Understanding Complex Systems: When Big Data meets Network Science

Ingo Scholtes

Abstract: Better understanding and controlling complex systems has become a grand challenge not only for computer science, but also for the natural and social sciences. Many of these systems have in common that they can be studied from a network perspective. Consequently methods from network science have proven instrumental in their analysis. In this article, I introduce the macroscopic perspective that is at the heart of network science. Summarizing my recent research activities, I discuss how a combination of this perspective with Big Data methods can improve our understanding of complex systems.

ACM CCS: Networks → Network properties → Network structure; Networks → Network properties → Network dynamics; Software and its engineering → Collaboration in software development; Computing methodologies → Machine learning approaches

Keywords: complex networks, data mining, socio-technical systems, temporal networks

1 Complex Systems: A Network Perspective

We live in a connected world, with complex networked systems all around us. Examples include globally distributed information systems like the World Wide Web, the increasingly digitized social fabric into which we are embedded, as well as critical infrastructures like telecommunication networks, electrical grids or transportation systems which we depend on. Better understanding, designing and controlling such systems, and preventing emerging systemic risks has become a grand challenge for computer science. Luckily, we are not alone in this endeavour. Complex systems consisting of thousands or millions of interconnected elements are studied in disciplines as diverse as sociology, economy, management science, biology, neuroscience or physics. And thanks to the ongoing trend towards *computational sciences*, these studies increasingly translate into quantitative research employing Big Data methods: Driven by the fact that social interactions increasingly manifest themselves as digital traces, *computational social science* uses the resulting fine-grained data to study the structure and dynamics of social systems. Large-scale data sets covering trade relations and market dynamics have resulted in a surge of data-driven modelling approaches in areas such as *Quantitative economics* or *Econophysics*. And the development of high-throughput genetic sequencing and high-resolution medical imaging techniques has triggered an explosion of data being studied in the different areas of *computational biology*.

What do these research themes have in common and why are they relevant for computer science? The answer is that the complex systems studied in these seemingly different areas give rise to *relational data* which can be represented as *complex networks*. Such a representation is not only useful to study how the elements of a system are connected to each other and which of those elements are most important. More importantly, network science provides us with a *macroscopic perspective* on complex systems which can be used to differentiate structure from noise, thus providing the foundation for pattern recognition, machine learning and statistical inference techniques. As such, it is not surprising that *network science* has become instrumental for the development of statistical and computational tools that facilitate Big Data analyses of complex systems not only in computer science, but also in the natural and social sciences [1, 2].

In this article, I introduce the macroscopic perspective which is at the heart of network science. Summarizing recent work, I demonstrate how it can be applied in data-driven studies of socio-technical systems. Highlighting challenges and limitations of this perspective, I finally comment on interesting current research topics as well as on opportunities for future research.

2 Network Science: The Macro-Perspective

How can we analyze data from networked systems? The reader may argue that *graph-theoretic methods* have been standard techniques since the earliest days of computer science. This is true, however these methods make the important assumption that we have *full knowledge* about the network, i.e. that we precisely know to which other nodes each node is connected. This detailed knowledge then allows us to apply a wealth of node- and network-level measures, algorithms and visualizations which can provide important insights about complex systems. However, the condition of full knowledge about microscopic details proves difficult in many real-world scenarios: For large decentralized systems like, e.g., the Internet or P2P systems, mapping network topologies can be anything from costly and time-consuming to impossible. Furthermore, as systems evolve such data are likely to be inaccurate as soon as they have been collected. And even if we succeed, we may end up with data sets comprised of Billions of nodes and hundreds of Billion of links. At this scale, graph-theoretic analyses may be possible in principle, however the time and resources they demand are often prohibitive.

