#bias: Measuring the Tweeting Behavior of Propagandists

Cristian Lumezanu
NEC Laboratories America, Georgia Tech
lume@nec-labs.com

Nick Feamster
Georgia Tech
feamster@cc.gatech.edu

Hans Klein
Georgia Tech
hans@gatech.edu

Abstract

Twitter is an efficient conduit of information for millions of users around the world. Its ability to quickly spread information to a large number of people makes it an efficient way to shape information and, hence, shape public opinion. We study the tweeting behavior of Twitter propagandists, users who consistently express the same opinion or ideology, focusing on two online communities: the 2010 Nevada senate race and the 2011 debt-ceiling debate. We identify several extreme tweeting patterns that could characterize users who spread propaganda: (1) sending high volumes of tweets over short periods of time, (2) retweeting while publishing little original content, (3) quickly retweeting, and (4) colluding with other, seemingly unrelated, users to send duplicate or near-duplicate messages on the same topic simultaneously. These four features appear to distinguish tweeters who spread propaganda from other more neutral users and could serve as starting point for developing behavioral-based propaganda detection techniques for Twitter.

1 Introduction

Twitter is a conduit for many different types of information, including breaking news (Kwak et al. 2010), political discourse (Conover et al. 2010), community events (Washington Post 2011a), and calls for protest (Los Angeles Times 2011). Twitter's reach and diversity of uses makes it a powerful tool for shaping public opinion: indeed Twitter is already being used to defame political candidates and discredit their views (Ratkiewicz et al. 2010; Metaxas and Mustafaraj 2010). Countries such as China are using censors to track Internet discussions and shape opinions (Directing Internet opinion 2005).

In this paper, we take a first step towards understanding how Twitter is used to spread propaganda. Propaganda is the systematic dissemination of information to support or discredit a cause, point of view, or topic (How to detect propaganda? 1937). We seek to understand how the publishing behavior of Twitter propagandists differs from that of users who express more neutral or balanced viewpoints. Behavioral differences between propagandists and balanced tweeters could be used to quickly detect extreme bias in content dissemination, without necessarily parsing the content first, and could ultimately inform future propaganda detection methods.

To study propaganda on Twitter, we must first define what constitutes propaganda. For the purposes of this initial study, we focus on a particular type of propaganda, which we call *hyperadvocacy* and define as the *consistent lack of impartiality*. While the term "propaganda" suggests the use of deception, confusion, and manipulation to change public opinion, hyperadvocacy refers to those users and content that are consistently biased towards a specific point of view, without necessarily having a malicious or subversive intent. If most tweets published by a user on a topic subscribe to a single ideology or opinion, we consider the user to be a hyperadvocate; otherwise the user is neutral. This definition serves as our "ground truth" for tweeters who we consider hyperadvocates and is appropriate for Twitter communities organized around partisan political issues, where many users consistently try to support their views and discredit opposing views. Throughout this paper we use both "hyperadvocacy" and "propaganda" to refer to the consistent lack of impartiality.

Given our definition of hyperadvocacy, our next step is to look for characteristics of tweets that are unique to hyperadvocates. One way to make hyperadvocacy, and propaganda in general, effective is to appeal to emotion by tweaking the content of tweets (*e.g.*, using words that express strong sentiment). Existing techniques analyze content to detect propaganda in traditional mass media (Herman and Chomsky 1988; mediaaccuracy.org 2008), but they are less likely to work in social media where the large number of publishers makes it difficult to establish standards for impartiality.

Another method to shape opinions on Twitter is to increase the visibility of a topic through extreme publishing behavior; we study this aspect in our paper. Consistent with Herman and Chomsky's views on spreading propaganda (Herman and Chomsky 1988), we are looking for users that are acting as amplifiers (or repeaters) of information on Twitter. We study four publishing patterns that could be associated with hyperadvocates: 1) sending high volumes of tweets over short periods of time, 2) retweeting while publishing little original content, 3) quickly retweeting others' content, and 4) colluding with other, seemingly unrelated, users to send similar content at the same time.

We analyze how these publishing patterns differ between

	NV Senate Race #nvsen	Debt ceiling debate #debtceiling
Tweets	43,032	53,282
Retweets	23,750 (55%)	3,667 (7%)
Users	6,080	26,784
- Avg tweets per user	7.1	2
Users (over 20 tweets)	326	165
- Avg tweets per user	83.5	39
- Hyperadvocates	266	59
- Neutral users	58	106

Table 1: Data sets

hyperadvocates and neutral users in two Twitter communities organized around US political issues: #nvsen, dedicated to the 2010 Nevada Senate race and #debtceiling, about the 2011 debt-ceiling debate. All of these features appear to distinguish hyperadvocates from neutral participants. Of course, the uniqueness of these features depends on the community and topic being studied; we show that volume-based features (sending high-volumes of tweets and exclusive retweeting) are more present in communities with higher fractions of retweets and higher daily volumes of tweets while time-based features (quick retweeting and collusion) appear more in groups with fewer retweets and tweets overall. That hyperadvocates and neutral Twitter users exhibit different publishing behavior patterns is extremely important for quick identification of biased content: rather than parsing the content, which may be expensive, we could monitor how content is sent.

