Data for
"Analysis of the Wikipedia Network of Mathematicians"

Bingsheng Chen, Zhengyu Lin, Tim S. Evans

Centre for Complexity Science, and Theoretical Physics Group,
Imperial College London, SW7 2AZ, U.K.

The data is taken from three snapshots of Wikipedia, taken in 2013, 2017 and 2018. Filenames indicate which data set has been used by incoporating 13, 17 or 18 into the name.

- `README.tex`, `README.pdf`

  This README file in orginal LaTeXformat and in readable pdf format with hyperlinks.

- `graph13(LCC).graphml`, `graph17(LCC).graphml`, `graph18(LCC).graphml`

  These are the networks of the largest connected components (LCC) for each year. The `graphml` format is used by the Visone network package amongst others, and can be produced by the python `networkx` library.

- `score13.xlsx`, `score17.xlsx`, `score18.xlsx`

  These contain information on the nodes, i.e. on all the mathematicians in each data set. The name is given along with other measures such as centrality measures.

  For the 2017 data set, each centrality measure has three columns, the column labelled "original" is the measurement on the original data, columns with "mean" and "std" in the name at the top of the column indicate the mean and standard deviation when the data is partially randomised as described in the paper.

- `Survey_Count.txt`

  A text files of the results from the student survey.

- The subdirectory `RawData` contains the following:

  - `Wikipedia-20131126132847_LM.xml`,
    `Wikipedia-2017_LM.xml`,
    `Wikipedia-20180922143921_LM.xml`

    These `LM` files contain the list of mathematicians. These are the 26 pages given on Wikipedia but they have been spliced together into a single xml file. It is processing of this file that gives us our list of mathematicians as defined by the crowd of editors who produced these pages. The entries for each mathematician contain the name of the Wikipedia webpage with the appropriate biography and we used that to extract that page from Wikipedia.

  - `Wikipedia-20131126132847_LMmathematicianslist.dat`,
    `Wikipedia-2017_LMmathematicianslist.dat`,
    `Wikipedia-20180922143921_LMmathematicianslist.dat`

    These are simple text files containing the List of Mathematicians (LM). The names used are those used within Wikipedia for the web page of the biography and contain spaces (underscores in the URL). These are the names extracted from the 26 pages given on Wikipedia.

– `Wikipedia-20131127_MathematiciansAll.xml`,
 `Wikipedia-2017_MathematiciansAll.xml`,
 `Wikipedia-20180922143921_MathematiciansAll.xml`

 Each file is an xml files containing all the biographies for one of the three years.

– `Wikipedia-20131127_MathematiciansAlledgelist.dat`,
 `Wikipedia-2017_MathematiciansAlledgelist.dat`,
 `Wikipedia-20180922143921_MathematiciansAlledgelist.dat`

 Each file is a list of every hyperlinks from one biography to another in one year of our data. Two columns, source node then target node separated by a tab character. The nodes are given by the Wikipedia name of the relevant page, a name which includes spaces.

 Each file is an xml files containing all the biographies for one of the three years.