There is a common theme in the scenarios described above: We lack information about system details and are thus required to reason under uncertainty. Combining methods from *random graph theory* and *statistical physics*, this problem can be addressed by a *macroscopic perspective* on networked systems, which is often subsumed under the umbrella of *network science*. Notably, statistical physics faces a similar problem, being required to reason about particle systems despite a lack of knowledge about microscopic details such as particle positions, velocities or interactions. Probabilistic methods, which are based on *aggregate statistics* of large populations rather than details of individual elements, have thus been developed for this purpose. The same ideas can be applied to analyze large-scale networked systems and it involves the following three steps: We first compute aggregate statistics of interest, such as the number of nodes, the density of links, the distribution of node degrees, the presence of clustering structures, correlations between neighboring nodes, etc. Notably, using distributed data processing techniques, such aggregate statistics can be computed efficiently even for massive data sets, while for distributed systems they can be estimated efficiently by means of sensible sampling techniques. Based on these statistics, in a second step we can then define *statistical ensembles*, i.e. probability spaces in which we assign probabilities to all possible network realizations which are consistent with the given aggregate statistics. Using analytical tools like generating functions, or computational statistics methods like Metropolis sampling, we can finally reason about the *expected properties* of a system given that we only know aggregate statistics of its network topology.

3 Applications in Socio-technical systems

In most of the literature, the macroscopic perspective outlined above is heavily associated with terminology and methods from statistical and computational physics. This can be daunting, however it should not prevent us from using these methods to address problems in domains relevant to computer science. One domain in which we applied these methods is *empirical software engineering*, which is concerned with the question how complex software systems evolve and how humans collaborate in their development. Here, a network science perspective can be applied to large-scale data covering both *technical* and *social* aspects of software systems.

Considering the technical dimension, we can mine software repositories to extract evolving networks of dependencies between software artifacts such as methods, classes or packages. We recently applied this method to quantify the *congruence* between developer-declared modules (e.g. packages in Java) in software architectures and cluster patterns emerging in their dependency networks [3]. A high degree of congruence corresponds to a reasonable modular structure in which dependencies are preferentially contained *within* modules. Here, the macro-perspective on networks is instrumental to establish a statistically significant measure which properly accounts for the level of congruence that we can already expect at random. The resulting measure provides interesting insights into the evolution of software architectures which typically start out in a state of high modular congruence which decreases as they grow. It can further be used to inform *refactoring* efforts that improve the modular structure and thus the maintainability of complex software projects [3].

While this work takes a network perspective on the technical dimension of software projects, it is equally important to consider social aspects emerging in teams of developers. How do social structures affect the performance of development teams? Are there indicators for emerging problems in social organizations? And how should work be organized to be most productive? Again, we can address these questions by a combination of large-scale data analysis and network science methods. Using data that cover the full history of large Open Source Software communities, we performed a statistical analysis of evolving networks that capture collaborations between community members [4]. Here, the macroscopic perspective on networks allowed us to go beyond a micro-level analysis of the (highly dynamic) collaboration structures. Using macroscopic and time-dependent measures which capture the efficiency of information flow and synchronization processes, we can instead identify regime changes which affect the performance of development teams, as well as the effect of a single individual leaving the project [4]. We foresee that project managers can use similar *macro-level monitoring techniques* to identify emerging risks in software projects.

Using a macro-perspective on networks, statistical patterns at the aggregate level can inform us about *collective properties* of a social organization, such as their robustness or the efficiency of information flow. A different yet related question is whether we can identify statistically significant patterns which show how individuals are influenced by the network structures around them. We studied this question using data on more than 5.8 Million time-stamped transactions recorded by the bug trackers of four major Open Source Software projects over a period of more than a decade [5]. A statistical analysis revealed that the position of community members in the evolving collaboration networks is a strong indicator for the quality of the bug reports they provide. Addressing the question how we can improve the design of collaboration tools, we utilized this finding to automatically identify *valid* bug reports that refer to actual software defects. For this, we combined our network perspective on collaboration structures with a machine learning classifier, thus obtaining an automated method that identifies valid bug reports with a precision of up to 90.3% [5]. Notably, taking a network perspective on the problem of identifying valid bug reports allowed us to achieve a significant improvement over previous approaches addressing the same question.

