#Snowden: Understanding Biases Introduced by Behavioral Differences of Opinion Groups on Social Media

Q. Vera Liao
University of Illinois at
Urbana-Champaign
Champaign,USA
liao28@illinois.edu

Wai-Tat Fu
University of Illinois at
Urbana-Champaign
Champaign,USA
wfu@illinois.edu

Markus Strohmaier
GESIS & U. of
Koblenz-Landau
Cologne, Germany
markus.strohmaier@gesis.org

ABSTRACT

We present a study of 10-months Twitter discussions on the controversial topic of Edward Snowden. We demonstrate how behavioral differences of opinion groups can distort the presence of opinions on a social media platform. By studying the differences between a numerical minority (anti-Snowden) and a majority (pro-Snowden) group, we found that the minority group engaged in a "shared audiencing" practice with more persistent production of original tweets, focusing increasingly on inter-personal interactions with like-minded others. The majority group engaged in a "gatewatching" practice by disseminating information from the group, and over time shifted further from making original comments to retweeting others'. The findings show consistency with previous social science research on how social environment shapes majority and minority group behaviors. We also highlight that they can be further distorted by the collective use of social media design features such as the "retweet" button, by introducing the concept of "amplification" to measure how a design feature biases the voice of an opinion group. Our work presents a warning to not oversimplify analysis of social media data for inferring social opinions.

Author Keywords

Social media; controversy; Twitter; social opinion; online opinion space; opinion minority; bias.

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Web-based interaction

INTRODUCTION

People discuss a broad range of controversial topics online, from social-political issues, current news and events, to reviewing consumer products. These data have become valuable resources for research and applications in many domains to assess, monitor and predict public opinions, such as sociology, political science, and marketing.

Recently, scholars have increasingly criticized the widely used analytics that rely on "pooled" contents from social media (e.g., extracted by keywords search) to study public opinions [39, 24], typically with measurements of volume or sentiment, which are especially popularized by convenient text analysis software such as LIWC. However, the mounting evidence that social media analytics may fail to predict elections and public polls (see review in [24]) suggests that, while these methods may be effective in assessing simple sentiment, they may prevent insights into complex social-political issues with the false assumption of motivations and behaviors of different opinion groups being generic and constant.

Understanding behavioral differences between opinion groups and the biases they introduce is no trivial issue. For example, differences in group *expressiveness* may introduce quantitative biases to the presence of opinions on a social media platform. Some groups may be more active in posting messages thus become over-represented [35]. Some groups may engage in different *communication patterns* and produce contents with distinctive characteristics [19], which are important to consider for content based analytics. For instance, groups that engage in more discussions or advocacy are more likely to produce expressive, opinionated contents.

While there could be many reasons for groups to exhibit different behavioral patterns, we highlight the impact of platform-specific social environments, including *interactions* with like-minded and different-minded others. Importantly, it may drive behavioral changes that are not necessarily associated with changes in opinions. For example, some groups may gain power through stronger alliance or more effective cooperation [19, 46], or by closely centering around influential individuals such as celebrities or activists [38, 50]. Such changes may systematically skew the opinion space in the long run, which is important to consider for opinion monitoring purposes.

To shed light on how social factors may shape group behavior changes, we study numerical majority and minority opinion groups, where the platform constitutes a fundamentally different social environment for the two. We are interested in whether the minority group would experience strong social pressure and withdraw their participation [36] or whether they would leverage the social connections enabled by social media to gather with like-minded people and enclave in "echo

CHI'16, May 07 - 12, 2016, San Jose, CA, USA.

ACM 978-1-4503-3362-7/16/05$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858422

chambers" [46]. In either case, social interactions could lead to changes of their presence in the opinion space.

In addition, we also study another factor that could distort an online opinion space— the asymmetrical effect of the collective use of a *platform design feature*. Collective use of social media design features often change the presence, visibility or influence of messages. For example, on Twitter some messages would be retweeted more, and thus be more visible than others. On Facebook or Reddit some messages would be liked or up/down voted more, which changes their algorithmic ranking and visibility. On a product review website some reviews might be endorsed more and appear to be more convincing. When such changes happen at an aggregate level, some groups' voices may become *amplified* while others' suppressed. For example, when there are numerical majority and minority groups, the sheer size difference may lead to varied effectiveness of the use of design features, and potentially put the minority group in further disadvantage.

Taking together, dissecting the complex, varying group dynamics that shape an online opinion space represents a pressing practical and scientific challenge. Such knowledge is crucial for developing better methods for using social media data to study public opinions. Moreover, by understanding how different opinion or ideological groups behave, it may enable us to identify more general mechanisms in behavioral and social phenomena concerning different groups.

To explore these issues, we present a case study of 10-month Twitter discussions on Edward Snowden, the former NSA subcontractor who made global headlines by leaking secret documents that expose the NSA's global surveillance program. Especially in the United States, Edward Snowden is a controversial figure who fueled much debate on government surveillance and information privacy. He has been considered a "hero" and "patriot" by many, but also called a "traitor" by some for the potential threat he posed to the national security. Although media polls in the US showed mixed results of public opinions on Snowden [1, 2, 4], there has been anecdotal evidence that Twitter users are dominantly leaning towards pro-Snowden [3]. Strikingly, when we looked at the top 100 most retweeted tweets in the dataset, we found only one of them casted slight doubt, but almost 40 of them expressed strong support for Snowden. We also expect the Snowden discussions to be especially suitable to study group behavior changes given the continuous interests it generated over time.

