Empirical analysis of complex social and financial networks

Theses

Dániel Kondor

Supervisor: Gábor Vattay, DSc
Consultant: István Csabai, DSc

Department of Physics of Complex Systems
Eötvös Loránd University, Budapest, Hungary

Doctoral School of Physics
Head of School: László Palla, DSc

Doctoral Program for Statistical Physics,
Biological Physics and Physics of Quantum Systems
Head of Program: Jenő Kürti, DSc

Budapest, 2015

In the past decades, the study of networks has emerged as a popular interdisciplinary research field. A main reason of this is that many complex systems around us can be effectively modeled using networks, where parts of the system are represented as nodes and possible interactions are edges. This approach can help in understanding the function of a complex system in many cases: analyzing the topology of the network can be a crucial step in understanding the possible processes taking place on it. This motivation is complemented by the increasing availability of relevant data about real-world networks, both regarding the topology and the behavior of individual parts. More generally, the amount of data openly available for scientific research has been displaying a high-paced increase in the past two decades. Some of these are databases storing the results of measurements performed on unprecedented scales, e.g. the SDSS database in the case of astronomy or the TCGA database of genome sequencing data. Other are the data collected by various sensors present with an increasing importance in society, e.g. conventional or smart mobile phones, GPS-traces collected by various devices, logs of Internet activity or financial transactions. Most of these are collected for commercial purposes, while some are shared with the scientific community too. A good example is the Twitter social network, where a sample of all public activity is available for anyone. An example from a quite different background is the Bitcoin virtual currency system, where a fundamental basis for its proper functioning is the availability of data for anyone.

Such increase in data volume has brought a new paradigm to science both in terms of possibilities and methodology. Data mining has become a new interdisciplinary scientific discipline with increasing influence both in science and in commercial applications. In science, this also means that existing scientific fields can hope to answer questions whose study was previously limited by the lack of data collection or measurement possibilities. Also, while performing analysis on the new datasets, new questions can arise, some becoming a separate sub-discipline by itself. In the case of social systems, the increase in data volume has been also accompanied with more fundamental changes in the terms of data collection. While previously, most empirical studies about human behavior were based on experiments conducted in a controlled, artificial environment and

surveys, nowadays, we have the possibility of performing *passive* measurements on massive scales. These can mean any data collection which does not require the active collaboration of participants, rather uses the records from the digital traces of their everyday activities, which they consent to and which can form a basis of research. A good example for this is the case of public activity on online social networks which are becoming an increasingly important part of everyday life.

Over the course of my doctoral studies, I focused on some of the complex networks which became available for scientific analysis in the past decades. Most of the data used in my research can be considered passive measurement, and the conducted analysis show some possible examples for their usability. The two main systems in the focus of the thesis are the Twitter social network and the Bitcoin digital currency system. In the case of Twitter, the main focus of research is the public posts of users, which can be augmented by the network of connections between the users. An especially interesting possibility is given by the users who choose to also include their GPS coordinates with their messages (as measured by the smartphones in most cases), thus enabling us to conduct large-scale spatial studies. On the other hand, the Bitcoin system presents a unique opportunity as a financial network where the complete list of transactions is publicly available, which is in contrast with traditional financial systems where this is usually considered highly sensitive and private (privacy of users is protected by having only limited possibilities for establishing a connection between Bitcoin address identifiers in the transaction list and their real identity). In my thesis, I show some examples how the possibilities presented by these new data sources can be exploited and what kind of questions can be efficiently answered by them.

In the first part of my thesis, I concentrate on analyzing the content of Twitter messages (tweets), and some of the related technical challenges. I present an effective method for the analysis of the spatial variation of the content of the tweets augmented with geographic coordinates. I demonstrate this method on a sample of over 500 million tweets from the United States of America. Apart from the analysis of the content of messages, I also present a related study focusing on the challenges of managing the dataset of billions

of tweets and classifying each of them by the administrative region in which they were posted. The next part focuses on the Bitcoin digital currency system: after giving an overview of the relevant technical details of it, I present a statistical analysis based on all transactions which happened in it over the course of six years since its inception. A main result is that preferential attachment plays an important role in the growth of the network and the evolution of the distribution of wealth among the Bitcoin users. I also present a study focusing on the transactions which happened in 2012 and 2013 where a possible method for relating the temporal evolution of the network structure and the bitcoin exchange price time series was evaluated.

In the last chapter of the thesis, I focus on the possibility of measuring the fractal dimension of networks. After reviewing previous definitions, I present a novel method especially tailored for spatially embedded networks. I demonstrate the possible use of this new dimension metric by measuring the dimension of various random and real-world networks, including samples of the Twitter user network.

