Vocal Minority versus Silent Majority: Discovering the Opinions of the Long Tail

Eni Mustafaraj, Samantha Finn, Carolyn Whitlock, and Panagiotis T. Metaxas
Department of Computer Science
Wellesley College
Email: (emustafa, sfinn, cwhitloc, pmetaxas)@wellesley.edu

***Abstract*—Social networks such as Facebook and Twitter have become the favorite places on the Web where people discuss real-time events. In fact, search engines such as Google and Bing have special agreements, which allow them to include into their search results public conversations happening in real-time in these social networks. However, for anyone who only reads these conversations occasionally, it is difficult to evaluate the (often) complex context in which these conversation bits are embedded. Who are the people carrying on the conversation? Are they random participants or people with a specific agenda? Making sense of real-time social streams often requires much more information than what is visible in the messages themselves. In this paper, we study this phenomenon in the context of one political event: a special election for the US Senate which took place in Massachusetts in January 2010, as observed in conversations on Twitter. We present results of data analysis that compares two groups of different users: the *vocal minority* (users who tweet very often) and the *silent majority* (users who tweeted only once). We discover that the content generated by these two groups is significantly different, therefore, researchers should take care in separating them when trying to create predictive models based on aggregated data.**

I. Introduction

Growing participation in social media and networks has resulted in an explosion of so-called user-generated content. Researchers are using these data to answer interesting questions which were impossible to tackle before: how does an idea become viral, who are the most influential users in a social network, or how to attract support for a good cause? In addition, there has been a lot of research in using these data for predictions of different kinds, for everything from predicting movie box office results based on Twitter chatter [1], to predicting the stock market based on Twitter users' mood [2], or predicting the outbreak of diseases [3]. One study has established a correlation between the sentiment of tweets and public opinion polls for job approval ratings [5], while yet another used chatter volume on Twitter to accurately predict the results of the parliamentary German elections in 2009 [6].

While researchers are aware of the power-law nature [7] of many of the datasets collected from the Web, they do not use sampling in their analysis, relying on the very large size of the datasets to smooth out differences in the data. In this paper, we argue that this technique might be flawed, especially if the data is used for predictions. Our analysis suggests that in particular occasions (such as a toss-up political election), where stakes are high and public opinion can shift in the space of hours, the largest amount of user-generated data is authored by a group of dedicated users, the "vocal minority", who go at great lengths to create the impression that they and their opinions are the majority. While this happens, the real majority remains silent and contributes to the conversation sporadically, mostly after an important event has concluded (for example, the results of the election are announced).

Our contribution in this paper is to support the hypothesis that content generated by users of the vocal minority and silent majority groups is different, and that it should not be aggregated together to try to build predictive models or infer other conclusions. Concretely, the users of the vocal minority compose their tweets by using more hashtags, links, and mentions, and retweet at double the rate of the silent majority. Furthermore, they tweet the same pre-fabricated content in a crowdsourced manner, resulting in so-called Twitter protests aimed at traditional media outlets to influence their reporting towards topics that the vocal minority believes are of most-important interest.

II. Elections and Social Media Buzz

When President Barack Obama was elected in 2008, much was written about his extensive use of social media and networks to mobilize and energize voters, e.g. [8]. Metrics such as his number of fans on Facebook, his number of followers on Twitter, the number of videos on his YouTube channel and the number of views for these videos, as well as the large community of bloggers under the umbrella of his own social network mybarackobama.com, all were interpreted as numbers which indicated the high level of interaction from volunteers and voters with his political message, which ultimately contributed to his election. Even the larger number of searches for his name on Google (as compared to that of his opponent, the Republican John McCain), was seen as a good predictor of his win, because it expressed the "wisdom of the crowd".

Traditional and new web-based media put a lot of emphasis in the interpretation of the correlation between these numbers and the election result, so much so, that political campaigns around the country and the globe took note. The take-away lesson was that to increase your chances of winning, you need to be prominently active in social media and networks. Not only this, but the metrics that measure this activity need to be

in your favor, so that one can point to them as indication of broader support from the voters.

The first election campaign after President Obama's win was the 2010 United States Senate special election in Massachusetts, held on January 19, 2010, to assign the seat held by the late Senator Ted Kennedy. For weeks and months before the voting day, the Democratic candidate, Martha Coakley, comfortably led her opponent, the Republican Scott Brown, in all polls. In fact, two weeks before the election, a Boston Globe poll predicted Coakley 50% - Brown 35%. However, a Rasmussen Reports poll only one week before the election, saw the result in a statistical dead heat: Coakley 49% - Brown 47% [9]. At that moment, the battle in the social media and elsewhere became even more intense. A major component of it was to create the impression that one of the candidates was genuinely leading the popularity contest, as measured by the usual metrics: number of fans, followers, volume of Google searches, etc. However, at this point it was difficult to verify the accuracy of such metrics, because supporters turned these numbers into a game. For example, in our Twitter corpus, a single user sent 115 tweets to urge users to join Scott Brown's page on Facebook (see Table I); and 54 tweets to urge users to follow Brown on Twitter (see Table II).