The rest of the paper is organized as follows. In Section 2, we present background on opinion shaping, hyperadvocacy, and spreading propaganda in social media. In Section 3, we describe the data sets we use in our analysis and how we quantify opinion bias in them. Section 4 describes our attempts to understand and evaluate the four behavioral patterns that could differentiate between hyperadvocates and neutral users. We conclude in Section 5.

2 Background

Propaganda has existed in traditional mass media for many years (psywar.org 2011; Herman and Chomsky 1988), and it is now slowly permeating social media. For example, the United States government is funding projects to both build online persona management software which would help disseminate propaganda (The Guardian 2011) and to detect and track popular ideas in social media (Washington Post 2011b). China has long employed teams of censors to track Internet discussions and quickly shape opinions (Directing Internet opinion 2005).

Spreading misinformation or lies could alter perceived public opinion on a topic and have real-world repercussions. Previous studies have shown that the results of elections are positively correlated with opinion about candidates expressed on Twitter (OConnor et al. 2010; Tumasjan et al. 2010). Other reports mention the significant role of Twitter and Facebook in helping organize and coordinate protests in Tunisia, Libya, or Egypt (Los Angeles Times 2011).

Quickly and accurately identifying users that exhibit bias towards or against certain topics is important for preserving the fairness and openness of social media. Some efforts have focused on analyzing the content of news and identifying publishers with strong sentiment or unusual choices of words in traditional news sources, such as newspapers (mediaaccuracy.org 2008). Several tools attempt to analyze the content of tweets by detecting positive or negative opinions (Twitter sentiment 2011). However, with thousands of publishers, it is difficult to establish a standard for the impartial way of presenting a piece of information on Twitter. Additionally, the lack of accountability allows propagandists to easily evade content-based filters by sending information from multiple identities. In contrast, detection techniques based on watching tweeting behavior could avoid these shortcomings because they focus on *how* to send content effectively rather than on *what* the content is (Ramachandran and Feamster 2006; Ramachandran, Feamster, and Vempala 2007).

The Truthy system detects suspicious memes of Twitter by studying the diffusion network of a topic (*i.e.*, who tweets and retweets about the topic) (Ratkiewicz et al. 2010). Truthy can detect suspicious topics but cannot always distinguish hyperadvocates from more neutral users, since legitimate users may unwittingly participate in a discussion concerning a suspicious topic. Our approach of studying the behavior of users is complementary to Truthy and is inspired by Herman and Chomsky's seminal work on propaganda (Herman and Chomsky 1988), which identifies repeating content as an effective way of spreading propaganda.

3 Collecting and Labeling Data

In this section, we describe the data sets used in our analysis and show how to identify hyperadvocates.

3.1 Data sets

We collect tweets from two online discussion groups: the 2010 Nevada Senate race (identified by the hashtag #nvsen) and the 2011 debt ceiling debate (identified by the hashtag #debtceiling) using the API offered by Topsy (Topsy 2011). The collection process is rate-limited by Topsy and may not retrieve all tweets in a community. Our estimates based on the number of results in direct searches for the corresponding hashtags, however, indicate that we processed more than 80% of the tweets in each community.

