**Networks of Political Communication**

‘Political Polarization on Twitter’

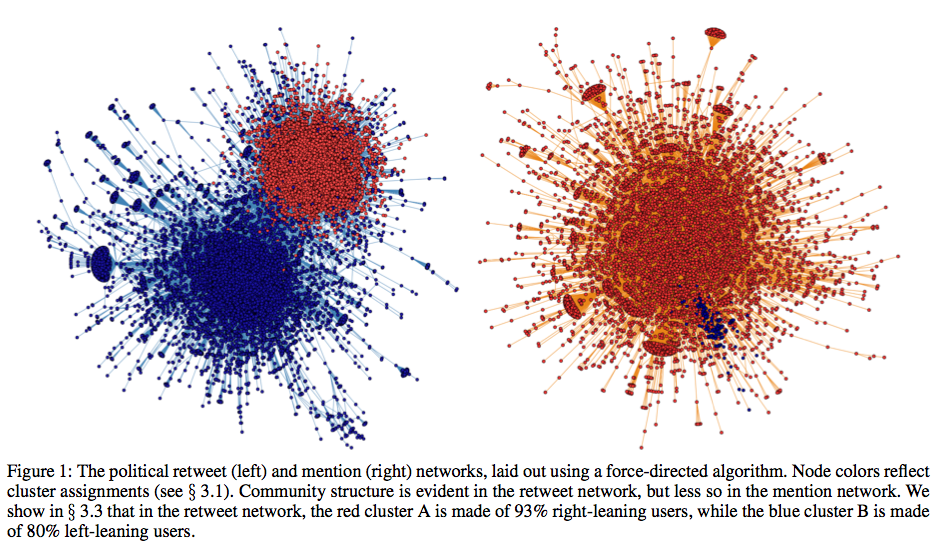
Center for Complex Networks and Systems Research

Indiana University

**Overview**

This dataset describes three networks of political communication between users of the Twitter social media platform in the six weeks prior to the 2010 Congressional midterm elections. This network is particularly interesting because one mode of communication, retweets, segregates users into two politically homogeneous communities of likeminded individuals, while mentions form a bridge between the two communities over which users are exposed to people and information they would not likely select ahead of time.

We encourage you to read the associated article, ‘[Political Polarization on Twitter](http://www.criticalinsight.net/publications/icwsm2011_polarization.pdf)’[[1]](#footnote-1), by Conover et al., for a more detailed discussion of the technical and sociological features of this communication network. A brief excerpt from the article, describing the provenance and structure of the data appears below, followed by a detailed technical specification of the available data formats.



**Data Source**

The present analysis leverages data collected from the Twitter ‘gardenhose’ API (dev.twitter.com/pages/ streaming\_api) between September 14th and November 1st, 2010 — the run-up to the November 4th U.S. congressional midterm elections. During the six weeks of data collection we observed approximately 355 million tweets.

**Identifying Political Content**

Let us define a political communication as any tweet containing at least one politically relevant hashtag. To identify an appropriate set of political hashtags and to avoid intro- ducing bias into the sample, we performed a simple tag co-occurrence discovery procedure. We began by seeding our sample with the two most popular political hashtags, #p2 (“Progressives 2.0”) and #tcot (“Top Conservatives on Twitter”). For each seed we identified the set of hashtags with which it co-occurred in at least one tweet, and ranked the results using the Jaccard coefficient. For a set of tweets S containing a seed hashtag, and a set of tweets T containing another hashtag, the Jaccard coefficient between S and T is:

σ(S,T)= |S∩T| / |S ∪ T|

Thus, when the tweets in which both seed and hashtag occur make up a large portion of the tweets in which either occurs, the two are deemed to be related.

**Political Communication Networks**

From the tweets containing any of the politically relevant hashtags we constructed networks representing political communication among Twitter users. Focusing on the two primary modes of public user-user interaction, mentions and retweets, we define communication links in the following ways. In the retweet network an edge runs from a node representing user A to a node representing user B if B retweets content originally broadcast by A, indicating that information has propagated from A to B. In the mention network an edge runs from A to B if A mentions B in a tweet, indicating that information may have propagated from A to B (a tweet mentioning B is visible in B’s timeline). Both networks therefore represent potential pathways for information to flow between users.

The retweet network consists of 23,766 non-isolated nodes among a total of 45,365. The largest connected component accounts for 18,470 nodes, with 102 nodes in the next-largest component. The mention network is smaller, consisting of 10,142 non-isolated nodes out of 17,752 total. It has 7,175 nodes in its largest connected component, and 119 in the next-largest. Because of their dominance we focus on the largest connected components for the rest of our analysis.

**DATA SPECIFICATION**

**Network Types**

There are three versions of the dataset, each specifying a network representing different types of tweets.

