] M. E. Newman, "Spread of epidemic disease on networks," *Physical review E*, vol. 66, no. 1, p. 016128, 2002.

[129] 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*, vol. 35, no. 4, pp. 583–592, 2013.

[130] 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.

[131] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical review letters*, vol. 86, no. 14, p. 3200, 2001.

[132] A. Fonseca, "Modeling political opinion dynamics through social media and multi-agent simulation," in *First Doctoral Workshop for Complexity Sciences*, 2011.

[133] 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*, pp. 137–146, ACM, 2003.

[134] 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*, pp. 1–8, 2013.

[135] M. McDonald and H. Wilson, *Marketing plans: How to prepare them, how to use them.* Wiley. com, 2011.

[136] R. Axelrod, "The dissemination of culture: A model with local convergence and global polarization," *Journal of conflict resolution*, vol. 41, no. 2, pp. 203–226, 1997.

[137] E. Yildiz, A. Ozdaglar, D. Acemoglu, A. Saberi, and A. Scaglione, "Binary opinion dynamics with stubborn agents," *ACM Transactions on Economics and Computation (TEAC)*, vol. 1, no. 4, p. 19, 2013.

[138] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM Sigmod Record*, vol. 42, no. 2, pp. 17–28, 2013.

[139] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," *Advances in Complex Systems*, vol. 3, no. 01n04, pp. 87–98, 2000.

[140] 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*, pp. 805–812, Citeseer, 2013.

[141] L. Deng, Y. Liu, and F. Xiong, "An opinion diffusion model with clustered early adopters," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 17, pp. 3546–3554, 2013.

[142] J. Cannarella and J. A. Spechler, "Epidemiological modeling of online social network dynamics," *arXiv preprint arXiv:1401.4208*, 2014.

[143] B. Karrer and M. E. Newman, "Competing epidemics on complex networks," *Physical Review E*, vol. 84, no. 3, p. 036106, 2011.

[144] A.-L. Barabási, "Scale-free networks: a decade and beyond," *science*, vol. 325, no. 5939, pp. 412–413, 2009.

[145] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 4, p. 21, 2012.

[146] A. Duma and A. Topirceanu, "A network motif based approach for classifying online social networks," in *Applied computational intelligence and informatics (SACI)*, pp. 311–315, IEEE, 2014.

[147] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, p. 5, 2007.

[148] H. T. Nguyen, T. N. Dinh, and M. T. Thai, "Cost-aware targeted viral marketing in billion-scale networks," in *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9, IEEE, 2016.

[149] A. Feldman, C. Konold, B. Coulter, and B. Conroy, *Network science, a decade later: The Internet and classroom learning*. Routledge, 2000.

[150] M. O. Jackson, B. W. Rogers, and Y. Zenou, "The economic consequences of social-network structure," *Journal of Economic Literature*, vol. 55, no. 1, pp. 49–95, 2017.

[151] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," in *Proceedings of the 21st international conference on World Wide Web*, pp. 1145–1152, ACM, 2012.

[152] J. J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks.," in *NIPS*, vol. 2012, pp. 548–56, 2012.

[153] J. Golbeck and D. Hansen, "A method for computing political preference among twitter followers," *Social Networks*, 2013.

[154] M. M. Hussain and P. N. Howard, "What best explains successful protest cascades? icts and the fuzzy causes of the arab spring," *International Studies Review*, vol. 15, no. 1, pp. 48–66, 2013.

[155] A. L. Hughes and L. Palen, "Twitter adoption and use in mass convergence and emergency events," *International Journal of Emergency Management*, vol. 6, no. 3-4, pp. 248–260, 2009.

[156] B. A. Conway, K. Kenski, and D. Wang, "Twitter use by presidential primary candidates during the 2012 campaign," *American Behavioral Scientist*, vol. 57, no. 11, pp. 1596–1610, 2013.

[157] B. Heredia, J. D. Prusa, and T. M. Khoshgoftaar, "Social media for polling and predicting united states election outcome," *Social Network Analysis and Mining*, vol. 8, no. 1, p. 48, 2018.

[158] L. Hufnagel, D. Brockmann, and T. Geisel, "Forecast and control of epidemics in a globalized world," *Proceedings of the National Academy of Sciences*, vol. 101, no. 42, pp. 15124–15129, 2004.

[159] H. Gladwin, J. K. Lazo, B. H. Morrow, W. G. Peacock, and H. E. Willoughby, "Social science research needs for the hurricane forecast and warning system," *Natural Hazards Review*, vol. 8, no. 3, pp. 87–95, 2007.

