DATA ANALYSIS OF TWITTER DATA

Nikola Kaspříková

Abstract

Social network analysis like the analysis of Twitter network is an attractive topic for many students of data analysis and statistics courses these days. The social network analysis provides a nice opportunity to introduce some useful concepts from data science and graph theory in the classes. A basic exploratory data analysis of the Twitter followers network is reported. The analysis addresses the network of followers of the Faculty of Informatics and Statistics of the University of Economics in Prague. The social network analysis tools available in the free R statistical software are used, making it a suitable case study for an exercise in classes.

After having discussed the motivation in the introduction, the remaining parts of this paper briefly introduce the Data Science as the new trend in analytic practice and teaching, some of the tools for social network analysis are then recalled and then the network analysis is reported, followed by some conclusions.

Key words: social network analysis, Twitter, R software

JEL Code: C80, C02

Introduction

A large amount of data is produced and collected in many industry fields and many modern jobs are based on data analysis. These facts should be reflected in the training of professionals in mathematics, data analysis and statistics. And many top-class universities indeed have already started including the Data Science courses in the curriculum. There are textbooks like (Nolan and Temple Lang, 2015) from UC Berkeley, (Guttag, 2013) from MIT and many others and there are initiatives focused at including computational thinking in the curriculum, see e. g. (Computer Based Math, 2019).

Social network analysis like the analysis of the Twitter network is an attractive topic for many students of data analysis and statistics courses these days. The students often ask questions like “How to get and analyse the Twitter data” in classes. The social network

analysis provides a nice opportunity to introduce some useful concepts from data science and graph theory in the classes. A basic exploratory data analysis of the Twitter followers network is reported in this paper. The analysis addresses the network of followers of the Faculty of Informatics and Statistics of the University of Economics in Prague. The social network analysis tools available in the free R statistical software are used, making it a suitable case study for an exercise in classes.

The social network analysis is often reported in papers, see e.g. (Shields, 2015), (Dikmen, 2018), (Junco et al., 2010), (Kousha et al., 2012) for social network analysis in education framework or (Said et al., 2008).

After having discussed the motivation in the introduction, the remaining parts of this paper briefly introduce the Data Science as the new trend in analytic practice and teaching, some of the tools for social network analysis are then recalled and then the network analysis is reported, followed by some conclusions.

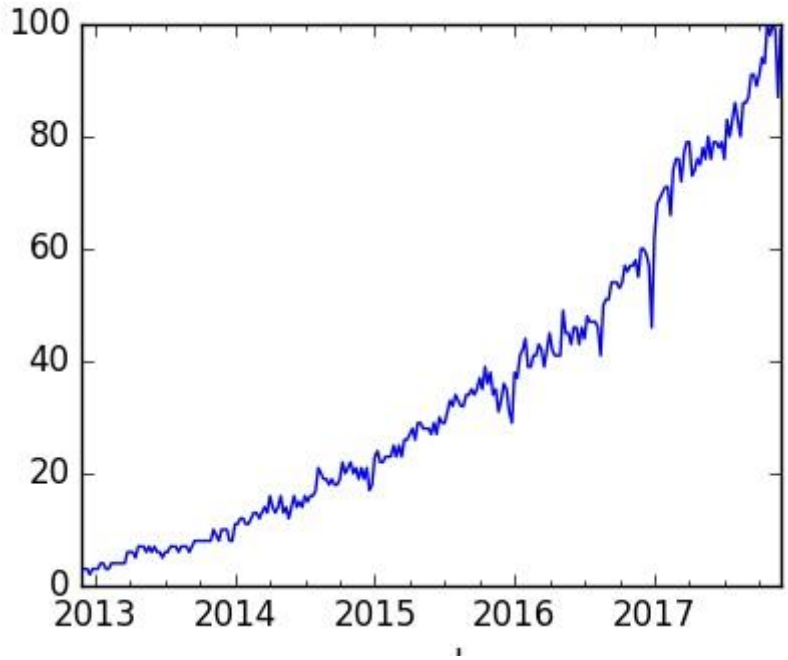
1 Data Science as the recent trend in analytical practice and education

There is the paper by John M. Chambers (Chambers, 1993), the author of the S programming language, the language which can be considered the predecessor of the widely used R software for the statistical computing (R Core Team, 2019). The paper (Chambers, 1993) is advocating the so-called greater statistics, which means everything related to learning from data – starting with the planning of the research, including collection of the data and organization of the data, to reporting the results. The paper (Chambers, 1993) was published as early as in 1993, and it was probably not the first and definitely not the last such paper. The Data Science has recently become popular as a new discipline related to data analysis. This discipline, actually developing the ideas in (Chambers, 1993), aims at the ability to efficiently operate with the data technologies and software tools, ask the right questions about the problem to be analysed and present the results appropriately.

