Twitter Data Collection: Crawling Users, Neighbors and Their Communication for Personal Attribute Prediction in Social Media

Svitlana Volkova
Center for Language and Speech Processing
Department of Computer Science
Johns Hopkins University
3400 North Charles, Baltimore MD, 21218 USA

May 8, 2014

0.1 Twitter Social Graph

This document describes the details on data collection from Twitter used in [Volkova et al., 2014]. Our goal was to explore six types of social relationships between Twitter users and construct social graphs from them as shown using an example in Figure 1. The definition of Twitter social graph presented below is given in the paper.

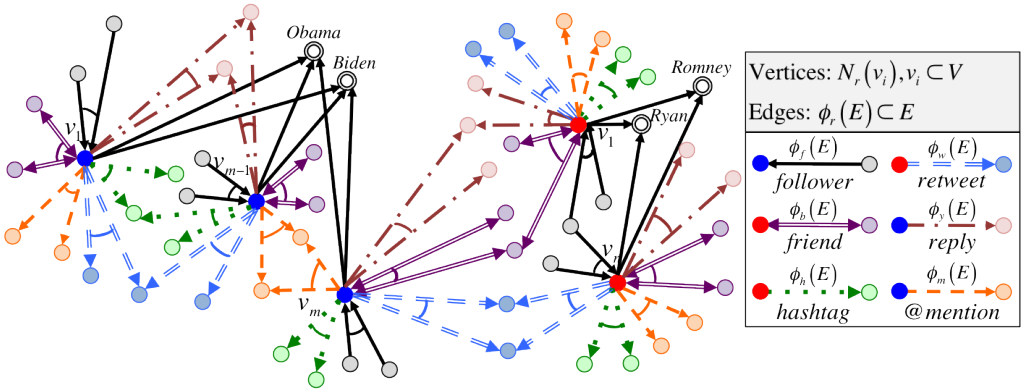


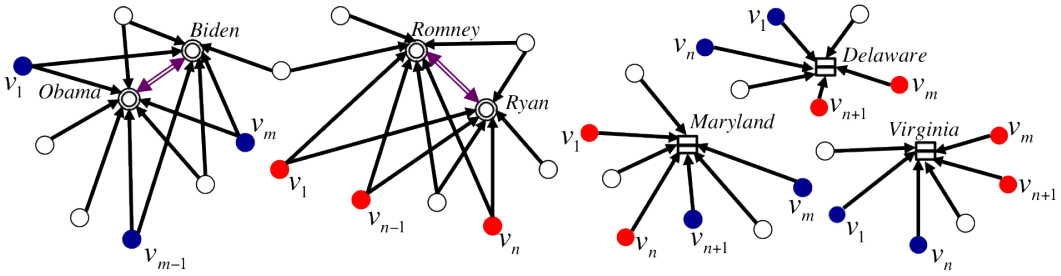
Figure 1: An example of a social graph with follower, friend, user mentions, reply, retween and hashtag social circles for each user e.g., blue: attribute a_1 and red: attribute a_2.

Political Preference Prediction: We consider several graphs including candidate-centric graph G_{cand} and geo-centric graph G_{geo} shown in Figure 2. We extend a G_{ZLR} graph crawled by [Pennacchiotti and Popescu, 2011] and further updated with friend relations by [Zamal et al., 2012]. We further extend G_{ZLR} social graph with follower, usermention and retweet relations.

	G_{cand}	G_{geo}	G_{ZLR}
D	516	135	193
R	515	135	172

Table 1: The distribution of Democratic D and Republican R users for which we construct social circles of local neighbors.

Political labels for G_{ZLR} graph were extracted from www.wefollow.com following the approach described in [Pennacchiotti and Popescu, 2011]. For



(a) Cand-centric graph G_{cand} (b) Geo-centric graph G_{geo}

Figure 2: Sampled users ●: D and R, ◎: candidates and □: geo-locations for G_{cand} and G_{geo}; ○: users not sampled; →: follower and ⇒: friend relations.