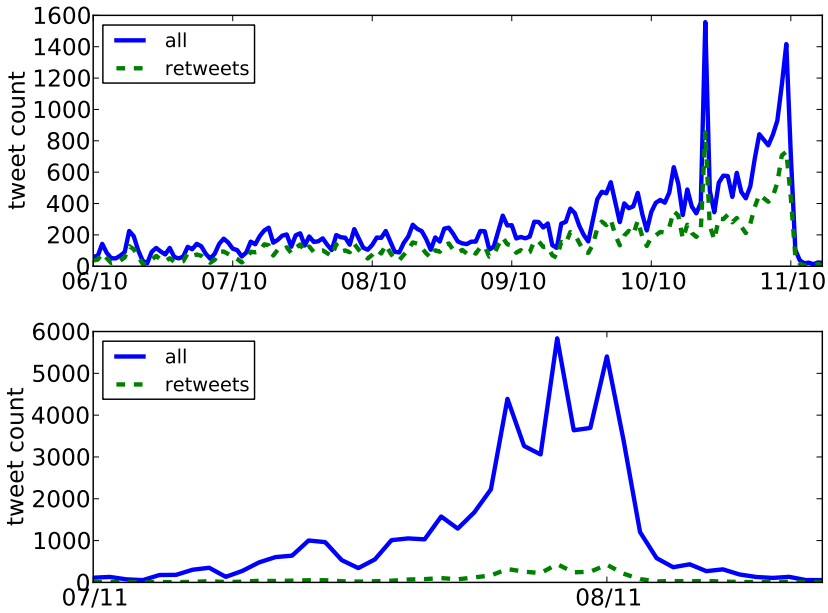

Figure 1: Number of tweets and retweets published each day in the #nvsen (top) and #debtceiling (bottom) communities.

Table 1 presents statistics about the data sets. Figure 1 shows the number of tweets and retweets per day for the two communities. We make two observations. First, in #nvsen, a much higher fraction of the tweets are retweets, meaning that #nvsen has fewer "original" publishers. Second, the number of tweets in each community increases as we approach the deadline around which the community is organized (November 2, 2010 for #nvsen and August 2, 2011 for #debtceiling); after the deadline has passed there are very few tweets with the corresponding hashtag. In addition, in #nvsen there is a high volume of tweets on October 15, 2010, the day of the candidates debate. Figure 2 shows the CDF of the number of tweets sent by each user; while, in both cases, the majority of users sent fewer than 10 tweets, there are a small number of high-volume tweeters in each case.

3.2 Labeling data

Because there is no precise quantitative definition of hyperadvocacy, obtaining ground truth for what messages constitute hyperadvocacy is challenging. Although some users are clearly hyperadvocates (*e.g.*, *dumpreid* in #nvsen, whose tweets continually defame democratic candidate Harry Reid), for most users such a clear-cut classification is more difficult. As previously mentioned, we define hyperadvocacy as lack of balance or impartiality: users that consistently express the same opinion about a topic are hyperadvocates, while those that express mixed opinions or views are neutral.

To classify a user as biased or neutral, we apply the following finding of Conover *et al.* (Conover et al. 2010), used to find political polarization on Twitter: *users with similar ideologies tend to retweet exclusively each other's messages*. To evaluate their method, Connover *et al.* used qualitative content analysis (Kolbe and Burnett 1991) on the text

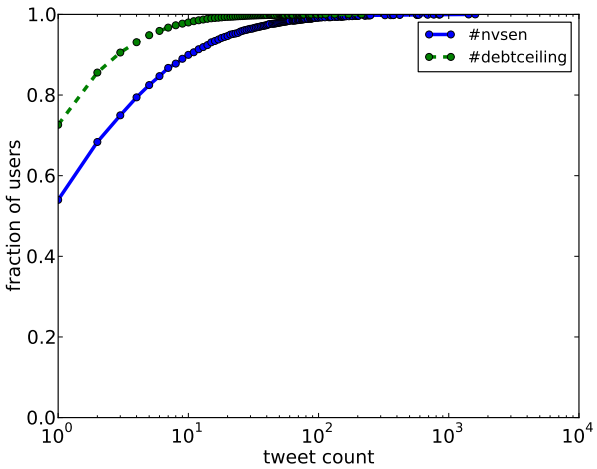

Figure 2: Cumulative distribution of the number of tweets sent by each user in the #nvsen and #debtceiling communities.

of tweets to annotate users with a specific political identity. Using 252,300 politically relevant tweets, they showed that the retweet communities identified by their algorithm match groups of users with similar political alignment.

Our method of labeling users has two steps. The first step is identical to the approach of Conover *et al.*: we begin by randomly assigning each user to one of two clusters[1], cor-

[1]We choose to start with two clusters because this fits well with the bipartisan US political landscape. The method works with any number of clusters.

