**For Use in Introduction to Social Network Analysis, Northwestern University**

**Please contact** **t@u.northwestern.edu** **with suggestions, corrections or additions!**

**Data Sources for the Social Network Analysis Project**

If you’re interested in using this data beyond this course, you may need IRB approval. Furthermore, some datasets that are available violate the terms of service of the company from which they were harvested; this makes them unusable in publications with ACM and some other academic organizations. Look closely at the dataset and check the rights associated with it before making any grand plans!

*Collect Your Own*

**WARNING:** due to rate-limiting by social networking platforms and limited availability over time (Twitter only gives you about a week of tweets), if you want to collect social network data, you’re probably going to need to start ASAP and collect data for a long period of time!

**NodeXL**- [NodeXL](http://nodexl.codeplex.com/) is a plug-in for the Windows version of Microsoft Excel. It is a powerful tool for calculating descriptive statistics and creating visualizations of network datasets. It also has the built-in ability to collect datasets from Twitter hashtags, Facebook groups, and more.

**twarc** - [Twarc](https://github.com/edsu/twarc) is an open source Python library that can collect Twitter data through the API. You’ll need to know how to convert network info from JSON to some other format for your analysis.

**NetVizz -** A [Facebook plug-in](https://apps.facebook.com/netvizz/?fb_source=search&ref=br_tf) that scrapes posts from Facebook Groups, Pages, and User profiles. It no longer collects friend-of-friend data, and has been on the fritz recently. Beware: it might not be there when you need it!

**APIs** - if you’re adventurous and can code, lots of social network datasets have APIs that you can pull data from directly. This approach is admittedly not for the faint of heart.

Here’s [a post](http://datascience.stackexchange.com/questions/422/publicly-available-social-network-datasets-apis) with information about some APIs.

[SOCSCIBOT](http://socscibot.wlv.ac.uk/) - a web crawler for the social sciences.

**VOSON -** talk to your TAs about how long your access to VOSON works.

[SOCILAB](http://socilab.com/) - download your linkedIn dataset.

[Lightbeam](https://www.mozilla.org/en-US/lightbeam/) or [Disconnect](https://chrome.google.com/webstore/detail/disconnect/jeoacafpbcihiomhlakheieifhpjdfeo?hl=en) - plug-ins that track websites that track you.

[BigBang](https://github.com/sbenthall/bigbang) - a package developed by Berkeley iSchool students to turn listserv interactions into network datasets.

*Pre-Assembled Datasets*

[University of Southampton](http://web-001.ecs.soton.ac.uk/) - lots of great collected datasets from Twitter, Wikipedia and more. Some sources require asking for permission.

[ICPSR](http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies?q=network&searchSource=find-analyze-home&sortBy=) - you have to dig around, but there’s some really interesting data in this international consortium for social science research, of which Northwestern is a part.

[KONECT](http://konect.uni-koblenz.de/) - large collection of datasets with longitudinal data, edge weights, directed/undirected, etc. Record labels, animal networks, club membership, infrastructure, citation networks, bio networks...

[**Stanford Large Network Dataset Collection**](https://snap.stanford.edu/data/index.html) - lots of great networks, often quite large. Citation networks, social networks, Amazon recommendation networks, product reviews, Internet networks, Google search networks.

[UC Irvine Network Data Repository](https://networkdata.ics.uci.edu/index.html) - all kinds of stuff. HAM radio networks, southern ladies, dining table partners at a school, CEO and club co-memberships, food webs.

[Christian Sommer’s datasets](http://www.sommer.jp/graphs/) - router info, web graphs, citation networks, protein interactions (!)

[Facebook datasets](http://datalinks.wikia.com/wiki/Facebook_Social_Network_for_100_Universities) - dataset of 100 collegiate facebook friendship networks.

 More information about that dataset is available [here](http://sociograph.blogspot.de/2011/03/facebook100-data-and-parser-for-it.html).

[Roldan Pozo](http://math.nist.gov/~RPozo/complex_datasets.html) - links to other datasets, Erdos collaborations, email networks, Sept. 11th news stories, thesaurus, corporate ownership

[NodeXL data gallery](https://nodexlgraphgallery.org/Pages/Default.aspx) - data collected through NodeXL, available for your use.

[U.S. film citation network](http://amaral-lab.org/resources/data-sets/us-film-citation-network) - Citations between films produced in the U.S.; description of data collection at website.

[Mark Newman](http://math.nist.gov/~RPozo/complex_datasets.html) - links to other datasets and datasets created by Newman. Karate clubs, neural networks, power grids, dolphin social networks, american football, book citations,

[Jake Hofman’s delicious links](https://delicious.com/jhofman/networks%2Cdata) - network researcher keeps track of interesting datasets.

[Laboratory for Web Algorithmics](http://law.di.unimi.it/datasets.php) - very large crawls of major websites such as Wikipedia, Facebook, etc.

[Bitcoin dataset](http://www.vo.elte.hu/bitcoin/) - Bitcoin transactions! anonymous but still interesting.

[epinions dataset](http://www.public.asu.edu/~jtang20/datasetcode/truststudy.htm) - info on trust over time within the rating community. You may need to contact the author for the full dataset.

[ASU’s data repository](http://socialcomputing.asu.edu/pages/datasets) - last.fm, livejournal, youtube, blogs, flickr, digg, friendster

[Early Youtube Datasets](http://netsg.cs.sfu.ca/youtubedata/) - Simon Fraser University.

[UCINET Data](http://vlado.fmf.uni-lj.si/pub/networks/data/Ucinet/UciData.htm)- generally featuring digitized versions of classic sociology studies.

[Longitudinal data from Stanford](http://web.stanford.edu/group/sonia/dataSources/index.html) - Debian projects, data of interaction in classrooms,

[InfoVis](http://iv.slis.indiana.edu/db/index.html) - citation datasets, patent and trademark data.

[Network-Science.org](http://www.network-science.org/) - a list of lists of network data

[Early Facebook New Orleans data set](http://socialnetworks.mpi-sws.org/data-wosn2009.html)

[Network Repository](http://networkrepository.com/) - a collection of networks of many different types.

[StackExchange data set list](http://datascience.stackexchange.com/questions/155/publicly-available-datasets) -- most are not in network format, but many could be converted to a network format.

[StackExchange Open Data](http://opendata.stackexchange.com/search?q=network&s=27af249e-d64e-4c10-9ce9-d018729c0a2e), search for “network” - you never know, there might be gems in here!

[Quora post](http://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public) - harnessing the power of networks to find more networks!

[data.gov](http://www.data.gov/) - generally not in network format, but lots of stuff could be converted to network data here.

[KDNuggets](http://www.kdnuggets.com/datasets/index.html) - Also generally not in network format, but could be converted.