While this user and others were using Twitter to increase the number of fans and followers, a social media analyst used these metrics to predict the election, wrote a blog post titled: "New social media polling data suggests Republican Scott Brown will trounce Democrat Martha Coakley in US Senate Race", and then tweeted 11 times to promote his own poll. Other users and websites caught up with this blog post, and started repeating and retweeting these results. In total, we found 765 tweets mentioning the "polls" based on social media metrics. Thus, certain Twitter users worked non-stop to increase the number of fans and followers on Facebook and Twitter, and then used these numbers as an evidence that Scott Brown was the most-supported candidate, even though the numbers represented individuals from all the United States, while the election was taking place in Massachusetts.

A. Facebook versus Twitter

At the end of the electoral race, Scott Brown had won the election and also the battle of social media in terms of fans and followers. As the text of the tweets in Table I and Table II show, in one week he doubled the number of his Twitter followers and quadrupled the number of Facebook fans. According to the tweets we have collected, the political engagement on Facebook was very heated, as some of the messages in Table III show.

Since many of these conversations take place on private Facebook pages and not in public forums, they cannot be collected by researchers not affiliated with Facebook. In this respect, the nature of Twitter as a public micro-blogging service with mostly public status updates, and an API that allows their collection, makes it easy for all researchers to tap into this data and study different problems. This is why all the studies we cited in the introduction are based on Twitter data.

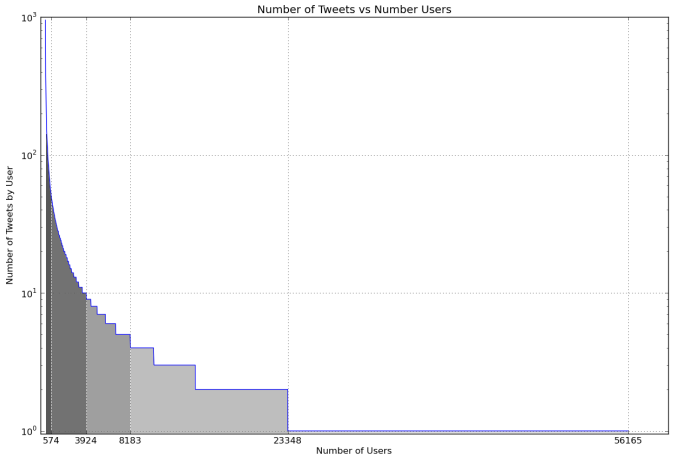

Fig. 1. Distribution of number of tweets for every user, ranked based on the most prolific user. Notice the long tail of users who tweeted only once. The five different shades of gray indicate the five categories of groups as established in Table IV.

B. Twitter Dataset for the Election

We became aware of the intensity of the social media battle around this election on January 12, when the names of Martha Coakley and Scott Brown became Twitter trending topics, as a result of the release of the Rasmussen Reports poll that saw the race as impossible to predict. Using the Twitter Streaming API, we collected all[1] tweets that contained the words "coakley" and "scott brown" starting on Jan 13 until Jan 20. The collection comprises 234,697 tweets contributed by 56,165 different Twitter accounts. We performed a series of analyses on this data to discover who is generating content on Twitter about the Senate election and whether all this content can be treated in the same way.

III. Is all user-generated content equal?

The answer to this question could be "yes" if all users on social media and networks would behave in more or less the same way, in terms of the nature and amount of generated content. But is that the case? Figure 1 shows how often users tweeted about the Massachusetts Senate election.

Immediately striking in this plot is its very long tail, which comprises 58.4% of all users in our dataset, who contributed only one tweet each. Our hypothesis is that the tweets generated by the users of the long tail are different from the tweets generated by users in other parts of the distribution, especially the so-called power users (the ones who tweeted a large number of times). In order to be able to talk about this difference, we divide users in different categories and then proceed to support our hypothesis by comparing the content of tweets generated from users in different categories.

[1]The Twitter Streaming API allows the downloading of all tweets containing a particular word. The particular keywords we chose were selected so as to maximize the number of tweets that were relevant to the candidates of the MA special elections while ignoring those that they may accidentally share a keyword (e.g., just "scott' or just "brown").