G_{cand} graph we labeled users as Democratic if they exclusively follow both Democratic candidates BarackObama and JoeBiden but do not follow both Republican candidates MittRomney and RepPaulRyan and vice versa. G_{geo} graph contains users from the Maryland, Virginia and Delaware region of the US with self-reported political preference in their biographies. The final number of labeled users per class for each graph is given in Table 1.

Gender and Age Prediction: We consider two graphs initially crawled by [Zamal et al., 2012]. From them, we recovered 383 out of 384 users labeled with gender, 381 out of 386 users labeled with age as reported in Table 2. We further extended these graphs with four type of neighbors including usermention, retweet, follower and friend. In Table 2 we report the number of users with at least one neighbor of each type (follower f, friend b, user mention m and retweet w). Moreover, in Table 3 we present mean and standard deviation for the number of neighbors per user.

Attribute	$\text{Users}_{\text{ZLR}}$	v	f	b	w	m
Gender	384	383	302	378	243	325
Age	386	381	310	359	217	324

Table 2: The number of recovered users v, and a subset of the same users with at least one neighbor – f followers, b friends, w retweets, and m user mentions in the dataset.

Gender labels were extracted using 100 most common baby boy and girl names on record with US social security department in 2011, following the

Type	Follower	Friend	Retweet	Mention
N_r^G	19 ± 3	19 ± 3	16 ± 4	8 ± 7
T_r^G	157 ± 79	193 ± 92	215 ± 72	217 ± 63
N_r^A	19 ± 7	18 ± 4	9 ± 7	20 ± 3
T_r^A	170 ± 100	174 ± 74	212 ± 66	212 ± 69

Table 3: Mean $\pm$ standard deviation for neighbors per user – N_r, and tweets per neighbor – T_r for each attribute G - gender, A - age.

technique suggested by [Mislove et al., 2011]. Age labels were extracted from user accounts with announced birthdays e.g., "Happy ##th/st/nd/rd birthday to me". Political labels were extracted from http://www.wefollow.com following the approach by [Pennacchiotti and Popescu, 2011].

0.2 Relationship Types on Twitter

In this section we describe our strategy to collect followers and friends, and extract @mentions, hashtags, replies and retweets from user communications.

- **Context-based relationships: followers and friends**

  To get *follower*[1] and *friend*[2] social circles we download lists of followers and friends for each user in Table 1. Twitter API rate limits[3] make it infeasible to download tweets for all followers and friends, so instead we randomly sample $k = 10$ followers and $k = 10$ friends per user.

- **Content-based relationships: replies, retweets, user mentions and shared hashtags**

  To get *user mention*[4] social circle we extract all @mentions from the tweets authored by users in Table 1. We eliminate 100 most frequent @mentions, treating them as stopwords (e.g., *@FoxNews* or *@youtube*) and @mentions that only occur once. For each user we randomly sample a subset $k = 10$ of their @mentions.

[1] A follower of user v_i is any v_j who subscribes to the tweets of v_i.

[2] A friend is a bidirectional following relationship – v_i follows v_j and v_j also follows v_i.

[3] Twitter API restricted queries to 3600 per day from a single IP which is now changed to 1 request per minute alloted via application only authorization as describe here: https://dev.twitter.com/docs/rate-limiting/1.1/limits.

[4] A user mention is when user v_i mentions another user v_j by screen name *e.g., @joe* in one of v_i's tweets.

Similarly, we build *retweet*[5] social circles from the retweets present in user self-authored tweets, eliminate 100 most frequent retweeted users (*e.g., @BarackObama*), and randomly sample a subset $k = 10$ usernames retweeted by each user.

To construct *reply*[6] social circle we consider tweets with filled *in reply to* field and extract reply user information. We collect all replies for users in D and R from user self-authored tweets, and randomly sample $k = 10$ reply users.

