**SOME NOTES ON SOCKPUPPET CREATION**

**vs@dartmouth.edu**

Sockpuppets and bots have traditionally been detected by analzying 4 broad classes of features:

1. *Meta-Data:* Meta-data about an account includes features such as age of the account, profile information (e.g. is there a photo or not), etc.
2. *Tweet Syntax:* Features in this category include factors such as
   1. Number of hashtags in a tweet
   2. Number of URLs embedded in a tweet
   3. Number of images/video/other media in a tweet
   4. Number of mentions in a tweet
   5. Number of emojis in a tweet
   6. All the above features and similar features can be associated with an author by aggregating their values across all tweets written by the author.
3. *Tweet Semantics:* Feature include an analysis of the linguistic content of a tweet such as
   1. *Sentiment* – is the tweet positive or negative, is the tweet positive or negative about a topic, what is the intensity of sentiment expressed in a tweet on a -1 to +1 scale?
   2. *Relationship between Sentiment in Tweet & Other Populations:* Such factors include how the sentiment of tweets written by an author compare with the sentiment of tweets written by authors belonging to various populations (e.g. all Twitter users; neighbors of the author in the follower-followee network).
4. *Network Analysis:* There are 3 broad types of networks that are created from a Twitter dataset These include follower-followee networks, co-mention networks, and hashtag co-occurrence networks. A number of features can be defined in each of these networks. In general, an author is sometimes considered suspicious if he belongs to a network neighborhood whose nodes have a high local clustering coefficient.
5. *Behavioral Analysis:* These include features such as the average time delay between consecutive tweets, measuring burstiness in tweets, total number of tweets posted during a given time window (e.g. a week) and various statistics linked to retweets.

**Sockpuppet Model 1 (The Unrestricted Model):** A puppetmaster operates *N* accounts. On average, he wants the accounts to have at least *K* followers but no more than *K’ < K* followers are from the *N* sockpuppets. The closer *K’* is to *K*, the more likely it is that the socks will be discovered. This means that the sockpuppet accounts must put out tweets that may not relate to their primary mission (e.g. influence opinion to be favorable to a given position) in order to garner support from non-sock accounts. In this model, we do not constrain the topics the socks discuss.

**Sockpuppet Model 2 (The Covert Model):** A puppetmaster operates *N* accounts with the goal of influencing opinion to be positive (or negative) about a specific topic *T*, e.g. positive toward cutting immigration. In this case, the puppetmaster wants **under** *K*% of his tweets to be positive toward cutting immigration – the other tweets may either be about other topics or be negative toward immigration. In fact, the puppetmaster may pretend to be against the position he really supports in order to – over time – change his position (to the true one) and tell a story for why he eventually supports his real position. The puppetmaster may use any strategy (e.g. putting out tweets on popular topics not related to *T*) in order to gain followers.

**Sockpuppet Model 3 (The Overt Model):** A puppetmaster operates *N* accounts with the goal of influencing opinion to be positive (or negative) about a specific topic *T*, e.g. positive toward cutting immigration. In this case, the puppetmaster wants **over** *K*% of his tweets to express the sentiment he wishes to push, e.g. over *K*% of his tweets are positive toward cutting immigration. In other words, he is quite overt about his desire to influence opinion about topic *T*.

There are many variants of these strategies we could consider. For example, how does the influence rate of a sock operator and/or his detection rate (ie. Probability of socks being discovered by 3rd parties) change with

1. *K*
2. *K’*
3. *N*
4. *Average number of links in socks’ tweets*
5. *Average number of links from leading news channels in socks’ tweets*
6. *Average number of media objects (images, video) in socks’ tweets*
7. *Average number of retweets by a sock*
8. *Average number of emojis/emoticons in socks’ tweets*

Note that these measurements measure performance of a SOCK OPERATOR rather than an individual sock.

Added on Sep 4 2020

**ANALYSIS OF SOCKS/SOCK OPERATORS**

The time frame of the experiment is *T=1,..,6* days (day 6 is added to capture accounts not detected after the end of day 6).

1. We define the *undetectability* of a sock account *a* to be where *T* is the total time (6 days) and is *T* if the sock is undetected or the earliest day on which *a* was reported to be a sock. Thus, an undetected sock would have a success rate of 1.
2. We define the *undetectability* of a sock creation strategy *S* to be where is the set of sock accounts created using strategy *S*.
3. The *direct influence di(a,a’,t,pol) of account a in influencing observer account a’ to have a desired polarity (pol = “+” or +-“) on topic t is:*

Where is the set of tweets posted by account *a* on topic *t* with polarity *pol* and *Followers(x)*  is the set of followers of *x* and *Observers* is the set of observers.

1. The direct influence of account *a* in spreading influence on topic *t*, polarity *pol* is: Sort of an expected value.
2. The direct influence of account *a* in spreading influence is: Sort of an expected value where *Assignments(a)* is the set of jobs *a* has, i.e. spreading influence to have polarity of *pol* on topic *t*.

The notion of “direct influence” above does not cover network effects. We could also learn a diffusion model and estimate influence *a la* Kleinberg et al. I’ll write something to this effect over the weekend.

For now, let’s write code to generate the following graphs:

**Graph 1:**

x-axis: time *t* = 1,..,6

y-axis:

*4 curves using S = the 3 strategies provided to sock operators plus ALL (all socks, regardless of strategy)*

**Graph 2:**

x-axis: time *t* = 1,..,6

y-axis:

*3 curves in all*

**Table 3:**

Undetectability score of socks generated using the 4 methods in Graph 1.

**Graph 4:**

x-axis: t=1,..,6

y-axis: *f(t)* is the average influence of sock accounts upto and including time *t*. Create a similar graph for non-sock accounts.