SNAP for C++
SNAP for Python
SNAP Datasets
BIOSNAP Datasets
What's new
People
Papers
Projects
Citing SNAP
Links
About
Contact us

Open positions

Open research positions in **SNAP** group are available at **undergraduate, graduate** and **postdoctoral** levels.

Higgs Twitter Dataset

Dataset information

The Higgs dataset has been built after monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012. The messages posted in Twitter about this discovery between 1st and 7th July 2012 are considered.

The four directional networks made available here have been extracted from user activities in Twitter as:

1. re-tweeting (retweet network)
2. replying (reply network) to existing tweets
3. mentioning (mention network) other users
4. friends/followers social relationships among user involved in the above activities
5. information about activity on Twitter during the discovery of Higgs boson

It is worth remarking that the user IDs have been anonimized, and the same user ID is used for all networks. This choice allows to use the Higgs dataset in studies about large-scale interdependent/interconnected multiplex/multilayer networks, where one layer accounts for the social structure and three layers encode different types of user dynamics .

Note that this dataset has been updated on **Mar 31 2015**. If you downloaded a previous version, please update it, results could differ.

For more information about data collection, please refer to our paper.

Dataset statistics are calculated for the graph with the highest number of nodes and edges:

Social Network statistics	
Nodes	456626
Edges	14855842
Nodes in largest WCC	456290 (0.999)
Edges in largest WCC	14855466 (1.000)
Nodes in largest SCC	360210 (0.789)
Edges in largest SCC	14102605 (0.949)
Average clustering coefficient	0.1887
Number of triangles	83023401
Fraction of closed triangles	0.002901
Diameter (longest shortest path)	9
90-percentile effective diameter	3.7
Retweet Network statistics	
Nodes	256491
Edges	328132
Nodes in largest WCC	223833 (0.873)
Edges in largest WCC	308596 (0.940)
Nodes in largest SCC	984 (0.004)
Edges in largest SCC	3850 (0.012)
Average clustering coefficient	0.0156
Number of triangles	21172
Fraction of closed triangles	0.0001085
Diameter (longest shortest path)	19
90-percentile effective diameter	6.8
Reply Network statistics	
Nodes	38918
Edges	32523
Nodes in largest WCC	12839 (0.330)
Edges in largest WCC	14944 (0.459)
Nodes in largest SCC	322 (0.008)
Edges in largest SCC	708 (0.022)
Average clustering coefficient	0.0058
Number of triangles	244

Fraction of closed triangles	0.0001561
Diameter (longest shortest path)	29
90-percentile effective diameter	10
Mention Network statistics	
Nodes	116408
Edges	150818
Nodes in largest WCC	91606 (0.787)
Edges in largest WCC	132068 (0.876)
Nodes in largest SCC	1801 (0.015)
Edges in largest SCC	7069 (0.047)
Average clustering coefficient	0.0825
Number of triangles	23068
Fraction of closed triangles	0.0002417
Diameter (longest shortest path)	18
90-percentile effective diameter	6.5

Data format - higgs-activity_time.txt

```
userA userB timestamp interaction
```

Interaction can be **RT** (retweet), **MT** (mention) or **RE** (reply). Each link is directed. The user IDs in this dataset corresponds to the ones adopted to anonymize the social structure, thus the datasets (1) - (5) can be used together for complex analysis involving structure and dynamics.

Note 1: the direction of links depends on the application, in general. For instance, if one is interested in building a network of how information flows, then the direction of RT should be reversed when used in the analysis. Nevertheless, the choice is left to the researcher and his/her own interpretation of the data, whereas we just provide the observed actions, i.e., who retweets/mentions/replies/follows whom.

Note 2: users mentioned in retweeted tweets are considered as mentions. For instance, if @A retweets the tweet “hello @C @D" sent by @B, then the following links are created: @A @B timeX RT, @A @C timeX MT, @A @D timeX MT, because @C and @D can be notified that they have been mentioned in a retweet. Similarly in the case of a reply. If for some reason the researcher does not agree with this choice, he/she can easily identify this type of links and remove the mentions, for instance.

Source (citation)

- M. De Domenico, A. Lima, P. Mougel and M. Musolesi. The Anatomy of a Scientific Rumor. (Nature Open Access) Scientific Reports 3, 2980 (2013).

Files

File	Description
social_network.edgelist.gz	Friends/follower graph (directed)
retweet_network.edgelist.gz	Graph of who retweets whom (directed and weighted)
reply_network.edgelist.gz	Graph of who replies to who (directed and weighted)
mention_network.edgelist.gz	Graph of who mentions whom (directed and weighted)
higgs-activity_time.txt.gz	The dataset provides information about activity on Twitter during the discovery of Higgs boson