Finally, to extract *hashtag*[7] social circles, we get a sample of hashtags e.g., *#tcot, #gop or #Obama* and users sharing a specific hashtag as follows: for each user in $D \cup R$, we eliminate hashtags that occur only once, then randomly sample a set of 5 hashtags; next, we download 100 most recent tweets per hashtag, extract user information associated with each tweet, and randomly sample $k = 10$ users per hashtag.

0.3 Code, Data and Model Files

Code, raw data and pre-trained log-linear models for gender, ager and political preference prediction is available at `http://www.cs.jhu.edu/~svitlana/`

0.3.1 Code

We release Python code for querying Twitter API[8] to get (i) user/neighbor tweets e.g., using tweetIDs or userIDs/Names; (ii) lists of user immediate friends/followers; (iii) k randomly sampled user neighbors and their 200 tweets; (iv) user/neighbor timelines (up to 3200 tweets) *etc.*.

[5] A retweet is a re-posting of someone else's tweet which helps the user to share that tweet with all of his/her followers *e.g., RT @simonsam: If AC/DC is on Paul Ryan's IPod, he downloaded it illegally cause they're not on iTunes.*.

[6] A reply is a tweet sent in direct response to another tweet.

[7] A hashtag is a convention to denote that a given tweet is related to a particular topic (e.g., *#obamacare*).

[8] More details available here: `https://dev.twitter.com/docs/api/1.1`

0.3.2 Data

Data for latent user attribute prediction includes user to neighbor relations described above, and user/neighbor communications. Twitter policy restricts to sharing only tweetIDs rather than complete tweets. Therefore, we provide code for downloading actual tweets using their tweetIDs.

Twitter network data for **political preference** (Democratic, Republican) prediction:

- Candidate-centric graph: `http://www.cs.jhu.edu/~svitlana/data/graph_cand.tar.gz`
- Geo-centric graph: `http://www.cs.jhu.edu/~svitlana/data/graph_geo.tar.gz`
- ZLR graph: `http://www.cs.jhu.edu/~svitlana/data/graph_zlr.tar.gz`

Twitter network data for **gender** (Male, Female) and **age** ("18 - 23", "25 - 30" years old) prediction includes four the most predictive user-neighbor relations – friend, follower, retweet and user mention:

- Gender and Age: `http://www.cs.jhu.edu/~svitlana/data/graph_gender_age.tar.gz`

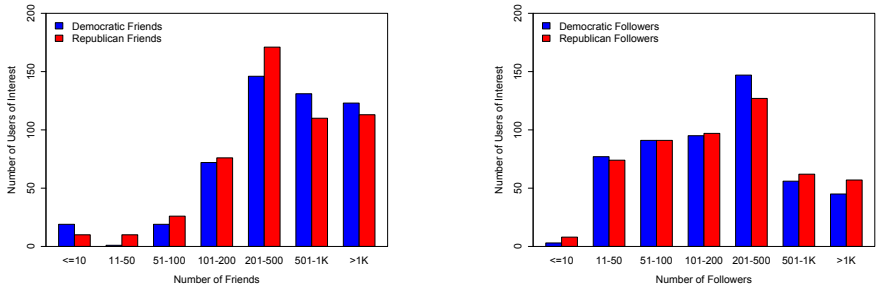
0.3.3 Models

We also release pre-trained log-linear model files that contain word unigrams and their weights for gender, age and political preference prediction. We learn these models from the data described above. We present separate models learned from user, friend, follower, retweet and usermention communications for each attribute:

- Gender: `http://www.cs.jhu.edu/~svitlana/data/models_gender.tar.gz`
- Age: `http://www.cs.jhu.edu/~svitlana/data/models_age.tar.gz`
- Political preference: `http://www.cs.jhu.edu/~svitlana/data/models_political.tar.gz`

