



Tutorial: Large Scale Network Analytics with SNAP

<http://snap.stanford.edu/proj/snap-icwsm>

Rok Sosič, Jure Leskovec
Stanford University

ICWSM-14, Ann Arbor, MI

June, 2014





SNAP Network Datasets

Rok Sosič, Jure Leskovec
Stanford University

SNAP Network Datasets

- <http://snap.stanford.edu/data/>
- **Public collection of large network datasets**
 - Over 15 network types
 - Over 70 datasets
 - Varying sizes from 20K up to 1.8B edges
- **Popular resource for network scientists**
 - Method development, study, benchmarking
- **Contribute your dataset**
 - We welcome new additions
- **SNAP Dataset Users mailing list**
<http://groups.google.com/group/snap-datasets>

Datasets in SNAP (1)

- **Social networks**
 - Online social networks, edges represent interactions between users
- **Location-based online social networks**
 - Social networks with geographic check-ins
- **Online communities**
 - Data from online communities such as Reddit and Flickr
- **Networks with ground-truth communities**
 - Ground-truth network communities in social/information networks
- **Online reviews**
 - Data from online review systems such as Amazon
- **Amazon networks**
 - Nodes represent products, edges link co-purchased products
- <http://snap.stanford.edu/data/>

Datasets in SNAP (2)

- **Twitter and Memetracker**
 - Memetracker phrases, links and 467 million Tweets
- **Signed networks**
 - Networks with positive and negative edges (friend/foe, trust/distrust)
- **Communication networks**
 - Email communication networks with edges representing emails
- **Wikipedia networks and metadata**
 - Talk, editing and voting data from Wikipedia
- **Citation networks**
 - Nodes represent papers, edges represent citations
- **Collaboration networks**
 - Nodes represent scientists, edges represent collaboration (paper co-authoring)
- <http://snap.stanford.edu/data/>

Datasets in SNAP (3)

- **Web graphs**
 - Nodes represent webpages and edges are hyperlinks
- **Internet networks**
 - Nodes represent computers and edges communication
- **Autonomous systems**
 - Graphs of the internet
- **Road networks**
 - Nodes represent intersections and edges roads connecting the intersections
- <http://snap.stanford.edu/data/>

Social Circles from Facebook

- **Friends lists from Facebook**
 - Includes user profiles, circles, ego networks
 - Collected via Social Circles App on Facebook
 - **Contribute your own social circles:**
<http://snap.stanford.edu/socialcircles/>
 - **Social circle detection Kaggle competition:**
 - <https://www.kaggle.com/c/learning-social-circles>

<http://snap.stanford.edu/data/egonets-Facebook.html>

| Dataset statistics | |
|----------------------------------|---------------|
| Nodes | 4039 |
| Edges | 88234 |
| Nodes in largest WCC | 4039 (1.000) |
| Edges in largest WCC | 88234 (1.000) |
| Nodes in largest SCC | 4039 (1.000) |
| Edges in largest SCC | 88234 (1.000) |
| Average clustering coefficient | 0.6055 |
| Number of triangles | 1612010 |
| Fraction of closed triangles | 0.2647 |
| Diameter (longest shortest path) | 8 |
| 90-percentile effective diameter | 4.7 |

Location Based Social Networks

Friendship network and check-ins in Gowalla location-based social network

| Dataset statistics | |
|----------------------------------|----------------|
| Nodes | 196591 |
| Edges | 950327 |
| Nodes in largest WCC | 196591 (1.000) |
| Edges in largest WCC | 950327 (1.000) |
| Nodes in largest SCC | 196591 (1.000) |
| Edges in largest SCC | 950327 (1.000) |
| Average clustering coefficient | 0.2367 |
| Number of triangles | 2273138 |
| Fraction of closed triangles | 0.007952 |
| Diameter (longest shortest path) | 14 |
| 90-percentile effective diameter | 5.7 |
| Check-ins | 6,442,890 |

<http://snap.stanford.edu/data/loc-gowalla.html>

Online Communities: Reddit

- **Post submissions to Reddit**
 - Includes an image with multiple submissions
 - **Features per posts:** number of ratings, the title, number of comments

| Dataset statistics | |
|---|----------------------|
| Number of submissions | 132,308 |
| Number of unique images | 16,736 |
| Average number of times an image is resubmitted | 7.9 |
| Timespan | July 2008 - Jan 2013 |

<http://snap.stanford.edu/data/web-Reddit.html>

Online Reviews: Amazon

- **18 years of Amazon reviews up to March 2013**
 - Product and user information, ratings, review text

| Dataset statistics | |
|--------------------------------|---------------------|
| Number of reviews | 34,686,770 |
| Number of users | 6,643,669 |
| Number of products | 2,441,053 |
| Users with > 50 reviews | 56,772 |
| Median no. of words per review | 82 |
| Timespan | Jun 1995 - Mar 2013 |

<http://snap.stanford.edu/data/web-Amazon.html>