The main results are summarized in the following:

1. Extending on the concept of spectral dimension of networks, I defined a new dimension concept which is applicable to networks with full or partial spatial embedding. Using this new dimension metric, I show that for a spatially embedded network, the distribution of link lengths is especially important with regards to network structure, and thus has a profound influence on the dimensionality of the network, too. Nevertheless, I have found that it is not in itself enough to completely control it, as further structural differences among the studied networks give rise to further, more subtle differences in network dimension as well. [1]

2. Applying the procedure of Robust Principal Component Analysis to a set of several hundred million tweets posted in the United States of America, I demonstrate that there are important language use differences among Twitter users from different regions in the USA, and these can be analyzed efficiently with this method. I also related these principal compo-

nents to the results of a previous study about the well-being of people. [2]

3. I gave an efficient solution for storing and indexing several billion Twitter messages augmented with geographic coordinates in our database system. The main challenge, which I was able to solve is classifying the coordinates by administrative regions defined by arbitrary complex polygons in a way which is well-integrated to the rest of the database system. The solution presented here gives an almost one hundred times speedup when compared to the original solution present in the database product and forms the basis of several ongoing research projects. [3]

4. Performing an analysis of the complete history of transactions in the Bitcoin digital currency system, I establish that preferential attachment is an important factor through the lifetime of the system, with regards to both the transaction network and the wealth of participants. The growth of the network exhibits linear preferential attachment while the growth of balances can be approximated by a sublinear process. I also show that the two are related; there is a positive correlation between the network degree and the balance of Bitcoin network users. [4]

5. Analyzing the time evolution of the Bitcoin network, I established a connection between the network structure and external factors; representing the network evolution with a time series matrix, I found significant correlation among some of the principal components of this matrix and the bitcoin exchange price time series, showing that external measures such as the exchange price which is entirely determined by market effects can be related to the internal structure of the network. [5]

Publications supporting the thesis

[1] Dániel Kondor, Péter Mátray, István Csabai, and Gábor Vattay. Measuring the dimension of partially embedded networks. *Physica A*, 392(18):4160–4171, 2013.

[2] Dániel Kondor, István Csabai, László Dobos, János Szüle, Norbert Barankai, Tamás Hanyecz, Tamás Sebők, Zsófia Kallus, and Gábor Vattay. Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages. *CogInfoCom 2013*, 2013.

[3] Dániel Kondor, László Dobos, István Csabai, András Bodor, Gábor Vattay, Tamás Budavári, and Alexander S. Szalay. Efficient classification of billions of points into complex geographic regions using hierarchical triangular mesh. *Proceedings of the 26th International Conference on Scientific and Statistical Database Management - SSDBM '14*, 2014.

[4] Dániel Kondor, Márton Pósfai, István Csabai, and Gábor Vattay. Do the rich get richer? An empirical analysis of the BitCoin transaction network. *PLoS ONE*, 9(2):e86197, 2014.

[5] Dániel Kondor, István Csabai, János Szüle, Márton Pósfai, and Gábor Vattay. Inferring the interplay between network structure and market effects in Bitcoin. *New Journal of Physics*, 16(12):125003, 2014.

Other publications

[6] Áron Szabó, Gábor Vattay, and Dániel Kondor. A cell signaling model as a trainable neural nanonetwork. *Nano Communication Networks*, 2012.

[7] Dániel Kondor and Gábor Vattay. Dynamics and structure in cell signaling networks: off-state stability and dynamically positive cycles. *PloS one*, 8(3):e57653, 2013.

[8] László Dobos, János Szüle, Tamás Bodnár, Tamás Hanyecz, Tamás Sebők, Dániel Kondor, Zsófia Kallus, József Stéger, István Csabai, and Gábor Vattay. A multi-terabyte relational database for geo-tagged social network data. *CogInfoCom 2013*, 2013.

[9] Zsófia Kallus, Norbert Barankai, Dániel Kondor, László Dobos, Tamás Hanyecz, János Szüle, József Stéger, Tamás Sebők, Gábor Vattay, and István Csabai. Regional properties of global communication as reflected in aggregated Twitter data. *CogInfoCom 2013*, 2013.

[10] János Szüle, Dániel Kondor, László Dobos, István Csabai, and Gábor Vattay. Lost in the City: Revisiting Milgram's Experiment in the Age of Social Networks. *PloS ONE*, 9(11):e111973, 2014.