[160] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 492–499, IEEE Computer Society, 2010.

[161] R. Nadeau, M. S. Lewis-Beck, and É. Bélanger, "Electoral forecasting in france: A multi-equation solution," *International Journal of Forecasting*, vol. 26, no. 1, pp. 11–18, 2010.

[162] P. Whiteley, "Electoral forecasting from poll data: the british case," *British Journal of Political Science*, vol. 9, no. 2, pp. 219–236, 1979.

[163] G. Weimann, "The obsession to forecast: Pre-election polls in the israeli press," *Public Opinion Quarterly*, vol. 54, no. 3, pp. 396–408, 1990.

[164] J. Wallinga and P. Teunis, "Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures," *American Journal of epidemiology*, vol. 160, no. 6, pp. 509–516, 2004.

[165] S. Myers and J. Leskovec, "On the convexity of latent social network inference," in *Advances in neural information processing systems*, pp. 1741–1749, 2010.

[166] W. F. Christensen and L. W. Florence, "Predicting presidential and other multistage election outcomes using state-level pre-election polls," *The American Statistician*, vol. 62, no. 1, pp. 1–10, 2008.

[167] C. P. Kiewiet de Jonge, G. Langer, and S. Sinozich, "Predicting state presidential election results using national tracking polls and multilevel regression with poststratification (mrp)," *Public Opinion Quarterly*, vol. 82, no. 3, pp. 419–446, 2018.

[168] W. Wang, D. Rothschild, S. Goel, and A. Gelman, "Forecasting elections with non-representative polls," *International Journal of Forecasting*, vol. 31, no. 3, pp. 980–991, 2015.

[169] P. Hummel and D. Rothschild, "Fundamental models for forecasting elections at the state level," *Electoral Studies*, vol. 35, pp. 123–139, 2014.

[170] J. Mellon and C. Prosser, "Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users," *Research & Politics*, vol. 4, no. 3, p. 2053168017720008, 2017.

[171] D. A. Graber and J. Dunaway, *Mass media and American politics*. Cq Press, 2017.

[172] M. Coppedge, J. Gerring, C. H. Knutsen, S. I. Lindberg, J. Teorell, D. Altman, M. Bernhard, M. S. Fish, A. Glynn, A. Hicken, *et al.*, "V-dem codebook v9," 2019.

[173] S. Y. Chan and J. Loscalzo, "The emerging paradigm of network medicine in the study of human disease," *Circulation research*, vol. 111, no. 3, pp. 359–374, 2012.

[174] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.

[175] A. R. Sonawane, S. T. Weiss, K. Glass, and A. Sharma, "Network medicine in the age of biomedical big data," *Frontiers in Genetics*, vol. 10, p. 294, 2019.

[176] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, 2015.

[177] A. Sharma, N. Gulbahce, S. J. Pevzner, J. Menche, C. Ladenvall, L. Folkersen, P. Eriksson, M. Orho-Melander, and A.-L. Barabási, "Network-based analysis of genome wide association data provides novel candidate genes for lipid and lipoprotein traits," *Molecular & Cellular Proteomics*, vol. 12, no. 11, pp. 3398–3408, 2013.

[178] O. Rozenblatt-Rosen, R. C. Deo, M. Padi, G. Adelmant, M. A. Calderwood, T. Rolland, M. Grace, A. Dricot, M. Askenazi, M. Tavares, S. J. Pevzner, F. Abderazzaq, D. Byrdsong, A. R. Carvunis, A. A. Chen, J. Cheng, M. Correll, M. Duarte, C. Fan, M. C. Feltkamp, S. B. Ficarro, R. Franchi, B. K. Garg, N. Gulbahce, T. Hao, A. M. Holthaus, R. James, A. Korkhin, L. Litovchick, J. C. Mar, T. R. Pak, S. Rabello, R. Rubio, Y. Shen, S. Singh, J. M. Spangle, M. Tasan, S. Wanamaker, J. T. Webber, J. Roecklein-Canfield, E. Johannsen, A. L. Barabási, R. Beroukhim, E. Kieff, M. E. Cusick, D. E. Hill, K. Münger, J. A. Marto, J. Quackenbush, F. P. Roth, J. A. DeCaprio, and M. Vidal, "Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins," *Nature*, vol. 487, no. 7408, pp. 491–495, 2012.

[179] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabasi, and M. Vidal, "Drug–target network," *Nature biotechnology*, vol. 25, no. 10, p. 1119, 2007.