TABLE I
TWEETS URGING USERS TO BECOME FANS OF SCOTT BROWN ON FACEBOOK

Timestamp	Tweet Text
01-13-10 23:49:06	Jan13 @ScottBrownMA Facebook Fans: 1:09AM = 29,713; 11:49PM EST = 39,717! + 10K in under 23 hrs! SCOTT IS SURGING!
01-16-10 14:32:03	Scott Brown has over 60,000 #Facebook fans. Support Scott 4 #ThePeoplesSeat: http://tcot.me/o7k4 #MASen #tcot #sgp
01-20-10 01:26:44	WOW! Scott Brown has over 120,000 #Facebook fans! JOIN! http://tcot.me/o7k4 @ScottBrownMA #MASen (my final update)

TABLE II
TWEETS URGING USERS TO FOLLOW SCOTT BROWN ON TWITTER

Timestamp	Tweet Text
01-14-10 01:04:13	YES! @ScottBrownMA has 7,000 Twitter followers. Over 7,500 on 1/14? Follow, RT & List Scott #MASen #tcot #sgp #41stVote
01-18-10 12:38:35	Scott Brown has over 10,500 Twitter followers! Follow, RT & List @ScottBrownMA #MASen #41stVote
01-20-10 01:11:24	WOW! Scott Brown has over 15,500 #Twitter followers! Follow, RT & List @ScottBrownMA #MASen (my final update)

TABLE III
TWEETS ABOUT USER EXPERIENCES ON FACEBOOK

Tweet Text
Martha Coakley's facebook fan page comments are highly entertaining. She is not well-liked online...
#tcot Scott Brown Facebook Challenge!: Please copy my profile picture and use it as your own until election day! -... http://bit.ly/6zBArU
Noticing lots of my friends on Facebook becoming fans of Scott Brown :) #ScottBrown
latest FaceBook poll: Brown 75%, Croakley 22% and Kennedy 3%: http://bit.ly/5LpJtE #masen #scottbrownma #tcot
So sad seeing Scott Brown supporters amongst my Mass friends on facebook. Ugh he opposes gay marriage & wants to destroy Teddy's legacy.
In wild MA Sen race, law enforcement looks into Facebook talk of violence against Dems Martha Coakley, state AG.
Scott Brown Supporters On Facebook Wish For Martha Coakley To Be Killed And Raped - http://j.mp/6J2ZYf
I read some of the defenses for Scott Brown on facebook and I'm left wondering how these people are even allowed to vote.
Cannot believe the number of my facebook "friends" are militant /vocal Scott Brown fans. I'm rethinking my "friendship" Vote Martha! #masen
"I wasn't even paying attention to this election until about 2 days ago... apparently Coakley wasn't either." - random facebook quote
I love my Republican friends, but I can't even log on to Facebook because everyone's uncorking goddamn champagne over Scott Brown's win.

TABLE IV
CATEGORIZATION OF TWITTER USERS WHO WROTE ABOUT THE MA SENATE RACE.

Group Name	Tweets per User	Users	Tweets per Category	Tweet Volume (%)
Silent	1	32817	32817	14.0%
Attentive	2-4	15165	38844	16.6%
Interested	5-9	4259	27505	11.7%
Engaged	10-49	3350	66252	28.2%
Vocal	50+	574	69279	29.5%
		Total: 56165	Total: 234697	Total: 100%

A. Categorizing Users

To divide users in different categories based on their tweeting behavior, we decided to make use of a five-point scale, which is usual when studying a continuous spectrum. The first group consisting of users with only one tweet was a simple choice for a threshold, while all the other threshold values: 2–4 tweets, 5–9, 10–49, and more than 50 tweets, are chosen subjectively, by roughly doubling the amount of tweets from group to group, whenever possible[2]. The stark contrast between the number of users in the one-tweet group (32,817) and that of the users in the 50+ tweets (574) gave us the idea of calling the two groups: silent majority and vocal minority. The other group names: "attentive", "interested", and "engaged" are chosen to describe the intensity of involvement based on the number of tweets. As the corresponding numbers for each category in Table IV show, the "vocal" group produced more than twice the volume of tweets from the "silent" group, though it is 57 times smaller in size. While we could observe that the groups are different from each other simply because their members tweet at different rates, that would not be a point of concern. If the nature of every tweet would be roughly the same in all groups (in terms of whom it targets and how it does that), the fact that some people tweet more than others would not be so important. However, if some users construct their tweets with the intention to reach the largest possible audience or repeat and retweet them constantly in order to keep a topic alive, then we cannot put an equal sign between content generated by users of different groups. We discuss now how tweets can be different from each other.