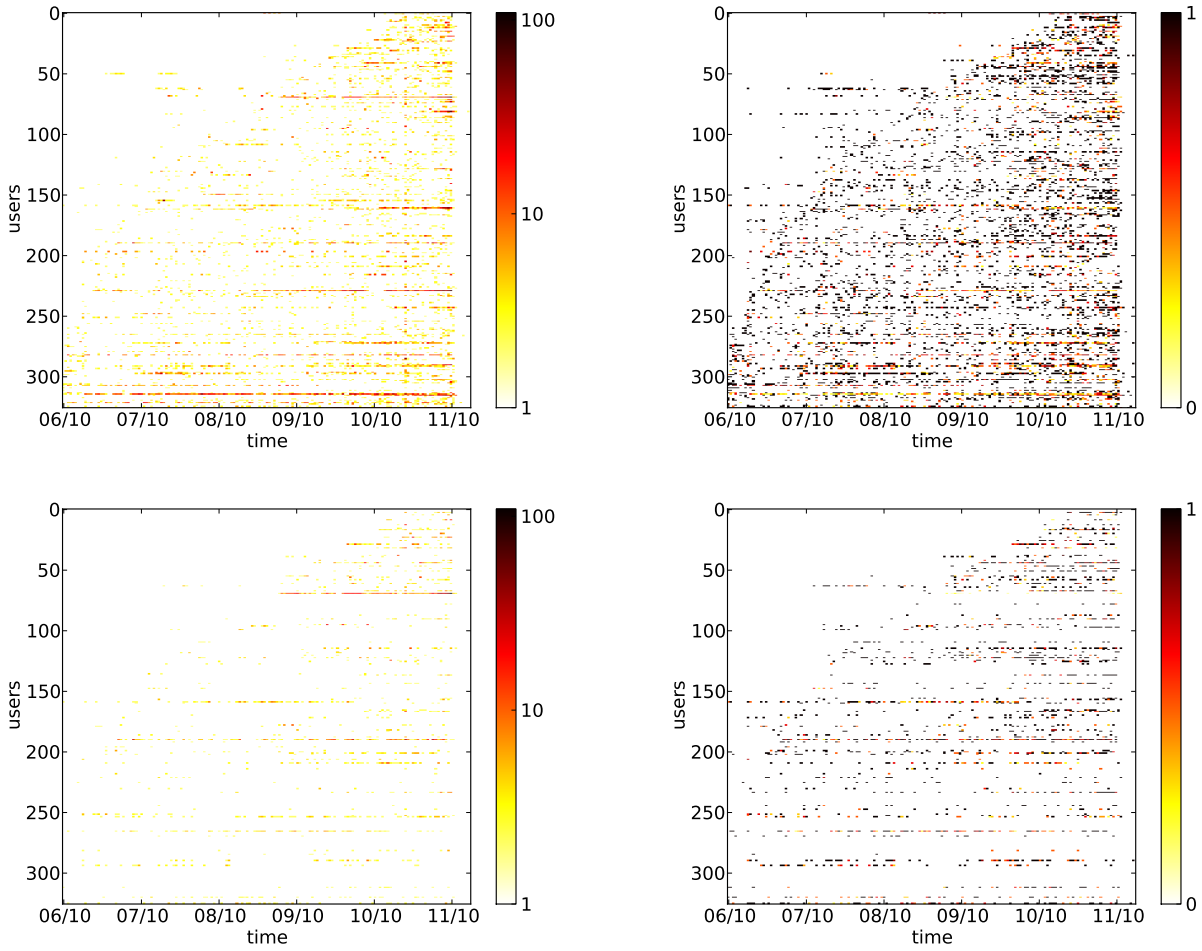

Figure 3: Publishing patterns for hyperadvocates (**top**) and neutral users (**bottom**) in the #nvsen community. Each point is associated with a user and a day. Its color intensity reflects the number of all tweets **(left)** or fraction of retweets **(right)** published by the user during the day. The color bar has logarithmic scale for the left-hand plots.

responding to one set of related opinions (*e.g.*, liberal and conservative). We seed the algorithm by assigning the users whose political views we know (such as the accounts of the two candidates, @harryreid and @sharronangle, in #nvsen) to the cluster corresponding to their viewpoint. We consider two users to be associated with one another if one of them retweets a message originally published by the other; then, we iterate through the users, reassigning every user to the cluster for which that user has the most associated users. We stop after 1,000 iterations.

In the second step, we inspect every user: if at least a fraction f of the connections are to users in the same cluster then the user is a hyperadvocate; otherwise, the user is neutral. The choice of f defines the sensitivity of the algorithm in labeling hyperadvocates. Lower values for f yield more hyperadvocates but the risk of false positives is higher. High values of f implicitly establish a more strict labeling. We experimented with values of f from 0.8 to 1. Table 1 presents statistics about the number of propagandists and neutral users for $f = 0.8$. Because it is based on retweeting behavior, this approach may lead to mislabeling some of the hyperadvocates in communities with few retweets (such as #debtceiling). We return to this limitation in Section 4.

4 Understanding Tweeting Behavior

In this section, we study how the tweeting behavior of hyperadvocates differs from that of neutral users. We start by making several observations about tweeting patterns in the the #nvsen and #debtceiling communities; we then use the labeled sets of users from Section 3 to study specific behaviors in more detail.