[180] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[181] J.-D. J. Han, "Understanding biological functions through molecular networks," *Cell Research*, vol. 18, no. 2, pp. 224–237, 2008.

[182] R. Faner, T. Cruz, A. López-Giraldo, and A. Agustí, "Network medicine, multimorbidity and the lung in the elderly," *European Respiratory Journal*, vol. 44, no. 3, pp. 775–788, 2014.

[183] M. J. Divo, C. Casanova, J. M. Marin, V. M. Pinto-Plata, J. P. de Torres, J. J. Zulueta, C. Cabrera, J. Zagaceta, P. Sanchez-Salcedo, J. Berto, R. Baz Davila, A. B. Alcaide, C. Cote, and B. R. Celli, "Chronic obstructive pulmonary disease comorbidities network," *European Respiratory Journal*, vol. 46, no. 3, pp. 640–650, 2015.

[184] 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*, vol. 165, no. 9, pp. 1217–1239, 2002.

[185] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of sleep-disordered breathing in adults," *American Journal of Epidemiology*, vol. 177, no. 9, pp. 1006–1014, 2013.

[186] R. Heinzer, S. Vat, P. Marques-Vidal, H. Marti-Soler, D. Andries, N. Tobback, V. Mooser, M. Preisig, A. Malhotra, G. Waeber, *et al.*, "Prevalence of sleep-disordered breathing in the general population: the hypnolaus study," *The Lancet Respiratory Medicine*, vol. 3, no. 4, pp. 310–318, 2015.

[187] M. R. Bonsignore, M. C. S. Giron, O. Marrone, A. Castrogiovanni, and J. M. Montserrat, "Personalised medicine in sleep respiratory disorders: focus on obstructive sleep apnoea diagnosis and treatment," *European Respiratory Review*, vol. 26, no. 146, p. 170069, 2017.

[188] W. T. McNicholas, C. L. Bassetti, L. Ferini-Strambi, J. L. Pépin, D. Pevernagie, J. Verbraecken, W. Randerath, M. R. Bonsignore, R. Farre, L. Grote, *et al.*, "Challenges in obstructive sleep apnoea," *The Lancet Respiratory Medicine*, vol. 6, no. 3, pp. 170–172, 2018.

[189] W. Randerath, C. L. Bassetti, M. R. Bonsignore, R. Farre, L. Ferini-Strambi, L. Grote, J. Hedner, M. Kohler, M. Martinez-Garcia, S. Mihaicuta, *et al.*, "Challenges and perspectives in obstructive sleep apnoea: Report by an ad hoc working group of the sleep disordered breathing group of the european respiratory society and the european sleep research society.," *European Respiratory Journal*, p. 1702616, 2018.

[190] S. Simon and N. Collop, "Latest advances in sleep medicine: obstructive sleep apnea," *CHEST Journal*, vol. 142, no. 6, pp. 1645–1651, 2012.

[191] P. Lévy, M. Kohler, W. T. McNicholas, F. Barbé, R. D. McEvoy, V. K. Somers, L. Lavie, and J.-L. Pépin, "Obstructive sleep apnoea syndrome.," *Nature Reviews. Disease Primers*, vol. 1, pp. 15015–15015, 2014.

[192] D. F. Kripke, L. Garfinkel, D. L. Wingard, M. R. Klauber, and M. R. Marler, "Mortality associated with sleep duration and insomnia," *Archives of general psychiatry*, vol. 59, no. 2, pp. 131–136, 2002.

[193] J. S. Floras, "Sleep apnea and cardiovascular risk," *Journal of cardiology*, vol. 63, no. 1, pp. 3–8, 2014.

[194] F. J. Nieto, T. B. Young, B. K. Lind, E. Shahar, J. M. Samet, S. Redline, R. B. D'agostino, A. B. Newman, M. D. Lebowitz, T. G. Pickering, *et al.*, "Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study," *Jama*, vol. 283, no. 14, pp. 1829–1836, 2000.

[195] A. R. Babu, J. Herdegen, L. Fogelfeld, S. Shott, and T. Mazzone, "Type 2 diabetes, glycemic control, and continuous positive airway pressure in obstructive sleep apnea," *Archives of internal medicine*, vol. 165, no. 4, pp. 447–452, 2005.