0.4 Candidate-Centric Graph Statistics

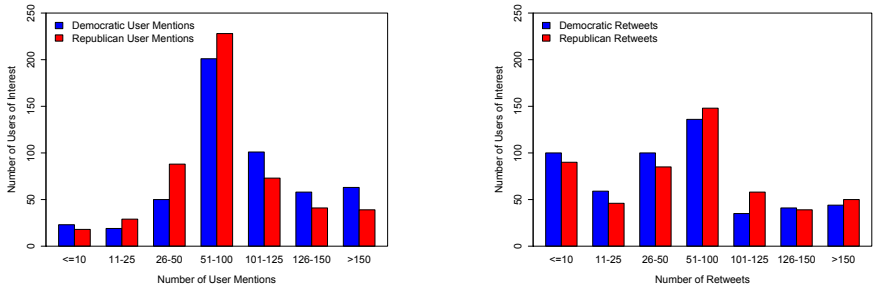
I. The distributions of follower and friend counts per user of interest are similar as shown in Figure 3 for Democratic and Republican users, and only 1-2% users have less than 10 followers or friends. Overall, both D and R users of interest have from 200 to 500 followers or friends. However, only 50% Democratic and 57% Republican users of interest have less than 500 friends compared to 80% Democratic and 77% Republican users who have less than 500 followers.



(a) Friends (b) Followers

Figure 3: Follower and friend distributions for G_{cand}.

II. The distribution of user mentions is presented in Figure 4. We find that 18% Democratic and 26% Republican users utilize less than 50 user mentions, 39% Democratic and 44% Republican users apply from 50 to 100 user mentions, 43% Democratic and 30% Republican users apply more than 100 user mentions per 200 tweets.

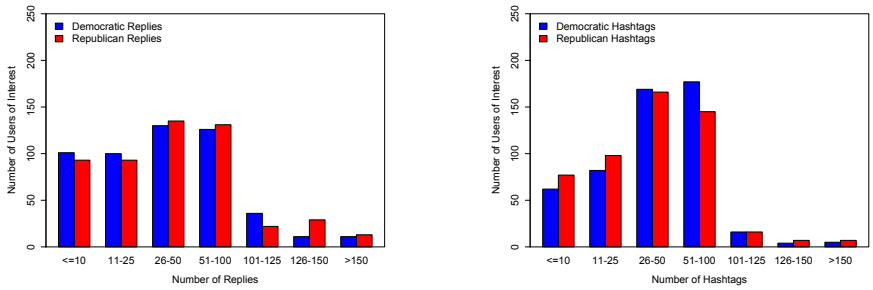


(a) User mentions (b) Retweets

Figure 4: User mention and retweet distributions for G_{cand}.

We observe that 77% Democratic and 72% Republican users of interest retweet less than 100 times per 200 tweets.

III. We observe that 39% Democratic and 36% Republican users replied less than 25 times per 200 tweets (1% - 12% reply rate), 50% Democratic and 52% Republican users replied 26 -100 times per 200 tweets (13% - 50% reply rate), and only 11% Democratic and 12% Republican users replied more than 100 times per 200 tweets (more than 50% reply rate).



(a) Replies (b) Hashtags

Figure 5: Reply and shared hashtag distributions for G_{cand}.

We find that 28% Democratic and 34% Republican users tweet less than 25 hashtags, 67% Democratic and 61% Republican users tweet 26 - 100 hashtags, and only 5% users tweet more than 100 hashtags per 200 tweets.

Bibliography

[Mislove et al., 2011] Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquis, J. N. (2011). Understanding the demographics of Twitter users. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 554–557.

[Pennacchiotti and Popescu, 2011] Pennacchiotti, M. and Popescu, A. M. (2011). A machine learning approach to Twitter user classification. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 281–288.

[Volkova et al., 2014] Volkova, S., Coppersmith, G., and Van Dume, B. (2014). Inferring user political preferences from streaming communications. In *Proceedings of the Association for Computational Linguistics (ACL)*.

[Zamal et al., 2012] Zamal, F. A., Liu, W., and Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 387–390.