4.1 Observations

Volume. Figure 2 presents the cumulative distribution of the volume of tweets sent by each user. Most accounts send very few tweets. Only 5% of users in #nvsen and less than 1% of users in #debtceiling send more than 20 tweets over the measurement period. Because it is unlikely that low-volume

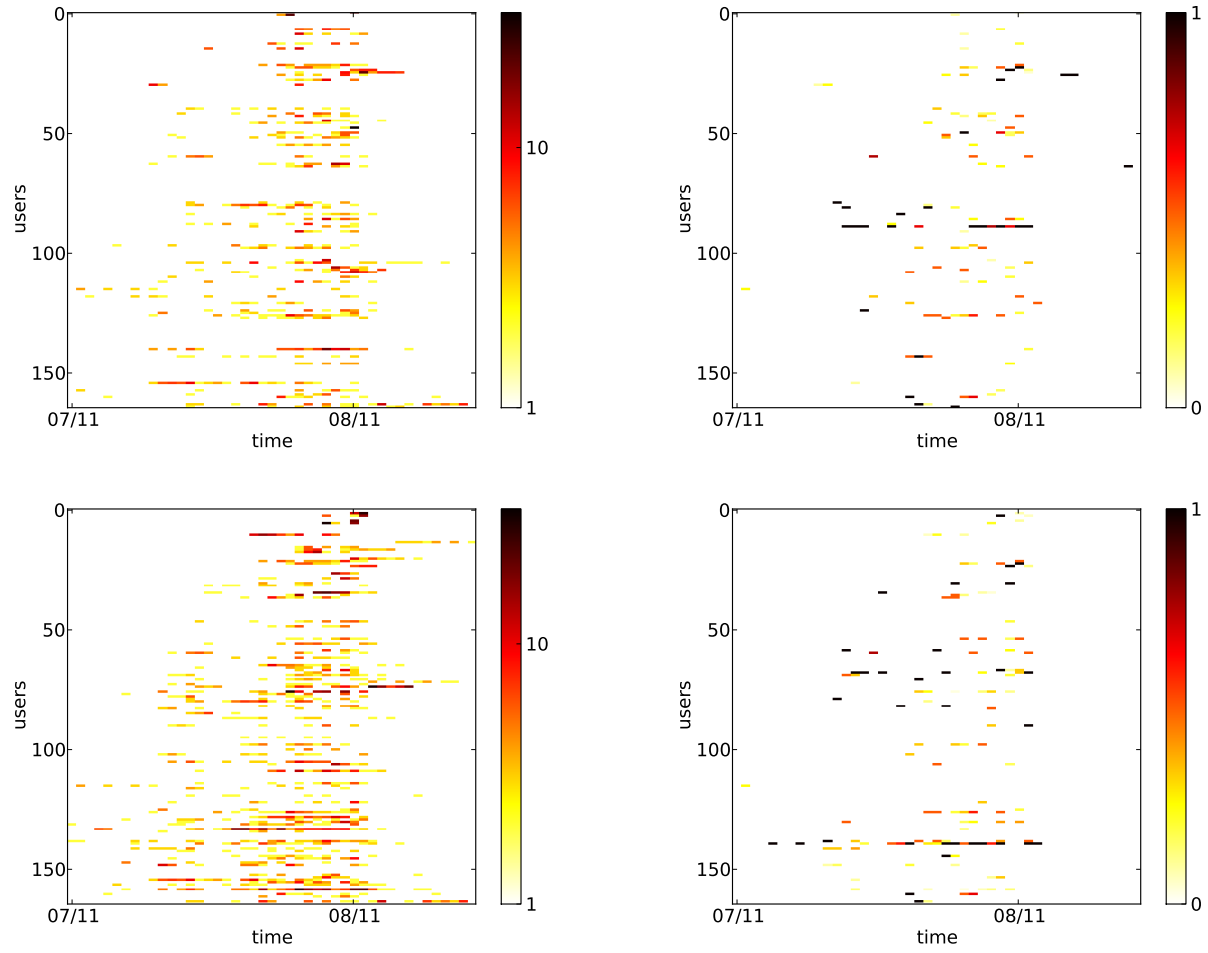

Figure 4: Publishing patterns for hyperadvocates (**top**) and neutral users (**bottom**) in the #debtceiling community. Each point is associated with a user and a day. Its color intensity reflects the number of all tweets **(left)** or fraction of retweets **(right)** published by the user during the day. The color bar has logarithmic scale for the left-hand plots.

accounts can be effective hyperadvocates by themselves, we focus on those users that send more than 20 tweets in either community. Of course, low-volume users might still collude to gain better exposure and higher audiences for their messages. We examine such scenarios in Section 4.5.