[196] F. Campos-Rodriguez, M. A. Martinez-Garcia, M. Martinez, J. Duran-Cantolla, M. d. l. Peña, M. J. Masdeu, M. Gonzalez, F. d. Campo, I. Gallego, J. M. Marin, *et al.*, "Association between obstructive sleep apnea and cancer incidence in a large multicenter spanish cohort," *American journal of respiratory and critical care medicine*, vol. 187, no. 1, pp. 99–105, 2013.

[197] O. Marrone, S. Battaglia, P. Steiropoulos, O. K. Basoglu, J. A. Kvamme, S. Ryan, J.-L. Pepin, J. Verbraecken, L. Grote, J. Hedner, *et al.*, "Chronic kidney disease in european patients with obstructive sleep apnea: the esada cohort study," *Journal of sleep research*, vol. 25, no. 6, pp. 739–745, 2016.

[198] T. Saaresranta, J. Hedner, M. R. Bonsignore, R. L. Riha, W. T. McNicholas, T. Penzel, U. Anttalainen, J. A. Kvamme, M. Pretl, P. Sliwinski, *et al.*, "Clinical phenotypes and comorbidity in european sleep apnoea patients," *PloS one*, vol. 11, no. 10, p. e0163439, 2016.

[199] S. G. Memtsoudis, M. C. Besculides, and M. Mazumdar, "A rude awakening—the perioperative sleep apnea epidemic," *N Engl J Med*, vol. 368, no. 25, pp. 2352–2353, 2013.

[200] W. McNicholas, M. Bonsignore, and M. C. of EU Cost Action B26, "Sleep apnoea as an independent risk factor for cardiovascular disease: current evidence, basic mechanisms and research priorities," *European Respiratory Journal*, vol. 29, no. 1, pp. 156–178, 2007.

[201] V. A. Rossi, J. R. Stradling, and M. Kohler, "Effects of obstructive sleep apnoea on heart rhythm," *European Respiratory Journal*, vol. 41, no. 6, pp. 1439–1451, 2013.

[202] 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, H. Hakovirta, T. M. Salo, and T. T. Laitio, "Unrecognised obstructive sleep apnoea is common in severe peripheral arterial disease," *European Respiratory Journal*, vol. 41, no. 3, pp. 616–620, 2013.

[203] M. Sánchez-de-la Torre, F. Campos-Rodriguez, and F. Barbé, "Obstructive sleep apnoea and cardiovascular disease," *The Lancet Respiratory Medicine*, vol. 1, no. 1, pp. 61–72, 2013.

[204] R. B. Berry, R. Budhiraja, D. J. Gottlieb, D. Gozal, C. Iber, V. K. Kapur, C. L. Marcus, R. Mehra, S. Parthasarathy, S. F. Quan, *et al.*, "Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the american academy of sleep medicine," *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, vol. 8, no. 5, p. 597, 2012.

[205] H. Marti-Soler, C. Hirotsu, P. Marques-Vidal, P. Vollenweider, G. Waeber, M. Preisig, M. Tafti, S. B. Tufik, L. Bittencourt, S. Tufik, *et al.*, "The nosas score for screening of sleep-disordered breathing: a derivation and validation study," *The Lancet Respiratory Medicine*, vol. 4, no. 9, pp. 742–748, 2016.

[206] W. T. McNicholas, M. R. Bonsignore, P. Lévy, and S. Ryan, "Mild obstructive sleep apnoea: clinical relevance and approaches to management," *The Lancet Respiratory Medicine*, vol. 4, no. 10, pp. 826–834, 2016.

[207] E. S. Arnardottir, J. Verbraecken, M. Gonçalves, M. D. Gjerstad, L. Grote, F. J. Puertas, S. Mihaicuta, W. T. McNicholas, and L. Parrino, "Variability in recording and scoring of respiratory events during sleep in europe: a need for uniform standards," *Journal of sleep research*, vol. 25, no. 2, pp. 144–157, 2016.

[208] N. C. Netzer, R. A. Stoohs, C. M. Netzer, K. Clark, and K. P. Strohl, "Using the berlin questionnaire to identify patients at risk for the sleep apnea syndrome," *Annals of internal medicine*, vol. 131, no. 7, pp. 485–491, 1999.

[209] G. E. Silva, K. D. Vana, J. L. Goodwin, D. L. Sherrill, and S. F. Quan, "Identification of patients with sleep disordered breathing: comparing the four-variable screening tool, stop, stop-bang, and epworth sleepiness scales," *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, vol. 7, no. 5, p. 467, 2011.

