Analysis of complex socio-economic systems

Theses

Eszter Bokányi

Supervisor: Gábor Vattay, DSc

Eötvös Loránd University, Budapest
Department of Physics of Complex Systems

Doctoral School of Physics
Head of School: Jenő Gubicza, DSc

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

Budapest, 2019

Introduction

Most systems surrounding us in everyday life consist of the collection of numerous elements and their interactions. Often, the interactions the smallest scale of these systems lead to unexpected outcomes at the observable level. This kind of emergent behavior is a hallmark of complex systems, whose patterns are often hard to predict from the properties of their constituents. Some universal properties common to various systems might also characterize these patterns, for example, the universal scaling properties in certain physical systems. Understanding this kind of universal behavior is a great theoretical challenge, while in many cases, it is also crucial for practical applications.

In this thesis, I aimed to understand and model human behavior with tools from complexity science. There are already several studies that use models borrowed from physics such as gravity or radiation for explaining phenomena like human mobility. For validating such models, rich data sources such as mobile call records and social media data have recently become available. In the present work, I used mobile call records, billions of the messages of the online social network Twitter, and a detailed historical database of the voting data of United States counties for explaining aggregate patterns.

In Chapter 1, I gave an extensive review of the background of using geographically tagged social media data for predicting real-world outcomes, and I also introduced the online social network Twitter and the structure of its messages in detail. The end of Chapter 1 summarizes the literature on urban scaling, that is the underlying theory in Chapters 2-3 of the thesis. In Chapters 2-3 I present how scaling laws describe electoral processes and word choice in geographical areas. Then, in Chapters 4-5, I show two applications for connecting geolocated Twitter messages to real-world labels or measures of certain geographical areas.

Theses

My results are summarized in the following theses:

1. **I showed that the results of the most recent and historical presidential elections in the metropolitan areas of the United States exhibit scaling properties, and I was able to explain the relationship between the scaling parameters with a model based on the notion of complexity.** [1]

   First, I show that votes given to the Republican and Democrat candidates in the 2016 presidential election of the United States follow the laws of urban scaling. This is in line with the observation that bigger cities voted more for Democrats. In the context of urban scaling, this means that the number of Democrat voters follows a power-law relationship as a function of the voter turnout in US metropolitan areas, and the exponent in the scaling relationship is superlinear. Moreover, the data of all elections since 1960 fits the urban scaling laws, and the exponents of the two parties are not independent of each other. I show that the dependence follows from substituting the scaling relationships into the conservation of the voting probability when summing up for all parties. The theoretical relationship derived from this probability conservation is confirmed by the measured exponents. The two remaining parameters for each year's fits, namely the two intercepts, can also be derived from the exponents. Therefore, by knowing the exponents of only one of the parties, it is possible to determine all other fit parameters. Because deviations from the fits only show a lognormal distribution for the Democrats, it is their exponent that comes from an urban scaling process, and the other parameters are the derivatives of the Democrat exponent.

Then, I adapt the scaling model of Gomez-Lievano et al. for the election process. According to the model, each voter has to accept a certain number of basic values or minorities from the program of the Democrats to vote for them. The bigger the city a voter lives in, the more probable it is that by living in the city, the voter is exposed to diverse groups or opinions. This increases the probability of developing a tolerance or acceptance towards these groups or opinions. The model explains the growth of the Democrat exponent since the 1980s through the increasing number of issues and groups that have to be accepted for a Democrat voter. Urban scaling is not only a valid model for the election of the United States, but also for other voting processes. In the 2016 EU referendum in the United Kingdom, the number of people voting for the Leave or Remain opinions also fit urban scaling laws. The model also sheds light on the so-called "immigration paradox", that is, cities voting for Remain with a greater percentage were the ones having the most diverse ethnic groups.

2. **I showed that the word frequencies in the messages of an online social media platform follow the laws of urban scaling in the metropolitan areas of the United States.** [2]

   With the tools of urban scaling, I investigate the relationship between the Twitter messages sent from the United States and the population of metropolitan areas. The total number of messages and the total number of words both have a slightly superlinear exponent, that is less than the literature value for exponents of other human interactions. Words having an exponent close to that of the total number of words belong to a core vocabulary of the language, for example, articles or pronouns. The most sublinearly or superlinearly scaling words have a meaning that is in

line with the sub- or superlinear urban measures described in the previous literature. Thus, population density has a quantifiable effect on word choice and word frequencies on Twitter. I also show that the number of distinct words scales sublinearly with population, that corresponds to Heaps law from linguistics. Overall, this sublinear scaling demonstrates that even though the number of sent messages grows superlinearly, there is a decreasing marginal need for new words for this increased amount of communication in cities.