Tweeting and retweeting behavior. Figure 3 shows when and how hyperadvocates and neutral users publish their messages (left) and retweets (right) in #nvsen. Each point is associated with a user and a day, and its color intensity is proportional to the number of tweets or fraction of retweets sent by the user during that day. Users are sorted according to their lifetime (*i.e.*, difference between the timestamps of the last and first tweets published in the community), with the long-lived users towards the bottom of the plots; within each community, the ordering of users is consistent across plots.

We make several observations about tweeting behavior in #nvsen:

- hyperadvocates send **higher daily volumes** of tweets (predominantly dark lines in left-hand plots);
- similarly, hyperadvocates consistently send a **higher daily fraction of retweets** (lines with many black regions in right-hand plots).

We also studied the tweeting and retweeting behavior in #debtceiling, depicted in Figure 4, but did not observe the same behaviors as in #nvsen. We return to the reasons of why the observations are not consistent in Section 4.5.

These results suggests that two tweeting patterns may be specific to some hyperadvocates: (1) they send more messages over short periods and (2) they send more retweets than original content. In addition, inspired by Herman and Chomsky's work on propaganda models (Herman and Chomsky 1988), we propose two more patterns that could help amplify the effect opinion shaping: (3) they retweet quickly, and (4) they collude to send similar messages simultaneously, to offer the illusion of volume and coverage for a specific topic. The last two tweeting patterns carry

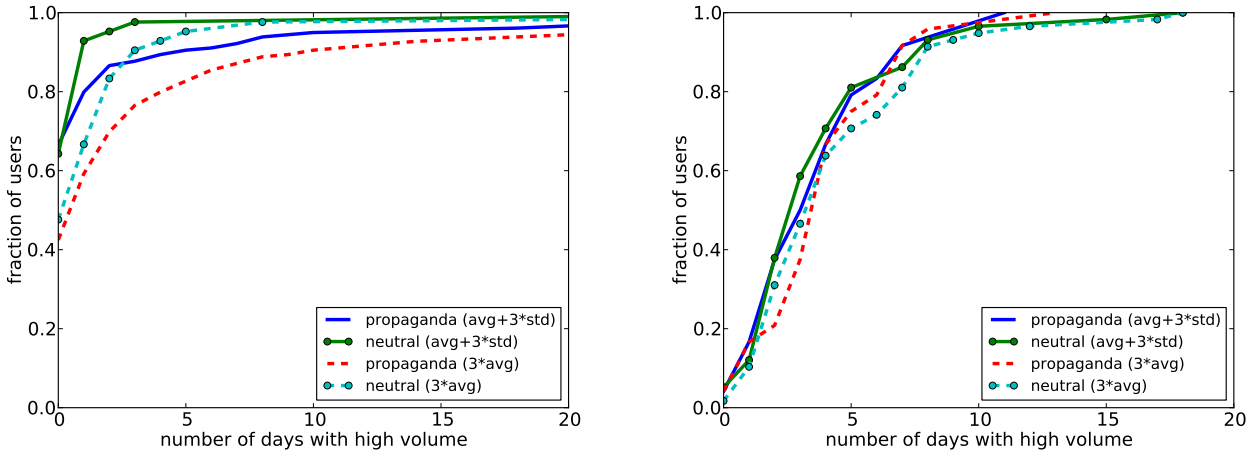

Figure 5: Cumulative distributions of the number of high-volume days for **(left)** #nvsen and **(right)** #debtceiling. Propagandists have more high-volume days than neutral users in #nvsen.

an implicit malicious intent and could help us identify who among hyperadvocates is truly propagandist. We evaluate these four features next.

4.2 Bursty volumes

We consider a user to have a high-volume day if it publishes more tweets than a predefined threshold θ. Because what constitutes normal publishing behavior varies from day to day (*e.g.*, users are expected to publish more tweets right before the election than six months before the election), the value of θ varies daily. We experiment with two formulations for the threshold θ. First, we consider the θ as three standard deviations over the daily average number of tweets ($\theta_1 = avg + 3 * std$), which, in a normal distribution would eliminate about 99% of the users. Second, we set θ to three times the daily average number of tweets ($\theta_2 = 3 * avg$), based on observations we made on the two data sets that this value filters out about 90% of all users.

Figure 5 shows the distribution of the number of high-volume days for both types of users. Hyperadvocates have more high-volume days than neutral users in #nvsen. Only one neutral user has more than three high-volume days when $\theta = avg + 3 * std$, while some hyperadvocates have over twenty such days. Because each user in #debtceiling sends fewer tweets, even hyperadvocates do not exhibit high-volume publishing patterns. As we show later in the section, publishing patterns that account for time between tweets are better suited for #debtceiling.