[2]The particular choice of group boundaries is not important as our results hold under any logarithmic scale of growth.

TABLE V
A TWEET MIGHT CONTAIN A COMBINATION OF DIFFERENT KINDS OF ENTITIES.

Content Type	Tweet Text
Only text	Women need to vote for Martha Coakley
Link	Rep. Frank, D-Mass., raises stakes in Coakley-Brown race; commonwealth most powerful force in union. http://bit.ly/85lLCL
Hashtags	A gop in Teddy's seat would be a travesty. Vote COAKLEY #p2 #MAsen
Retweet	RT @gpollowitz: I'm thinking of all the Catholics providing er care in Haiti and wondering what Coakley thinks about that
Mention	Scott Brown - that's right, MA should have universal health care - NO ONE ELSE SHOULD! Hope @maddow does a story here.
Reply	@MarthaCoakley I have you on my home page... Go getum Martha...
Hashtags + link	Scott Brown posed nude in Cosmo. http://bit.ly/MyQ8A HA! #p2 #tcot
Retweet + hashtags	RT @NorsU: Its awesome here in ma talk radio being flooded with callers mocking Coakley ads #masen #tcot go Scott Brown

TABLE VI
DISTRIBUTION OF CONTENT OF TWEETS ACCORDING TO THE USER CATEGORIES.

	Silent	Attentive	Interested	Engaged	Vocal
Hashtags	14.1% (4625)	21.0% (8145)	29.3% (8081)	38.9% (25825)	53.0% (36785)
Links	29.7% (9754)	35.7% (13885)	42.1% (11583)	47.0% (31165)	49.4% (34197)
Retweets	29.6% (9698)	31.7%(12321)	36.3% (9972)	44.3% (29361)	60.3% (41769)
Replies	6.2% (2026)	7.5% (2910)	7.1% (1952)	6.5% (4323)	7.6% (5275)
Only text	42.0% (13770)	31.3% (12155)	22.2% (6119)	14.4% (9545)	8.0% (5564)

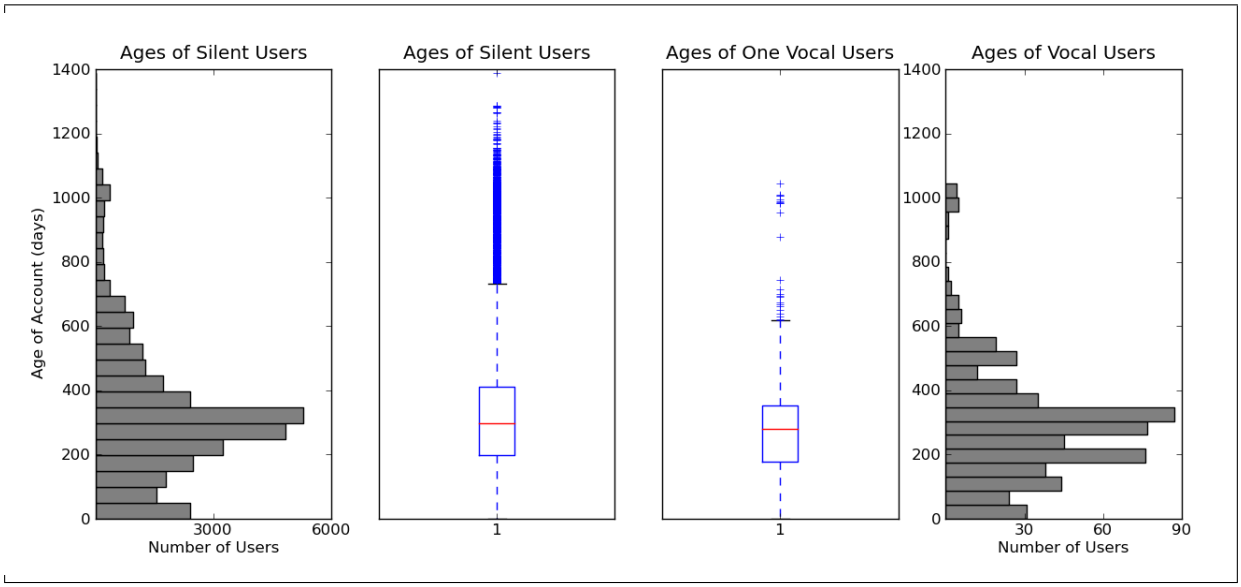

Fig. 2. Distribution of the account ages for the Silent group and the Vocal group. The statistical test indicates that the difference in the average age is significant, and in particular that members of the Silent group have been using Twitter for a longer time than those of the Vocal group.