[210] R. J. Farney, B. S. Walker, R. M. Farney, G. L. Snow, and J. M. Walker, "The stop-bang equivalent model and prediction of severity of obstructive sleep apnea: relation to polysomnographic measurements of the apnea/hypopnea index," *Journal of Clinical Sleep Medicine*, 2011.

[211] F. Chung, B. Yegneswaran, P. Liao, S. A. Chung, S. Vairavanathan, S. Islam, A. Khajehdehi, and C. M. Shapiro, "Stop questionnairea tool to screen patients for obstructive sleep apnea," *Anesthesiology: The Journal of the American Society of Anesthesiologists*, vol. 108, no. 5, pp. 812–821, 2008.

[212] F. Chung, B. Yegneswaran, P. Liao, S. A. Chung, S. Vairavanathan, S. Islam, A. Khajehdehi, and C. M. Shapiro, "Validation of the berlin questionnaire and american society of anesthesiologists checklist as screening tools for obstructive sleep apnea in surgical patients," *The Journal of the American Society of Anesthesiologists*, vol. 108, no. 5, pp. 822–830, 2008.

[213] F. Chung, Y. Yang, R. Brown, and P. Liao, "Alternative scoring models of stop-bang questionnaire improve specificity to detect undiagnosed obstructive sleep apnea," *J Clin Sleep Med*, vol. 10, no. 9, pp. 951–958, 2014.

[214] J. Hedner, L. Grote, M. Bonsignore, W. McNicholas, P. Lavie, G. Parati, P. Sliwinski, F. Barbé, W. De Backer, P. Escourrou, I. Fietze, J. A. Kvamme, C. Lombardi, O. Marrone, J. F. Masa, J. M. Montserrat, T. Penzel, M. Pretl, R. Riha, D. Rodenstein, T. Saaresranta, R. Schulz, R. Tkacova, G. Varoneckas, A. Vitols, H. Vrints, and

J. Zielinski, "The european sleep apnoea database (esada): report from 22 european sleep laboratories," *European Respiratory Journal*, vol. 38, no. 3, pp. 635–642, 2011.

[215] R. Santos-Silva, L. S. Castro, J. A. Taddei, S. Tufik, and L. R. A. Bittencourt, "Sleep disorders and demand for medical services: evidence from a population-based longitudinal study," *PloS one*, vol. 7, no. 2, p. e30085, 2012.

[216] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," 1968.

[217] A. Noack, "An energy model for visual graph clustering," in *International symposium on graph drawing*, pp. 425–436, Springer, 2003.

[218] A. Cairns, G. Poulos, and R. Bogan, "Sex differences in sleep apnea predictors and outcomes from home sleep apnea testing," *Nature and Science of Sleep*, vol. 8, p. 197, 2016.

[219] I. E. Gabbay and P. Lavie *Sleep and Breathing*, vol. 16, no. 2, pp. 453–460, 2012.

[220] M. W. Johns, "A new method for measuring daytime sleepiness: the epworth sleepiness scale," *sleep*, vol. 14, no. 6, pp. 540–545, 1991.

[221] M. R. Bonsignore, W. Randerath, R. Riha, D. Smyth, C. Gratziou, M. Goncalves, and W. T. McNicholas, "New rules on driver licensing for patients with obstructive sleep apnoea: Eu directive 2014/85/eu," 2016.

[222] M. Dickson and J. P. Gagnon, "The cost of new drug discovery and development," *Discovery Medicine*, vol. 4, no. 22, pp. 172–179, 2009.

[223] X.-Q. Chen, M. D. Antman, C. Gesenberg, and O. S. Gudmundsson, "Discovery pharmaceutics—challenges and opportunities," *The AAPS journal*, vol. 8, no. 2, pp. E402–E408, 2006.

[224] A. Mullard, "2016 fda drug approvals," *Nature Reviews Drug Discovery*, vol. 16, no. 2, pp. 73–76, 2017.

[225] A. Graul, P. Pina, E. Cruces, and M. Stringer, "The year's new drugs & biologics 2016: Part i.," *Drugs of today (Barcelona, Spain: 1998)*, vol. 53, no. 1, p. 27, 2017.

[226] F. Pammolli, L. Magazzini, and M. Riccaboni, "The productivity crisis in pharmaceutical r&d," *Nature reviews Drug discovery*, vol. 10, no. 6, pp. 428–438, 2011.

[227] 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*, vol. 138, no. 3, pp. 333–408, 2013.

[228] S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, *et al.*, "Drug repurposing: progress, challenges and recommendations," *Nature Reviews Drug Discovery*, vol. 18, no. 1, p. 41, 2019.