4.3 Exclusive retweeting

As Conover *et al.* discovered, retweets are a popular method for spreading messages in Twitter political communities (Conover et al. 2010). Retweeting amplifies the visibility of a topic by exposing it to a different audience or by increasing the volume of tweets that mention it. Next, we study whether the fraction of retweets that a user sends can help determine whether the user is a hyperadvocate.

We define a user's *repeater score* as the ratio of the number of retweets to the total number of tweets. Figure 6(left) shows the distribution of the repeater score for users in #nvsen and #debtceiling. Few users in #debtceiling are repeaters; of those, hyperadvocates send slightly more retweets than neutral users. The gap is wider for #nvsen: 75% of hyperadvocates and only 35% of neutral users have a repeater score of at least 0.5 (*i.e.*, more than half of their messages are retweets).

4.4 Quick retweeting

To increase the effectiveness of a message, hyperadvocates could use automated accounts to quickly retweet content shared by users they support. We define the *reaction time* for a retweeted message as the difference between the time of the retweet and the time of the original message. We consider all retweets in each community and, for every user that has retweeted at least five times in the community, compute the average reaction time across all her retweets. We do not consider the retweets where we cannot identify the time of the original message (*e.g.*, because the original appeared before we started data collection or is not part of the community). For messages retweeted multiple times we compute the reaction time based on the most recent retweet.

In Figure 6(right), we show the cumulative distribution for the average retweet time for all users with more than five retweets and 20 tweets overall. The results for #nvsen match our intuition only partially. On one hand, no neutral user retweets less then 15 minutes after the original message, while a few hyperadvocates have an average retweet time of around one minute. On the other hand, however, hyperadvocates tend to have a higher retweet reaction time than neutral users. For #debtceiling, hyperadvocates are slightly quicker in retweeting although no one retweets less than one hour after the original message.

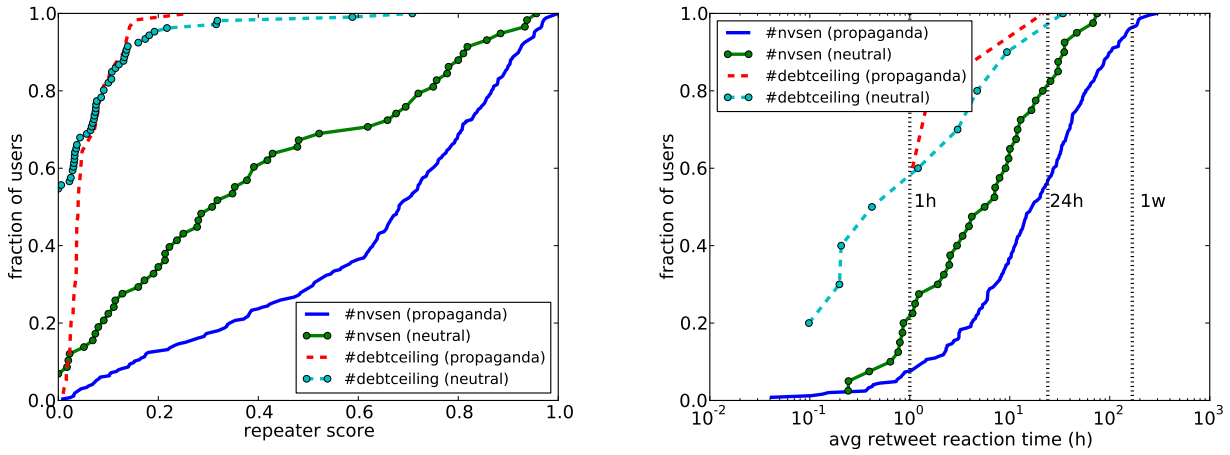

Figure 6: CDFs for **(left)** repeater score and **(right)** average retweet reaction time for active users in #nvsen and #debtceiling. Hyperadvocates send more retweets and less original content than neutral users but do not necessarily retweet faster.

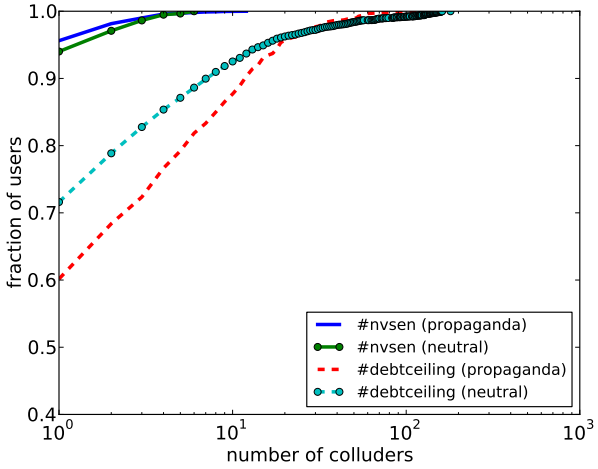

Figure 7: Cumulative distribution of the number of users that send similar messages less than three minutes apart, for every user in #nvsen and #debtceiling.