It is rather difficult to find an objective way of evaluating the popularity of some concept. We try using the Google Trends tool (Google Trends, 2019) to analyse the "Data Science" search query popularity in recent years. The Figure 1 is the visualization of the data obtained from Google Trends and it represents the relative interest in the search query at the Google search engine in comparison to the highest value in the time series (the value 100 represents the highest popularity). Obviously "Data Science" has been gaining popularity quite rapidly in the recent years.

The Data Science concept, which has recently started to be used much more often than earlier, may be still finding its precise meaning, but often it refers to, see (Nolan and Temple Lang, 2015) parallel computing, dealing with non-standard data formats (such as access log files, e-mail messages), data technologies like XML, JSON, web scraping, NoSQL database technologies (including Map Reduce, Key - value database systems or graph database systems) and others, all aimed at dealing the data efficiently and ability to solve complex real world problems using data analytics.

Fig. 1: Data Science search term in Google Trends



Source: Google Trends

2 Social Networks Analysis

Social network analysis (SNA) has been one of the major research tools used in sociology and social psychology for a long time. Introduction of a sociogram as early as in 1930's marked the beginning of sociometry. A concept of opinion leaders and other concepts used within a social network analysis framework are still of major importance even nowadays, when the number of participants in various web-based online networking communities is increasing and SNA is also used to support business decisions in finance or telecommunication companies, among others. SNA provides methods for description of structure of relations in a group of interest and may be also used for identification of most popular or influential actors. SNA

methods are often combined with other data analysis tools, such as text mining). SNA may aim at one of the following goals:

- find important characteristics and interesting properties of the network, learn what makes this network different from other networks
- identify the key actors in the network
- investigate the dynamics of the network, i. e. get to know something about how the network evolves in time.

For example, we may take a group of people attending a particular concert and if two people have ever met before, there is a tie between them. In this case, the resulting network representation would be an undirected graph. Or we can take a group of students attending a particular course and place an edge between two actors if one of them would choose the other to seek an advice regarding the subject if needed. In this case, it could make sense to use directed graph to distinguish the roles in the pair (giving an advice - asking for an advice).

2.1 Some concepts in SNA

We briefly discuss the standard and mostly descriptive tools and concepts of the social network analysis, and we refer the reader e.g. to (Butts, 2008) for further details. In this section we recall definitions of selected concepts used within SNA framework first. A graph G is an ordered pair G= (V, E), where V is the set of vertices (also called nodes or actors) and E is the set of edges (edge is a tie between two nodes). The edges may be either directed (we distinguish the source node and the receiving node) or undirected (the order of nodes in the pair does not matter). Basic graph level characteristics of social networks include density, connected-ness, reciprocity and transitivity.

The density of a graph refers to the number of edges in the graph expressed as a proportion of the maximum possible number of edges. It is natural that larger social networks have lower density, because the number of possible edges increases rapidly with the number of nodes in a graph and at the same time the number of connections which each person can maintain is usually limited. Another social network characteristic is its reciprocity. Edgewise reciprocity is the proportion of edges which are reciprocated. Degree is a node-level characteristic. Indegree is the number of nodes that are adjacent to particular node, outdegree of a node is number of nodes adjacent from the node, (total) degree of a node refers to number of nodes adjacent from or to the node. A centrality is a node-level characteristic, as opposed to centralization, which is a graph-level property. Most often used centrality measures c refer to the degree and betweenness. Some of the centrality scores may be used as a measure of

prestige of a node, among others the indegree centrality or the domain centrality, which is indegree centrality within the corresponding reachability graph (Butts, 2010). An interesting measure of a social network is transitivity, which is the fraction of connected triplets of vertices which also form triangles, so it refers to probability that there is a tie between A and C if there is a tie between A and B and a tie between B and C.