[229] B. Munos, "Lessons from 60 years of pharmaceutical innovation," *Nature Reviews Drug Discovery*, vol. 8, no. 12, pp. 959–968, 2009.

[230] A. F. Shaughnessy, "Old drugs, new tricks," *BMJ*, vol. 342, p. d741, 2011.

[231] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning," *Briefings in bioinformatics*, vol. 17, no. 1, pp. 2–12, 2015.

[232] M. Lotfi Shahreza, N. Ghadiri, S. R. Mousavi, J. Varshosaz, and J. R. Green, "A review of network-based approaches to drug repositioning," *Briefings in Bioinformatics*, p. bbx017, 2017.

[233] T. Nugent, V. Plachouras, and J. L. Leidner, "Computational drug repositioning based on side-effects mined from social media," *PeerJ Computer Science*, vol. 2, p. e46, 2016.

[234] M. Zhao and C. C. Yang, "Mining online heterogeneous healthcare networks for drug repositioning," in *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*, pp. 106–112, IEEE, 2016.

[235] K. Shameer, B. Readhead, and J. T Dudley, "Computational and experimental advances in drug repositioning for accelerated therapeutic stratification," *Current topics in medicinal chemistry*, vol. 15, no. 1, pp. 5–20, 2015.

[236] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, and J. Zheng, "Drug–target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2012.

[237] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.

[238] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *bioRxiv*, p. 100305, 2017.

[239] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "Drugbank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D901–D906, 2007.

[240] S. Ekins, J. Mestres, and B. Testa, "In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling," *British journal of pharmacology*, vol. 152, no. 1, pp. 9–20, 2007.

[241] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, "Dock 4.0: search strategies for automated molecular docking of flexible molecule databases," *Journal of computer-aided molecular design*, vol. 15, no. 5, pp. 411–428, 2001.

[242] 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-)*, vol. 185, no. 2, pp. 443–451, 2016.

[243] T. Arentze, P. van den Berg, and H. Timmermans, "Modeling social networks in geographic space: approach and empirical application," *Environment and Planning-Part A*, vol. 44, no. 5, p. 1101, 2012.

[244] V. Sekara, A. Stopczynski, and S. Lehmann, "Fundamental structures of dynamic social networks," *Proceedings of the national academy of sciences*, vol. 113, no. 36, pp. 9977–9982, 2016.

[245] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," *arXiv preprint arXiv:1105.0697*, 2011.

[246] S. I. Fierăscu, M. Pârvu, A. Topîrceanu, and M. Udrescu, "Exploring party switching in the post-1989 romanian politicians network from a complex network perspective," *Romanian Journal of Political Science*, vol. 18, no. 1, pp. 108–136, 2018.

[247] G. Barina, M. Udrescu, A. Barina, A. Topirceanu, and M. Vladutiu, "Agent-based simulations of payoff distribution in economic networks," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–18, 2019.

[248] G. Barina, M. Udrescu, A. Topirceanu, and M. Vladutiu, "Simulating payoff distribution in networks of economic agents," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 467–470, IEEE, 2018.

[249] M. Topirceanu, A. Topirceanu, and M. Udrescu, "Exploring currency exchange dynamics from a complex network perspective," in *2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pp. 63–68, IEEE, 2019.

[250] Y. Sun, L. Ma, A. Zeng, and W.-X. Wang, "Spreading to localized targets in complex networks," *Scientific reports*, vol. 6, 2016.

[251] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, vol. 524, no. 7563, pp. 65–68, 2015.

[252] K. Börner, W. B. Rouse, P. Trunfio, and H. E. Stanley, "Forecasting innovations in science, technology, and education," *Proceedings of the National Academy of Sciences*, vol. 115, no. 50, pp. 12573–12581, 2018.

[253] M. Gong, C. Song, C. Duan, L. Ma, and B. Shen, "An efficient memetic algorithm for influence maximization in social networks," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 22–33, 2016.

[254] J. Tang, R. Zhang, P. Wang, Z. Zhao, L. Fan, and X. Liu, "A discrete shuffled frog-leaping algorithm to identify influential nodes for influence maximization in social networks," *Knowledge-Based Systems*, vol. 187, p. 104833, 2020.

[255] A. Topirceanu, "Genetically driven optimal selection of opinion spreaders in complex networks," *Machine Learning and Artificial Intelligence: Proceedings of MLIS 2020*, vol. 332, p. 3, 2020.