B. What's in a tweet?

A tweet is a message of no more than 140 characters that a user sends via Twitter to the world. Mostly, people tweet to communicate their thoughts or feelings; to participate in conversations; or to share some interesting story, picture, or video, by providing a link to it. Once a user has sent a tweet, only users who have subscribed to receive updates from the sender will be able to see it. What to do if you want your tweet to be viewed by as many users as possible? There are several available strategies:

1) Insert in the text of the tweet as many hashtags[3] as possible. A hashtag becomes automatically a hyperlink on Twitter, which allows everyone who clicks on it to view the search results of all other tweets containing the same hashtag;
2) Ask other people to retweet your tweet (by writing "Please RT");
3) Send your tweet to a famous (or influential) Twitter user by replying or mentioning them and hope that they will retweet it to all their followers or take some other action.

Thus, a tweet might contain any combination of the following entities: text, links, hashtags, and Twitter handles that can mean: a retweet, a reply, or a mention. For examples of tweets which use and combine these entities, refer to Table V.

[3]A hashtag is a word preceded by the pound sign, e.g., #wimbledon.

The example *Reply* in Table V shows a tweet that starts with a Twitter handle (@MarthaCoakley). It is different from the example *Mention*, because a reply is not viewed by the followers, but only by the user to whom is directed[4]. A tweet that contains a mention, on the other hand, can be viewed by all the followers of the sender, and in addition, the mentioned user gets a notification. The handle @maddow in the *Mention* example belongs to MSNBC journalist Rachel Maddow, who hosts a nightly political show on TV; thus, users try to interest her on political developments of the day, by mentioning her in their tweets. A retweet is different from a reply or *mention* tweet, because it shows that the Twitter handle (usually proceeded by RT) is actually the originator of the tweet and that her message is being forwarded by other Twitter users, who think it is a valuable or interesting message. In the Twitter API, links, hashtags, and handles are commonly referred as entities.

As the examples show, it is possible to combine several entities in a single tweet and many users do that. The question is whether all users of the above-mentioned categories are equally likely to do that. We refer to the process of creating tweets, retweeting, and replying as user behavior and proceed to compare the behaviors of users in the different groups.

C. Comparing User Behavior

Hypothesis: *The members of the "vocal group" tweet differently from the members of the "silent group".*

To test this hypothesis, we inspect the content of tweets (by users of every category) for the aforementioned entities: hashtags, links, and handles (which might correspond to retweets or replies). A summary of the breakdown for each category is shown in Table VI. We consider every entity as an independent feature (which is what the examples in Table V indicate), and check whether the feature is present or not in every tweet. We calculate the mean for every group, as the ratio of the counts (for every present feature) with the total number of tweets in a category. For example, in the "silent group" only 14.1% of all tweets contain hashtags (the number of tweets with hashtags is given in parentheses).

We performed Tukey's HSD ("Honestly Significant Difference") test to find whether differences of the mean for the five categories are statistically significant. The test shows that with a $p < .001$, this is the case for all rows but the *Replies* (which shows only some pairwise significance at .05 level). However, since replies are the form of tweeting which gets the smallest audience (remember, it is only viewed by the person to which the tweet is addressed), they are not part of strategies to maximize the reach and impact of tweets, which (as we discussed previously) would consist in including as many hashtags as possible, providing links to external content which doesn't fit in 140 characters, and retweeting in order to reach a larger number of users.

In all these three features, the "vocal group" shows a clear dominance, which is significantly different from that of the other groups. In fact, the very small number of tweets which contain only text (8.0%, see last row of Table VI), reveals that the members of this group are not interested in simply expressing themselves, but in getting the largest audience for a message.

One could try to explain the difference between the two groups by hypothesizing that members of the "silent group" are less experienced, because they haven't been using Twitter for as long as the members of the "vocal group". In order to test this hypothesis, we calculated the age of all accounts in both groups. This was possible because every tweet contains the timestamp of the message as well as information for the sender (the creation date of the account). Thus, the age of the account was defined as the difference in days between the date of account creation and the day of the (last) message sent which is inside our dataset. The distributions of account ages for members of both groups are plotted in Figure 2. The mean values for the age are: 330 days for the "silent group" and 288 days for the "vocal group". This difference is statistically significant ($p < 0.001$). The results show that users of the "silent group", in average, have been using Twitter longer than users of the "vocal group", which means that they are probably not behaving differently because they have less knowledge of how to use the medium. As it will be discussed in Section IV, one of the reasons for the sophisticated use of Twitter features by users of the "vocal group" could be the training they received as part of using pre-fabricated content from members of their community. We will briefly discuss such communities in the following subsection.