An interpretation of selected measures may be as follows. A degree may be interpreted as a measure of activity of a node in the network. Actors with high betweenness scores may be considered to have good control over information flow in the network, as betweenness refers to the number of shortest paths between nodes in the network, which go through the particular node. Actors with high closeness scores (i. e. actors with the low average of geodesic (the shortest path) distances to all other nodes) have good access to other members of the network. Note that the closeness in its standard form is defined only for connected graphs (i.e. graphs with no isolated nodes).

The data analysis may be performed using the igraph or sna package for social network analysis (Butts, 2008) and the tidyverse suite of packages (Wickham and Grolemund, 2017) for the R environment for statistical computing (R Core Team, 2019).

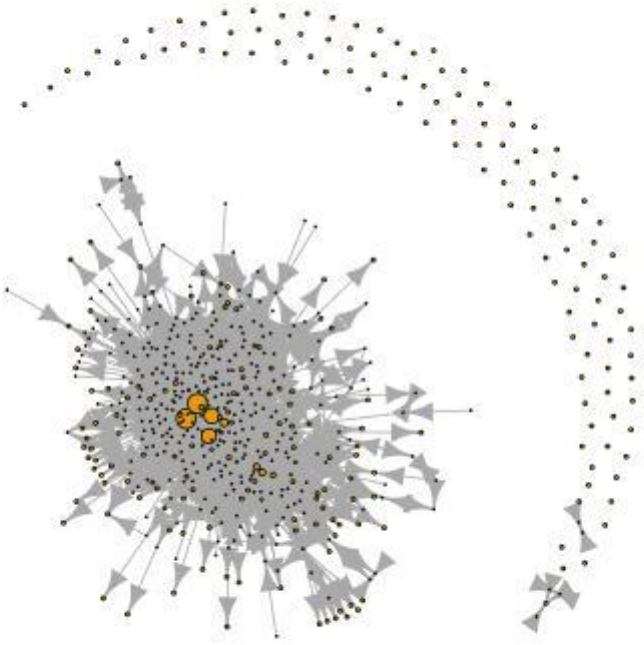
3 FIS VŠE Twitter Network Analysis

We analyse the network of followers of the Faculty of Informatics and Statistics of the University of Economics in Prague (FIS VŠE). That is the graph is built using the set of followers of FIS VŠE as the set of vertices and the set of follower relationships as the edges between the actors in this set. It makes sense to distinguish between the source and the receiving node of a vertex, i.e. we are going to work with directed edges. We use an edge from A to B if A is followed by B. It is possible to download the set of (artificial identifiers of) the followers of a user (e.g. FIS VŠE) from Twitter. It is then possible to get the followers for each of the actors in this set. Then it is a nice Data Science exercise to build the network. This exercise among other steps includes dropping the edges leading to actors which are outside the group. Especially for larger networks, an efficient building and analysis of the graph may become quite challenging.

Due to the way how the data for building the network may be retrieved from the source website, considering that the network may be changing (i.e. the edges may be added or dropped nearly any time, even while the data transfer is in process) it cannot be guaranteed that one gets 100% correct snapshot of all the graph vertices and edges valid at a particular

time. But for most analytic purposes, just a near-correct picture is usually enough. As of April 2019, this network has 605 nodes and some 2400 edges. The graph density value is around 0.006, the reciprocity is 0.39 which is quite high, and the transitivity is 0.12. The network has over 100 components. The size of the largest component is 487, so it is a component, covering majority of the actors. The remaining components are just small groups or isolated actors. For the network layout, see Figure 2, where the vertex size of actors is proportional to the number of followers.

Fig. 2: FIS VŠE followers network



Source: own work

The basic statistics for the followers count and for the count of the followed actors (called friends) is shown in Table 1. Clearly the mean value must be the same for both characteristics.