4.5 Collusion

To appear neutral while still increasing the visibility of their message, hyperadvocates might collude to send similar messages simultaneously. We study how often and how far apart in time pairs of users send duplicate and near-duplicate messages. Two messages are near-duplicates if they have more than 80% of the words in common. We do not consider retweets because we want to focus only on similar messages where the connection between senders is not explicit.

For each user in #nvsen and #debtceiling, we compute the number of *colluders*, the users in the same community that send duplicate or near-duplicate messages very close in time. Figure 7 presents the cumulative distribution of the number of colluders for messages sent less than three minutes apart. (Results for one, five, and ten minutes apart were similar.) 16% of all users (including those with less than twenty tweets) in #nvsen and 46% of those in #debtceiling have at least one colluder. Hyperadvocates and neutral users in #nvsen exhibit similar behavior. On the other hand, hyperadvocates collude more often than neutral users in #debtceiling: 40% of propagandists and fewer than 30% of neutral users have more than one colluder. These results offer a possible explanation for why the volume-based behavioral patterns appear more in #nvsen than in #debtceiling: #debtceiling has more users and fewer messages per user where propaganda is spread through collusion of low-volume users rather than retweeting or high-volume tweeting over short intervals.

4.6 Summary

We studied four tweeting patterns that could help amplify the effect on hyperadvocacy on Twitter and showed that their presence or absence depend on the properties of the community being analyzed. Volume-based patterns, such as sending many tweets over short periods or retweeting without publishing much original content are better suited for communities with higher average number of tweets and fraction of retweets. On the other hand, time-based patterns, such as quickly retweeting or sending similar messages close in time appear in communities with fewer retweets and smaller volume of tweets.

5 Conclusion

This paper presented a first step towards measuring how Twitter is used to disseminate propaganda. We focused on a particular type of propaganda: hyperadvocacy, or the consistent dissemination of content that subscribes to a single ideology or opinion. Using observations of tweeting behavior in two Twitter communities, as well as intuition from Herman and Chomsky's seminal work on propaganda models (Herman and Chomsky 1988), we described four

tweeting patterns that could amplify the visibility of propaganda on Twitter. We evaluated these patterns using tweets about the 2010 Nevada senate race and the 2011 debt-ceiling debate, and showed that their presence depends on the properties of the Twitter community being analyzed, but that, ultimately, they could provide a starting point towards behavioral-based detection of Twitter propaganda.

References

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonalves, B.; Flammini, A.; and Menczer, F. 2010. Political polarization on Twitter. In *AAAI ICWSM*.

2005. The Practical Aspects of Directing Internet Opinion. http://goo.gl/wzlfr.

Herman, E. S., and Chomsky, N. 1988. *Manufacturing Consent: The Political Economy of the Mass Media*. Pantheon Books.

1937. How to detect propaganda? http://goo.gl/Wg4O9.

Kolbe, R. H., and Burnett, M. S. 1991. Content analysis research: An examination of applications with directives for improving research reliability and objectivity. *The Journal of Consumer Research* 18(2):243–250.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *WWW*.

2011. Tunisia protesters use Facebook, Twitter and YouTube to help organize and report. http://goo.gl/PfP3C.

2008. Colombia and Venezuela: Testing the Propaganda Model. http://goo.gl/ybHsy.

Metaxas, P. T., and Mustafaraj, E. 2010. From obscurity to prominence in minutes: Political speech and real-time search. In *WebSci*.

OConnor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *ICWSM*.

2011. Psywar. http://goo.gl/SVvWL.

Ramachandran, A., and Feamster, N. 2006. Understanding the network-level behavior of spammers. In *ACM Sigcomm*.

Ramachandran, A.; Feamster, N.; and Vempala, S. 2007. Filtering spam with behavioral blacklisting. In *ACM CCS*.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A.; and Menczer, F. 2010. Detecting and tracking the spread of astroturf memes in microblog streams. In *AAAI ICWSM*.

2011. US spy operation. http://goo.gl/cJVOM.

2011. Topsy. http://www.topsy.com.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *AAAI ICWSM*.

2011. Twitter Sentiment. http://goo.gl/TocPH.

2011a. Obama holds first White House Twitter Town Hall. http://goo.gl/t55GQ.

2011b. Pentagon puts out a call for the socially savvy. http://goo.gl/Ce7aH.