

INTRODUCTION TO COMPLEX SYSTEMS PROBLEM SHEET 1

REPRESENTING DATA AS A NETWORK, 2019/2020

A very good book for this chapter is the book *Social Network Analysis for Startups* by Maksim Tsvetovat and Alexander Kouznetsov, first published in 2011. You can learn from this book how to analyze social media data and where these kind of data can be found. You can also learn more about NetworkX environment (Python), which will be very useful tool during this course.

1. Visit the following sites and check what kind of data can you acquire and what kind of visualization can you make:

- <http://networkrepository.com/>
- <http://snap.stanford.edu>
- <http://www-personal.umich.edu/~mejn/netdata/>

Download a chosen dataset and draw a network.

2. Networks are often used to represent collaboration between researchers. Using Scopus database (<https://www.scopus.com>) create a database for researchers working in the Department of Computer Science WPPT, PWR and in the Department of Theoretical Physics WPPT, PWR, something similar that can be found at <http://www.kft.pwr.edu.pl/?Publications> (short database for KFT). In fact you can start with this short database for KFT for simplicity. Create a network: node = researcher, link between two researchers if they have a common paper. Plot the network.

3. We will analyze the book by Lewis Carroll *Alice's Adventures in Wonderland*.

Visit <https://www.gutenberg.org/browse/scores/top> and download a book in an appropriate format. Among others pdf and txt versions are available. Conduct two types of analysis:

- (a) Classical Zipf analysis, checking the frequency distribution of words. This method was presented during the Introduction to Complex Systems: (1) compute the frequency of each word in the book, (2) sort them by frequency and give ranks – the most frequent word should have rank 1, etc. (3) plot the dependence between frequency and rank in 2 scales: linear and log-log scale, (4) find an exponent that describes a power-law for this book.
- (b) Present a text as a network. Words will be nodes and relations between them will be represented by links. There are several possibilities to do this. Here we use the simple idea of undirected, unweighted network: two words will be linked if they appear next to each other (word-adjacency technique). Do you have any other idea? Plot this network and the distribution of the degree of nodes. List 10 nodes (words) with the highest degree and 10 with the lowest.