3. **Using a linear model, I introduced a new measure for geographical areas based on online social network activity, that shows a significant correlation with employment and unemployment levels.** [3]

   The new measure is based on the daily activity profiles measures from public posts of the online social network Twitter. The main idea is that the activity in a United States county is the superposition of the activity patterns of two groups of people. While one group has a regular daily schedule because of work or school, the other group's lifestyle lacks this kind of regularity. Therefore, I assume that the measured activity pattern is the linear combination of two base patterns, and I formulate a linear model based on this assumption. By minimizing the squared error between this model and the data, I obtain the two base patterns, that correspond to a daily activity pattern for people who get up early and go to bed early, the "active" group, and to another pattern for people getting up later and going to bed later, the "inactive" group. The coefficient giving the mixing ratio of the two groups in each county correlates more with employment and unemployment levels than any previous measures in the literature. Therefore, instead of using the activity of an arbitrarily

chosen time window throughout the day, it is possible to use the whole shape of the daily profile for predicting employment, and it gives better results than the former approach. This result is easily extendable to other data types.

4. **I showed that supervised learning can retrieve mobile call based land use clusters of different cities from social media data.** [4]

   For this work, the land use clusters (e.g. business, commercial, recreational, residential and mixed) of three different cities, London, New York City, and Los Angeles are given. These clusters are based on mobile call data by aggregating weekly average activity patterns of small pixels, and then clustering similar patterns into the same land use label.

   Then, I collect messages from the online social network Twitter spanning a year in time. I sort these messages into the pixels using their coordinates and determine the average frequency of words in these pixels, as well as their average weekly timelines. First, I show that words with a weekly activity profile correlating highly with that of the clusters have a meaning related to the cluster label. Words being significantly more frequent within a cluster are also related to land use.

   Next, I train a supervised learning algorithm, the so-called Random Forest classifier to separate land use clusters based on word frequencies within a city. The classifier can reliably retrieve the cluster types. Then, I make another classifier for the problem of separating pixels having the same land use type but belonging to different cities. The algorithm almost perfectly predicts the cities for the pixels in this case, even though the weekly activity profiles for business clusters is indistinguishable from the

mobile-based clustering. This approach makes it possible to connect results from different data sources, and to cross-check the extent to which the results based on one dataset are retrievable in a dataset from a different source.

Publications supporting the theses

[1] Bokányi, E., Szállási, Z. & Vattay, G. "Universal scaling laws in metro area election results". *PLOS ONE* **13** (ed Braha, D.) e0192913 (2018)

[2] Bokányi E., Kondor, D. & Vattay, G. "Scaling in words on Twitter". *submitted to Royal Society Open Science* (2019)

[3] Bokányi, E., Lábszki, Z. & Vattay, G. "Prediction of employment and unemployment rates from Twitter daily rhythms in the US". *European Physical Journal Data Science* **6,** 14 (2017)

[4] Bokányi E., Kallus, Zs. & Gódor, I. "Collective Sensing of Evolving Urban Structures: from Activity-based to Content-Aware Social Monitoring". *accepted in Environment and Planning B* (2019)

Other publlications

[5] Bokányi, E., Kondor, D., Dobos, L., Sebők, T., Stéger, J., Csabai, I. & Vattay, G. "Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States". *Palgrave Communications* **2,** 16010 (2016)

[6] Kmetty, Z., Koltai, J., Bokányi, E. & Bozsonyi, K. "Seasonality Pattern of Suicides in the US – a Comparative Analysis of a Twitter Based Bad-mood Index and Committed Suicides". *Intersections* **3** (2017)

[7] Kallus, Z., Kondor, D., Stéger, J., Csabai, I., Bokányi, E. & Vattay, G. in *ICT Innovations 2017: Data-Driven Innovation. 9th International Conference, ICT Innovations 2017, Skopje, Macedonia, September 18-23, 2017, Proceedings* (eds Trajanov, D. & Bakeva, V.) 3–12 (Springer International Publishing, Cham, 2017)

[8] Radnóczi, G., Bokányi, E., Erdélyi, Z. & Misják, F. "Size dependent spinodal decomposition in Cu-Ag nanoparticles". *Acta Materialia* **123,** 82–89 (2017)

[9] Sóti, A., Bokányi, E. & Vattay, G. "Urban scaling of football followership on Twitter". *Acta Polytechnica Hungarica* **15,** 239–250 (2018)