Clearly, we don't have to limit the use of these methods to problems in empirical software engineering. The finding that the position of individuals in social networks is correlated with the (perceived) quality of information they provide, points to interesting general questions that are relevant in the design of information and recommender systems. We thus applied our approach of combining machine learning and network science techniques to a data set of more than $100,000$ scholarly publications linked by Millions of citations [6]. Studying correlations between citation and coauthorship networks, our results show that the citation counts of scholars depend in a statistically significant way on their position in the coauthorship network. This dependence actually allows for an automated classifier, which - solely based on the network position of their authors - predicts with high precision which papers will be highly cited in the future. While one needs to interpret these results carefully, they provide interesting insights about mechanisms of social cognition and social information filtering at work in large social information systems [7]. They can further be seen as a cautionary tale, contributing to the debate about the fallacies of citation-based *information ranking mechanisms* and *impact measures*.

4 Challenges and Research Perspectives

Without doubt, methods from network science provide interesting opportunities for data-driven studies of complex systems. On the one hand, they help us to better understand the complex systems that we design. On the other hand, the network perspective allows to identify analogies to complex systems studied in other disciplines, thus generating insights that go beyond our own field. At the same time, it cannot be denied that our understanding of network-based methods is still in its infancy, thus posing both challenges and opportunities for future research.

Limitations of Ensemble Studies A first challenge results from the interpretation of findings that are based on the macro-perspective outlined in section 2. This approach helps us to derive *expected properties* of networks based on aggregate statistics, which is particularly handy in situations where we cannot use information about a system's details. However we need to be careful when interpreting these *expected* properties. More precisely, the realizations subsumed in a statistical ensemble can exhibit large variances. As such, the properties of a particular realization observed in reality can differ quite substantially from the expected properties computed based on the ensemble. This may seem trivial, however a number of works in network science have failed to clearly express this fundamental limitation of ensemble studies, thus triggering a lively scientific debate [8].

As for any other methodology, considering the limitations and implicit assumptions of network science methods is crucial to arrive at the right conclusions. I believe the best way to address this issue is by means of better education. At ETH Zürich we have taken up the challenge by developing an interdisciplinary course on network science which teaches students from computer science, engineering, neuroscience, management, and physics to benefit from these methods, while being alert to their limitations.

Reasoning about Temporal Networks A second major challenge is associated with the fact network topologies of real systems are not static, but rather change continuously. This may seem obvious, however it poses a problem for how we typically represent such dynamic systems: We mostly consider them as static networks, aggregating all links that occur within a certain time interval. However, with this we neglect the temporal dimension of networked systems, possibly arriving at wrong conclusions about the robustness of systems, the importance of nodes or dynamical processes [9].

Addressing the analysis of such *temporal networks*, much of my latest research has focused on the question how we can better incorporate time in network-based studies. A simple yet important question which we studied recently is how the *ordering of links* affects causality in complex systems. As an example, consider a simple network with three nodes a, b, and c, connected by two time-stamped links (a, b, t) and (b, c, t') at times t and t'. Clearly, node a can only influence c, if (a, b) appears *before* (b, c), i.e. if $t < t'$. Studying data from social, biological and technical systems, we showed that neglecting the order of links

severely limits our understanding of dynamical processes in networked systems [10]. To overcome these limitations, we developed *higher-order aggregate networks*, a generalization of the commonly used static network abstraction that allows to incorporate both the topology and the order of links in the analysis of time-stamped relational data [11]. These works highlight the presence of a largely unexplored *temporal-topological* dimension of complex systems with interesting opportunities for new data mining techniques which I look forward to further explore in future research.

Towards Multi-Layer Network Models Finally, a third major challenge is due to the rather simplistic way in which most current network-based studies represent data from complex systems: Elements of a system are mostly represented as featureless nodes connected by a single type of links (possibly with different strengths). This clearly is an oversimplification of real systems, which often exhibit multiple types of links or nodes with heterogeneous characteristics. Furthermore, for many systems such as communication networks or the electrical grid it is non-trivial to define system boundaries which would allow to study them in isolation. By means of mutual dependencies, engineered systems are rather increasingly *interwoven*, thus requiring novel modeling approaches which capture their multi-layered structure. [12]. So far, reconciling the macro-perspective of network science with the complex characteristics of real-world systems has proven to be a challenge. However, recent advances in the theory of *interconnected* and *multi-layer networks* can be seen as promising steps in the direction of developing a multi-layer network science.