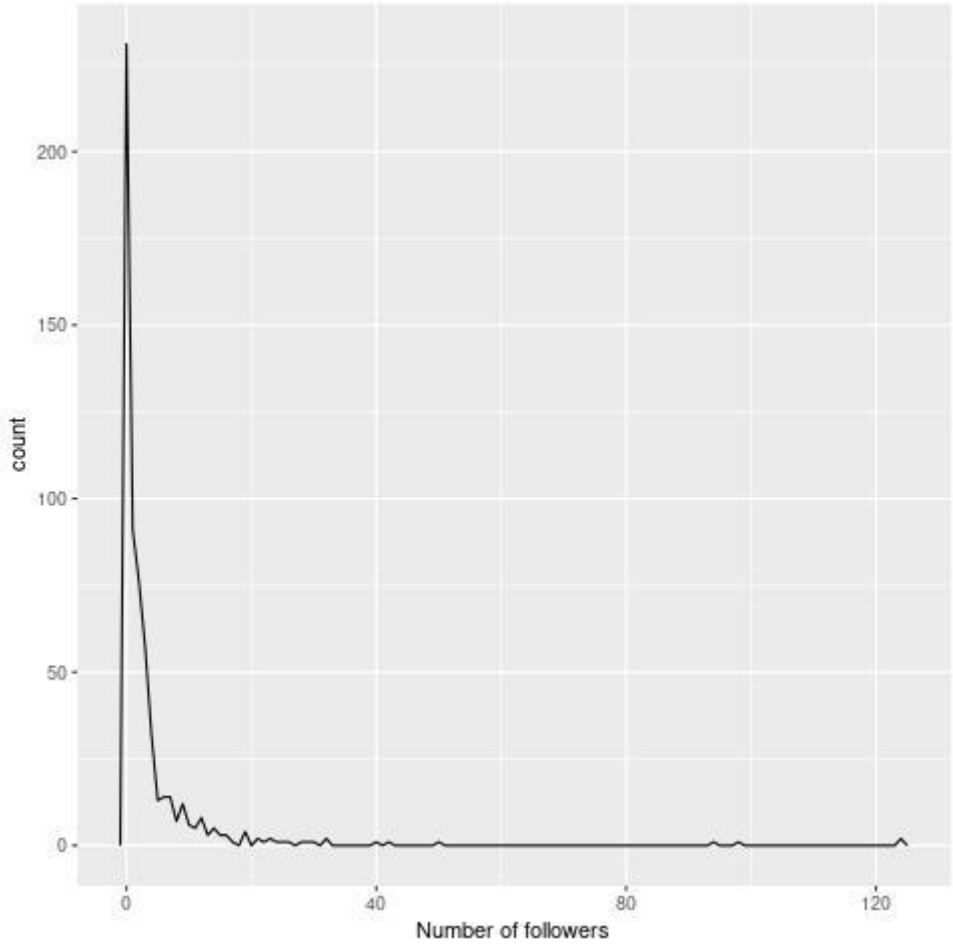
Tab. 1: Followers and friends count statistics

	Min	1st Q.	Median	Mean	3rd Q.	Max
Followers count	0	0	1	3.91	3	124
Friends count	0	0	2	3.91	5	128

Source: Own work

The Figure 3 then shows the frequency plot of the followers count. The distribution of the friends count in the network is similar. The distribution of the indegree and outdegree in the network may be considered a rather common distribution in social groups.