[256] A. S. Sunar, N. A. Abdullah, S. White, and H. C. Davis, "Personalisation of moocs: the state of the art," 2015.

[257] L. Maggio, A. Saltarelli, and K. Stranack, "Crowdsourcing the curriculum: A mooc for personalized, connected learning," *EDUCAUSE Review*, 2016.

[258] R. M. Anderson, R. M. May, and B. Anderson, *Infectious diseases of humans: dynamics and control*, vol. 28. Wiley Online Library, 1992.

[259] M. Keeling, "The implications of network structure for epidemic dynamics," *Theoretical population biology*, vol. 67, no. 1, pp. 1–8, 2005.

[260] M. J. Keeling and P. Rohani, *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[261] M. Salathé and J. H. Jones, "Dynamics and control of diseases in networks with community structure," *PLoS Comput Biol*, vol. 6, no. 4, p. e1000736, 2010.

[262] A. Siu and Y. R. Wong, "Economic impact of sars: the case of hong kong," *Asian Economic Papers*, vol. 3, no. 1, pp. 62–83, 2004.

[263] J. W. Elston, C. Cartwright, P. Ndumbi, and J. Wright, "The health impact of the 2014–15 ebola outbreak," *Public health*, vol. 143, pp. 60–70, 2017.

[264] M. Nicola, Z. Alsafi, C. Sohrabi, A. Kerwan, A. Al-Jabir, C. Iosifidis, M. Agha, and R. Agha, "The socio-economic implications of the coronavirus and covid-19 pandemic: a review," *International journal of surgery*, 2020.

[265] J. Hellewell, S. Abbott, and et al., "Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts," *The Lancet Global Health*, 2020.

[266] J. Cohen and K. Kupferschmidt, "Countries test tactics in 'war' against covid-19," *Science*, vol. 367, no. 6484, pp. 1287–1288, 2020.

[267] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, H. Coupland, T. A. Mellan, H. Zhu, T. Berah, J. W. Eaton, P. N. Guzman, *et al.*, "Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in european countries: technical description update," *arXiv preprint arXiv:2004.11342*, 2020.

[268] A. J. Kucharski, T. W. Russell, and et al., "Early dynamics of transmission and control of covid-19: a mathematical modelling study," *The Lancet Infectious Diseases*, 2020.

[269] H. Markel, H. B. Lipman, J. A. Navarro, A. Sloan, J. R. Michalsen, A. M. Stern, and M. S. Cetron, "Nonpharmaceutical interventions implemented by us cities during the 1918-1919 influenza pandemic," *Jama*, vol. 298, no. 6, pp. 644–654, 2007.

[270] R. J. Hatchett, C. E. Mecher, and M. Lipsitch, "Public health interventions and epidemic intensity during the 1918 influenza pandemic," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7582–7587, 2007.

[271] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and corona virus disease-2019 (covid-19): the epidemic and the challenges," *International journal of antimicrobial agents*, p. 105924, 2020.

[272] K. Kupferschmidt and J. Cohen, "China's aggressive measures have slowed the coronavirus. they may not work in other countries," *Science. Mar*, 2020.

[273] J. Koo, A. Cook, M. Park, and et al., "Interventions to mitigate early spread of covid-19 in singapore: a modelling study," *Lancet Infect Dis.*, 2020.

[274] J. Liu and T. Zhang, "Epidemic spreading of an seirs model in scale-free networks," *Communications in Nonlinear Science and Numerical Simulation*, vol. 16, no. 8, pp. 3375–3384, 2011.

[275] F. D. Sahneh, C. Scoglio, and P. Van Mieghem, "Generalized epidemic mean-field model for spreading processes over multilayer complex networks," *IEEE/ACM Transactions on Networking*, vol. 21, no. 5, pp. 1609–1620, 2013.

[276] F. D. Sahneh and C. Scoglio, "Competitive epidemic spreading over arbitrary multilayer networks," *Physical Review E*, vol. 89, no. 6, p. 062817, 2014.

[277] K. Prem, Y. Liu, T. W. Russell, A. J. Kucharski, R. M. Eggo, N. Davies, S. Flasche, S. Clifford, C. A. Pearson, J. D. Munday, *et al.*, "The effect of control strategies to reduce social mixing on outcomes of the covid-19 epidemic in wuhan, china: a modelling study," *The Lancet Public Health*, 2020.

[278] P. Diaz, P. Constantine, K. Kalmbach, E. Jones, and S. Pankavich, "A modified seir model for the spread of ebola in western africa and metrics for resource allocation," *Applied Mathematics and Computation*, vol. 324, pp. 141–155, 2018.