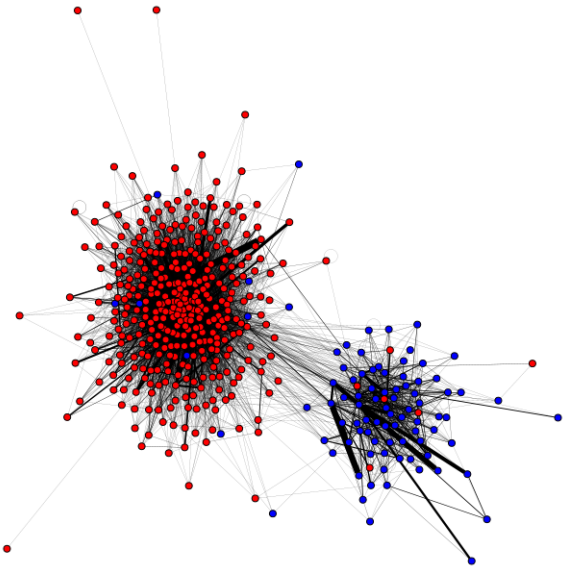

Fig. 3. The two communities within the "vocal group" of users, discovered based on the retweeting behavior. Conservative users are in red, progressive users are in blue. Thicker edges indicate multiple retweets for an account. The labels were assigned using an automatic clustering process, in which every user was represented by a vector of hashtags used in the tweets.

[4]If a third user happens to follow both the sender and receiver, then she will view this message as well.

D. The communities of retweeters

As the results in Table VI indicate, the "vocal group" users retweeted at a rate double that of the "silent group" users. We analyzed the retweeting behavior of both categories and discovered several interesting facts:

- Users of the "silent group" do not retweet users from the "vocal group". In fact, only 6% of their retweets originate from users in the "vocal group".
- Users of the "silent group" retweet famous people and traditional news organizations. Some of the most retweeted accounts by this group were: @barackobama, @cnnbreak, @nytimes, @breakingnews, @senjohnmccain, @gavinnewsom (Mayor of San Francisco), @sethmeyers21 (comedian), @amandapalmer (singer).
- Users of the "vocal group" retweet users with whom they agree politically, which are not famous but are members of their community. This can be seen in the graph shown in Figure 3.

The graph shows clearly two communities of users (inside the "vocal group") who predominantly retweet within the community. Indeed, the graph was drawn by using the force-directed layout, which is commonly used to discover communities. The data underlying the graph consists of all the retweeting pairs (and their weight - the number of times one user retweeted the other) inside the "vocal group". The colors of the nodes were established by a previous process of automatic clustering for these users. Each user was represented by a vector of frequencies of the hashtags they used in their tweets. Users who on the political spectrum are conservatives made frequent use of hashtags such as #tcot (Top Conservatives on Twitter) or #teaparty, while users who see themselves as progressives, use hashtags such as #p2 or #ofa (Organizing For America). In the graph, conservatives are shown in red and progressives in blue. The reasons for some nodes being in the opposite group might be related to errors in the automatic clustering. The fact that there are edges between two communities doesn't mean that the members of the two different communities retweet each-other to increase the audience of their message, but because they retweet to answer to each other, by commenting the original tweet, a phenomenon we discuss in [10]. The polarization in the retweeting behavior evident in Figure 3 is studied in more detail in [11].

IV. Repetition via pre-fabricated Tweets

One of the things we noticed by inspecting tweets of the "vocal group", were repeated examples of messages with identical content, which contained the same hashtags, URLs, and mentions. Messages were sent by different accounts and at different times, thus, it seemed difficult to understand what the relation between these messages was, until serendipitously we came across this tweet:

> RT @RedDevilRio: EVERY TWEET HELPS SCOTT BROWN http://bit.ly/ALLLists #tcot #teaparty #sgp #rush

Fig. 4. Screenshot of a page from the website Patriot Network Action, retrieved June 19, 2011. It shows a message posted on January 11, 2010, that announces creation of new lists with Twitter messages that the members of the website can copy and paste to send from their accounts. Each list has around 60 messages, so that by sending them, users do not risk to get suspended by Twitter for excessive updates. The shortened URL to get access to these lists is: http://bit.ly/ALLLists.

which led us to the website "Patriot Action Network" (http://www.patriotactionnetwork.com), a screenshot from which is shown in Figure 4. The message in the page indicates that tweets have been already composed by an individual (maybe the author of the post) and that members of the website are invited to post them by copying and pasting the text into their Twitter statuses. The author has taken care to put only 60 tweets per list (roughly the hourly limit on Twitter status updates) and to not use a very repetitious language, since Twitter throws an error if two subsequent tweets are very similar. The tweets are targeting news organizations in Massachusetts in order to attract their attention to alleged election fraud by organizations such as ACORN or SEIU (traditionally linked to Democrats). For example, the following tweets:

> WE THE PEOPLE WANT A FAIR ELECTION http://bit.ly/acRNFraud @ACORN_Nat @SEIU @GlobeSenateRace @wwlp #masen
>
> DO YOUR JOB SHINE THE LIGHT ON ACORN http://bit.ly/DoYourJob @ACORN_Nat @SEIU @GlobeSenateRace @wwlp #masen

are intended for the Boston Globe section that covers the election and the WWLP TV station, based on Springfield, Massachusetts. We downloaded the 30 tweets lists available in the website, which contained 2758 ready-to-paste tweets. By parsing the content of the tweets, we counted the number

of Twitter accounts of Massachusetts and national news organizations that were mentioned in the messages. There were 170 account names in total, belonging to organizations and individual journalists. Every account was mentioned in average 24 times, though some of the accounts had more than 100 mentions (e.g., @globesenaterace (183), @wbznewsradio (144), @masslivenews (135), etc.). If dozens of members of the Patriot Network tweeted and retweeted these messages daily during the election week, the total amount of such messages would mimic an event similar to a distributed denial attack of service (DDoS) to news organization accounts, because it floods them with the same message, making it difficult to find other reports and concerns from citizens in their area.

By following one of the URLs in the tweets, http://bit.ly/DoYourJob, we encountered an anonymous message which amounts to a kind of blackmailing for news organization:

> If you want us to stop our Twitter Protest, start reporting what is going on. Start taking some risks in doing investigative journalism. Take off your left-wing rose-colored hippy glasses and see reality for what it is ...

Unfortunately, we cannot report on how successful in terms of user participation this kind of protest was. Since we used the Twitter Streaming API to collect only tweets which mentioned Scott Brown and Martha Coakley, other tweets from the lists of Patriot Network have not ended up in our collection. However, our dataset contains several dozens of the pre-fabricated tweets with the name of the Scott Brown, such as:

> WE THE PEOPLE WILL ELECT SCOTT BROWN http://bit.ly/HereWEcome #masen #tcot #tweet-congress @SEIU @ACORN_Nat

Lack of capturing the other tweets, raises the question of how wide to cast the net when one collects data surrounding an event based on a set of some chosen keywords. We couldn't predict in advance that some users would think that election fraud was a problem and then send thousands of messages to raise awareness about it. However, if in the future access to historical Twitter data becomes available to researchers, it might be possible to study the broader context where the conversation about an election was situated, collecting topics which were peripheral to the candidates, but still tangible in terms of potential impact on the electorate and media.

V. Discussion

In a previous paper [4], we had uncovered a different kind of political campaigning on Twitter, which we dubbed a "Twitter bomb". It consisted of 9 accounts created within minutes of each other, with names such as "CoakleySaidThat" or "CoakleyAgainstU", which in the interval of 138 minutes sent 929 reply messages to Twitter users who had been tweeting about the election in the hours previous to this attack. Since this was an orchestrated and automatic attack, Twitter discovered it and shut down the accounts quickly. However, because the messages were retweeted by other users, they lived longer than the accounts. Creating such attacks on Twitter is easy in terms of programming, but it comes with the risk of your IP address being banned by Twitter, thus, it is not an appealing approach. A solution is to use real people from different computers to do the same thing, without attracting Twitter's algorithms attention for suspicious activity. This is what seems to have happened with the Twitter protests discussed in Section IV.

The kind of attacks mentioned here (either by automatic scripts, or crowd-sourced) raise the need for sense-making tools that are able to discover their nature. Such sense-making tools would be very useful to political campaigns, political analysists, and journalists covering political events. An example (inspired by our work in [4]) is the web-based system Truthy (http://truthy.indiana.edu), which is able to discover astro-turfed meme difusion and their origin [12].

The fact that a vocal minority was able to generate sufficient media buzz and financial support for Scott Brown's electoral victory, makes it particularly interesting to study its power. In [13], the authors offer a thorough and careful explanation of how minority opinions or actions can influence a large majority when the situation is uncertain, such as during an election. Even if the majority does not necessarily agree with the opinion of the minority, it does not wish to be left out, and so the majority acts in accordance with what they think normal group behavior is, even if they do not agree with it. In other words, the Twitter bombing or protest campaigns had the right idea: a strong showing of one opinion can influence the majority even if they disagree with it, especially if there is a relatively narrow majority as the results of the Rasmussen's Reports one week before the election had indicated.