5 Conclusion

As networked systems affect more and more aspects of our lives, a solid understanding of their structure and dynamics is of paramount importance. The macroscopic perspective on networks, along with the associated statistical, computational and analytical techniques provide us with a rich set of tools helping us to better understand, and thus design, complex systems. Analyzing large-scale relational data from systems occurring in nature and society, they further allow us to address a broad range of scientific questions. I thus believe that the combination of methods from network science and Big Data provides an exciting and fertile field of study for a new generation of interdisciplinary computer scientists. I further believe that computer science is particularly adept to provide the required combination of theoretical expertise in the modeling of systems and practical skills in the analysis of massive data sets.

Through a combination of *theoretical research* improving network-based data-mining methods, and *applied research* demonstrating applications in socio-technical systems, I look forward to contribute to this exciting field. I am further convinced that my GI Junior Fellowship will help me to foster the interdisciplinary exchange required to better understand the complex systems that increasingly influence our lives.

Literature

[1] M. van Steen. *On the Complexity of Simple Distributed Systems*. IEEE Distributed Systems Online, 5(5), 2004

[2] A. Vespignani. *Modelling dynamical processes in complex socio-technical systems*. Nature Physics, 8:32-39, 2012

[3] M.S. Zanetti, C.J. Tessone, I. Scholtes, F. Schweitzer. *Automated Software Remodularization Based on Move Refactoring*. 13th International Conference on Modularity, 2014.

[4] M.S. Zanetti, I. Scholtes, C.J. Tessone, F. Schweitzer. *The rise and fall of a central contributor: Dynamics of social organization and performance in the Gentoo community*. International Workshop on Cooperative and Human Aspects of Software Engineering, pp. 49-56, 2013.

[5] M.S. Zanetti, I. Scholtes, C.J. Tessone, F. Schweitzer. *Categorizing bugs with social networks: A case study on four open source software communities*. 35th International Conference on Software Engineering (ICSE), pp. 1032-1041, 2013

[6] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, F. Schweitzer. *Predicting Scientific Success Based on Coauthorship Networks*. EPJ Data Science, 3, 2014.

[7] I. Scholtes, R. Pfitzner, F. Schweitzer. *The Social Dimension of Information Ranking: A Discussion of Research Challenges and Approaches*. Socioinformatics - The Social Impact of Interactions between Humans and IT, Springer Proceedings in Complexity, pp. 45-61, October 2014

[8] J.C. Doyle, D.L. Anderson, L. Li, S. Low, M. Roughan, S. Shalunov, W. Willinger. *The "robust yet fragile" nature of the Internet*. PNAS, 102(41):14497-14502, 2005.

[9] P. Holme, J. Sarimäki. *Temporal networks*. Physics Reports, 519(3):97-125, 2012

[10] R. Pfitzner, I. Scholtes, A. Garas, C.J. Tessone, F. Schweitzer. *Betweenness Preference: Quantifying Correlations in the Topological Dynamics of Temporal Networks*. Phys. Rev. Lett., 110(19):198701, 2013.

[11] I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C.J. Tessone, F. Schweitzer. *Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks*. Nature Communications, 5:5024, 2014.

[12] S. Tomforde, J. Hähner, B. Sick. *Interwoven Systems*. Informatik Spektrum, 37(5):483-487, 2014

Dr. Ingo Scholtes is a senior researcher at the Chair of Systems Design at ETH Zürich. Following studies in computer science and mathematics, he completed his doctorate studies in the Systems Software and Distributed Systems group at the University of Trier in 2011. He was involved in the Large Hadron Collider experiment at CERN, designing and implementing a Peer-to-Peer-based framework for large-scale data distribution which is since being used to monitor particle collision data from the ATLAS detector. Inspired by this experience, he turned his attention to the modeling and analysis of complex networked systems. His latest research addresses applications of network science in the analysis of data from socio-technical systems, but also from biology and sociology. In a theoretical line of research he further studies new methods in the analysis of time-stamped network data. At ETH Zürich he developed a course on network science which bridges the curricula of engineering and natural sciences. He previously held a scholarship from the Studienstiftung des Deutschen Volkes and was awarded a Junior-Fellowship from the Gesellschaft für Informatik in 2014.

Address: ETH Zürich, Chair of Systems Design, CH-8092 Zürich, Switzerland, E-Mail: ischoltes@ethz.ch