* **retweet**: Contains only directed retweet edges, where an edge runs from A to B if B retweets content produced by A.
* **mention**: Contains only directed reply edges, where an edge runs from A to B if A mentions B in a tweet.
* **all**: Contains both types of edges listed above.

**Edge Attributes**

* time: Unix epoch timestamp for the associated tweet, by local timezone.
* type: Specifies the type of tweet, either retweet or mention.
* urls: A count of the number of hyperlinks appearing in the tweet.
* tags: a list of the hashtags appearing in the tweet.

**Node Attributes**

* cluster: Specifies which of the two retweet network clusters the node appeared in, based on assignments using the method from [1]. Values can be ‘left’, for users who were in the cluster with a majority of left-leaning users, ‘right’ for users who were in the cluster with a majority of right-leaning users, or ‘-‘ for users who appear in the mention network but not in the retweet network, and thus do not have a cluster assignment.

**Data Formats**

We provide the network data in three formats, an igraph pickle, tab-delimited flat files, and as GraphML.

igraph Pickle

We provide several [Python pickles](http://docs.python.org/library/pickle.html), each containing an [igraph](http://igraph.sourceforge.net/doc/python/index.html) object representing one of the networks. This is the easiest way to get started with the data, as you can load the network into memory and begin manipulating it directly. We refer to the following tutorials [[pickle](http://wiki.python.org/moin/UsingPickle), [igraph](http://www.cs.rhul.ac.uk/home/tamas/development/igraph/tutorial/tutorial.html)] for detailed instructions, and provide the code sample below as a starting point.

import cPickle as pickle

g = pickle.load(open('retweet\_igraph.pickle'))

print "# of Vertices: %s" % len(g.vs)

print "# of Edges: %s" % len(g.es)

# Iterate over edges, printing cluster ids of adj. vertices and tags contained in tweet

for e in g.es()[:10]:

print "%s --> %s\t| %s" % (g.vs[e.source]['cluster'], g.vs[e.target]['cluster'], e['tags'])

igraph can write to many different graph formats, so if you need something we don’t provide here, check the igraph documentation for output specifications.

Flat Files

For a network with type xxx, the text-only description of the graph depends on a pair of tab-delimited files, xxx.nodes, and xxx.edgelist.

*xxx.nodes* specifies the attributes and indices of nodes, and contains two columns:

* 0: node index
* 1: cluster id

*xxx.edgelist* specifies pairs of adjacent nodes connected by a directed edge, in addition to several attributes relating to the type and content of the tweet the edge represents. The format contains five fixed columns and a variable-length trailing list of hashtags.

* 0: source node index
* 1: target node index
* 2: edge type
* 3: number of hyperlinks contained in the tweet
* 4: unix epoch timestamp (time at origin)
* 5-N: tab-delimited list of hashtags contained in the twee

GraphML

We also provide [GraphML](http://graphml.graphdrawing.org/) versions of all the networks which can be loaded into the [Gephi](http://gephi.org/) network visualization tool and explored directly.

**Conditions of Use**

This data is licensed under the Creative Commons Attribution-ShareALike 3.0 ([CC BY-SA 3.0](http://creativecommons.org/licenses/by-sa/3.0/us/)) license. You are free to do whatever you’d like with this data as long as you:

* Cite the paper, “Political Polarization on Twitter”, by Conover et al.
* Distribute any derivative works under the same or similar license.

**Related Works**

Conover, M; Goncalves, B; Flammini, A; Menczer, F. 2011. Partisan Asymmetries in Online Political Activity. *Social Networks: Special Issue on Political Networks (invited submission)*

Conover, M.; Ratkiewicz, J.; M. Francisco; Goncalves, B.; Flammini, A.; and Menczer, F. 2011. Political Polarization on Twitter. *Proc of the 5th Intl. Conf on Weblogs and Social Media (ICWSM)*

Ratkiewicz, J.; Conover, M.; Meiss, M.; Goncalves, B.; Patil, S.; Flammini, A.; and Menczer, F. 2011. Truthy: Mapping the spread of astroturf in microblog streams. *Proc. of the 20th Intl. World Wide Web Conf. (WWW).*

Ratkiewicz, J.; Conover, M.; Meiss, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2011. Detecting and Tracking Political Abuse in Social Media. *Proc of the 5th Intl. Conf on Weblogs and Social Media (ICWSM)*

Conover, M.; Ratkiewicz, J.; Gonc alves, B.; Flammini, A.; and Menczer, F. 2011. Visualizing Political Communication on Twitter. *Journal of Information Technology and Politics 2011 (JITP)*

1. Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., Menczer, F. Political Polarization on Twitter, in *Proc. of the 5th Intl. Conf. on Weblogs and Social Media 2011 (ICWSM).* [↑](#footnote-ref-1)