Fig. 3: Followers counts

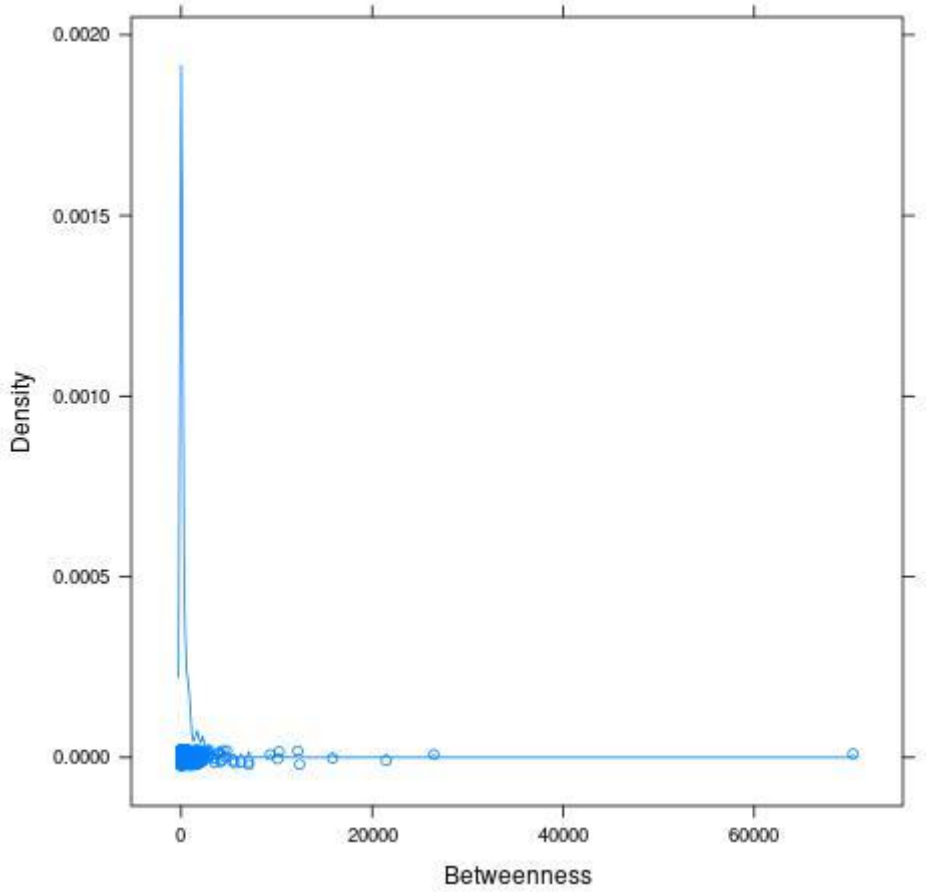


Source: own work

Further analysis has been performed just using the largest component of the network. The density plots of the betweenness scores are shown in Figure 4. It may be noted that one of

the actors has rather extreme high value of the betweenness score. Regarding the community structure in the network, the community detection analysis has shown that there does not seem to be any strong community structure in the network.

Fig. 4: Betweenness scores



Source: own work

Conclusion

The basic data analysis of FIS VŠE Twitter followers network has been reported, showing some of the elementary social network analysis tools. The process of building the graph and preparing the data for the analysis provides an attractive example for discussing some of the Data Science techniques in classes.

References

Butts, C. T. (2008). Social Network Analysis with sna. *Journal of Statistical Soft-ware*, *24*(6).

Dikmen, E. S. (2018). Video Sharinc Strategies of Higher Ecucation Institutions: A research on Universities YouTube Channels in Turkey. ILEF DERGISI, 5(2), 29-52.

Computer Based Math (2019). Retrieved April 22, 2012, from http://www.computerbasedmath.org.

Google Trends. (2019). Retrieved April 22, 2012, from https://trends.google.com.

Guttag, J. V. (2013). *Introduction to computation and programming using Python.* The MIT Press.

Junco, R., Heiberger, G., & Loken, E. (2010). The effect of Twitter on college student engagement and grades. *Journal of Computer Assisted Learning, 27*(2), 119-132. doi:10.1111/j.1365-2729.2010.00387.x

Kousha, K., Thelwall, M., & Abdoli, M. (2012). The role of online videos in research communication: A content analysis of YouTube videos cited in academic publications. *Journal of the American Society for Information Science and Technology, 63*(9), 1710-1727. doi:10.1002/asi.22717

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Said, Y. H., Wegman, E. J., Sharabati, W. K., & Rigsby, J. T. (2008). Social networks of author–coauthor relationships. *Computational Statistics & Data Analysis, 52*(4), 2177-2184. doi:10.1016/j.csda.2007.07.021

Shields, R. (2015). Following the leader? Network models of "world-class" universities on Twitter. Higher Education, 71(2), 253-268. doi:10.1007/s10734-015-9900-z

Temple Lang, D. T., & Nolan, D. A. (2015). *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving.* CRC Press.

Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data*. Beijing: OReilly.

Contact

Nikola Kaspříková

University of Economics in Prague, Department of Mathematics

Nám. W. Churchilla 4, 130 67 Praha

nb33@tulipany.cz