In fact, recent theoretical work based on simulation of opinion spreading in networks has demonstrated that when the size of the minority opinion-holding group increases to more than 10% of the network size, then the minority opinion takes hold and becomes the opinion of the majority [14]. An important research contribution would then be to find how these theoretical results could explain real-world events, such the one discussed in this paper, and whether such a model could have predicted in advance how the dedicated group of vocal users was influencing the general network.

VI. Conclusion and Future Work

Collecting user-generated data on the web is particularly appealing, because of the large amount that is available. This quantity might encourage researchers in believing that all content is equal, since it is generated by random users in a very large sample. However, in this paper we showed that there is a spectrum of users who engage in different ways with social media. At least between the two groups at the extremes of this spectrum, to which we refer as the vocal minority and the silent majority, there are significant differences in the tweeting behavior. The vocal minority users link more to outside content, use more hashtags, and retweet more, all activities intended to broaden the impact and reach of tweets. Because of this difference between the content generated by these two groups, one should be aware of aggregating data and

building models upon them, without verifying the underlying model that has generated the data.

There are several ways in which we are trying to extend the results shown in this paper. While the focus here was on user behavior and tweet structure, different kinds of content analysis can provide even more information. For example, one can distinguish between subjective tweets (epxression of opinions about the candidates or certain policy issues) or objective tweets (newspaper headlines, fact statements). Our initial explorations have indicated that silent users are more likely to write subjective tweets expressing personal opinions, while the vocal users are more likely to tweet about particular happenings (though the headlines used in such occasions can be hardly labeled as objective because they are politically biased, e.g., *In Coakley Hullabaloo, Assaulted Reporter's Question Are Still Unanswered http://dr.tl/263273 #tcot #masen*).

Additionally, one can perform topical modeling to discover topics which capture the interest of a large group of different users. Because topical analysis it is usually performed over an aggregrated set of tweets, it would benefit by the separation of tweets by different groups of users, which will then allow the discovery of what is important to these different groups.

ACKNOWLEDGMENT

The work of E. Mustafaraj and P. T. Metaxas was supported by NSF grant CNS-1117693. The work of S. Finn and C. Whitlock was supported by a Wellesley College Science Center grant. We thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] S. Asur and B. A. Huberman, *Predicting the future with social media*. CoRR abs/1003.5699, 2010. http://arxiv.org/abs/1003.5699

[2] J. Bollen, H. Mao and X.-J. Zeng, *Twitter mood predicts the stock market*. CoRR abs/1010.3003, 2010. http://arxiv.org/abs/1010.3003

[3] V. Lampos, T. D. Bie and N. Cristianini, *Flu detector - tracking epidemics on Twitter*. Machine Learning and Knowledge 6323, 599–602, 2010.

[4] P. T. Metaxas and E. Mustafaraj, *From obscurity to prominence in minutes: political speech and real-time search*. In WebSci10: Extending the Frontiers of Society OnLine, April, 2010. http://bit.ly/h3Mfld

[5] B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, *From tweets to polls: linking text sentiment to public opinion time series*. In Proc. of 4th ICWSM, AAAI Press, 122–129, 2010.

[6] A. Tumasjan, T. Sprenger, P. G. Sandner, and I. M. Welpe, *Predicting elections with twitter: what 140 characters reveal about political sentiment*. In Proc. of 4th ICWSM, AAAI Press, 178–185, 2010.

[7] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[8] M. Wagner, *Obama Election Ushering In First Internet Presidency*. In Information Week, November 5, 2008. http://www.informationweek.com/news/government/212000815.

[9] E. Swanson, *MA: Coakley 49 Brown 47 (Rasmussen 1/11)*, January 12, 2010, http://www.pollster.com/blogs/ma_coakley_49_brown_47_rasmuss.php.

[10] E. Mustafaraj and P. T. Metaxas, *What edited retweets reveal about online political discourse*. In Proc. of AAAI Workshop "Analyzing Microtext", August, 2011.

[11] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. *Political polarization on Twitter*. In Proc. of 5th ICWSM , AAAI Press, 2011.

[12] J. Ratkiewicz, M. Conover, M. Meiss, B. Goncalves, S. Patil, A. Flammini, F. Menczer. *Truthy: mapping the spread of astroturf in microblog streams*. WWW (Companion Volume), 249-252, 2011.

[13] C. Bicchieri and F. Yoshitaka, *The Great Illusion: Ignorance, Informational Cascades, and the Persistence of Unpopular Norms*. Business Ethics Quarterly, Vol. 9, No. 1, 127-155, 1999.

[14] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, B. K. Szymanski, *Social consensus through the influence of committed minorities*. 2011. http://arxiv.org/abs/1102.3931.