[279] C. Dye and N. Gay, "Modeling the sars epidemic," *Science*, vol. 300, no. 5627, pp. 1884–1885, 2003.

[280] A. Arenas, W. Cota, J. Gomez-Gardenes, S. Gómez, C. Granell, J. T. Matamalas, D. Soriano-Panos, and B. Steinegger, "A mathematical model for the spatiotemporal epidemic spreading of covid19," *MedRxiv*, 2020.

[281] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke, "Strategies for mitigating an influenza pandemic," *Nature*, vol. 442, no. 7101, pp. 448–452, 2006.

[282] P. Block, M. Hoffman, I. J. Raabe, J. B. Dowd, C. Rahal, R. Kashyap, and M. C. Mills, "Social network-based distancing strategies to flatten the covid-19 curve in a post-lockdown world," *Nature Human Behaviour*, pp. 1–9, 2020.

[283] L. Thunström, S. C. Newbold, D. Finnoff, M. Ashworth, and J. F. Shogren, "The benefits and costs of using social distancing to flatten the curve for covid-19," *Journal of Benefit-Cost Analysis*, pp. 1–27, 2020.

[284] A. Atkeson, "What will be the economic impact of covid-19 in the us? rough estimates of disease scenarios," tech. rep., Nat. Bureau of Economic Research, 2020.

[285] J. Chen, H. Zhang, Z.-H. Guan, and T. Li, "Epidemic spreading on networks with overlapping community structure," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 1848–1854, 2012.

[286] K. Sun, W. Wang, L. Gao, Y. Wang, K. Luo, L. Ren, Z. Zhan, X. Chen, S. Zhao, Y. Huang, *et al.*, "Transmission heterogeneities, kinetics, and controllability of sars-cov-2," *Science*, vol. 371, no. 6526, 2021.

[287] W. Van den Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, "The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale," *BMC infectious diseases*, vol. 11, no. 1, p. 37, 2011.

[288] C. Viboud, K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, A. Vespignani, *et al.*, "The rapidd ebola forecasting challenge: Synthesis and lessons learnt," *Epidemics*, vol. 22, pp. 13–21, 2018.

[289] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[290] A. D. Nguyen, P. Sénac, V. Ramiro, and M. Diaz, "Steps-an approach for human mobility modeling," in *International Conference on Research in Networking*, pp. 254–265, Springer, 2011.

[291] M. G. Moore and G. Kearsley, *Distance education: A systems view of online learning*. Cengage Learning, 2011.

[292] J. A. Lara, D. Lizcano, M. A. Martínez, J. Pazos, and T. Riera, "A system for knowledge discovery in e-learning environments within the european higher education area–application to student data from open university of madrid, udima," *Computers & Education*, vol. 72, pp. 23–36, 2014.

[293] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS quarterly*, pp. 1165–1188, 2012.

[294] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study.," *International Working Group on Educational Data Mining*, 2009.

[295] M. M. Quadri and N. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global Journal of Computer Science and Technology*, 2010.

[296] S. Suthaharan, "Decision tree learning," in *Machine Learning Models and Algorithms for Big Data Classification*, pp. 237–269, Springer, 2016.

[297] L. Guàrdia, M. Maina, and A. Sangrà, "Mooc design principles: A pedagogical approach from the learner's perspective," *elearning papers*, no. 33, 2013.

[298] A. Topirceanu, M. Udrescu, R. Avram, and S. Mihaicuta, "Data analysis for patients with sleep apnea syndrome: A complex network approach," in *International Workshop Soft Computing Applications*, pp. 231–239, Springer, 2014.

[299] F. Bischof, J. Egresits, R. Schulz, W. J. Randerath, W. Galetke, S. Budweiser, G. Nilius, M. Arzt, A. Hetzenecker, G. Investigators, *et al.*, "Effects of continuous positive airway pressure therapy on daytime and nighttime arterial blood pressure in patients with severe obstructive sleep apnea and endothelial dysfunction," *Sleep and Breathing*, pp. 1–11, 2019.

[300] K. Skaltsa, L. Jover, and J. L. Carrasco, "Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty," *Biometrical Journal*, vol. 52, no. 5, pp. 676–697, 2010.

[301] G. Altay and F. Emmert-Streib, "Inferring the conservative causal core of gene regulatory networks," *BMC systems biology*, vol. 4, no. 1, p. 132, 2010.