The Twitter population is more likely to fall into the known pro-Snowden categories, e.g., younger, tech-savvy, from foreign countries. Our focus is to explore how this "biased" environment drives behavioral differences, including activity levels and communication patterns, between the numerical majority and minority groups. We will explore how the difference, if any, was driven by social factors by studying their in-group (with opinion-similar users) and out-group (with opinion-different users) interactions. We will also study how their presence was skewed by their potentially different use of Twitter design features. Although previous research studied opposing opinion groups on social media [18, 43, 47, 49], the case of numerical majority and minority groups has not yet been well explored. Specifically, we ask:

- RQ1: How did the pro-Snowden and anti-Snowden groups' activity levels in the opinion space change over time?
- RQ2: Did the pro-Snowden and anti-Snowden groups engage in different communication patterns? How did they change over time?
- RQ3: How did the pro-Snowden and anti-Snowden groups engage in social interactions with in-group and out-group members? How did they change over time?
- RQ4: How did the collective use of platform design features distort the opinion space? How did it change over time?

An important argument we make is, when studying online opinion spaces, one should separate the production of original messages from the effect enabled by platform design features (such as likes, retweets) as it may amplify the presence of certain groups. As social media provides a plethora of social features, such as sharing, recommending, and rating, we urge to study how different design features amplify online opinion spaces and their implications for opinion groups, especially the marginalized and minority groups. The knowledge could inform more user-friendly and also more ethical design of social media. To facilitate such effort, we introduce a measurement for the amplification effect of a design feature, and demonstrate its use with a Twitter feature — the retweet button. We follow up with discussions on developing a method to assess the amplification effect on a more general level, for other design features.

In the remainder of the paper, we will first review related literature, and then discuss the dataset and methodology we used to identify pro-Snowden and anti-Snowden users. To answer RQ1-RQ3, we will examine the original tweets dataset by excluding retweets. To answer RQ4, we will first introduce the metric to measure amplification effect, and then use it to analyze the retweets data. Based on the results, we will discuss possible underlying factors driving the behavioral differences of majority and minority groups we observe, and their implications for using social media data to study public opinions.

RELATED WORK

Previous research studying online opinion spaces mainly targeted two goals: to understand social behaviors, and to study public opinions. Considerable research effort was made to explore ideological segregation and polarization. Adamic et al. studied the citation network of political blogs and found a structure divided by ideological affiliations [5]. Conover et al. demonstrated high homophily in the retweeting network of Twitter users' political discussions but more heterogeneity in conversational network [18]. Recent studies looked at event based controversy on Twitter and draw similar conclusions that Twitter is primarily used for spreading information to like-minded others [43]. While people occasionally engage in conversations with different-minded users, they are often to reinforce group affiliation, e.g.,expressing disagreement, rather than to engage in meaningful deliberation [49].

For many years scholars have been studying and improving the capabilities of using social media data to study public opinions. Many attempted to predict election results (see review in [24]), public polls [37], stock market [7], and consumer opinions [27, 44]. They often took a wholistic view and performed analysis on the "pooled" contents extracted by keywords. However, such method received increasing critics for its underlying assumption of individual behaviors being uniform and invariable. Gayo-Avello systematically reviewed the evidence that Twitter data failed to predict election results and pointed out a major flaw is treating all parties equally regardless of members' demographics, motivation and self-selection bias in tweeting a topic [24]. Ruths et al. warned that large-scale studies using social media data need to be held to higher methodological standard to account for population bias, biases in individual behaviors and biases driven by platform and algorithm designs [39].

Aligning with this view, we aim to use a case study to highlight that an online opinion space could be skewed by behavioral differences of opinion groups, which could be driven by social factors and platform design. While limited, a few previous studies provided relevant evidence by comparing the Tweeting activities of opposing or competing groups. [19] found that right-leaning Twitter users, compared to left-leaning ones, exhibit greater levels of political activity, tighter social bonds, and a communication network topology that facilitates rapid dissemination of information. [15] found that leaders of a minority party tend to engage in more conversations with their followers. By analyzing Secular versus Islamic polarization on Twitter, [8] found that they use distinct sets of hashtags to frame political issues in different ways. We also highlight recent studies on "vocal group" and "silent group" on Twitter, which found them to differ significantly on communication patterns — with the former more likely to adopt tweeting strategies that intend to broaden impact, e.g., sharing links, using multiple hashtags and retweeting more [35]. The vocal and silent groups were also found to differ in their power to predict election results [14].

We also draw attention to the potentially unequal impact of design features on different opinion groups. Research has highlighted that collective use of social features may create unintended and undesired inequality. For example, many warned that social sharing and collaborative filtering features may trigger uninformed conformity and create information cascade [6, 40], which leads to even greater inequality between popular and unpopular items. We argue that these design features may potentially create effect that differs among opinion groups thus inequity that skews the opinion space. In this paper, we study the *outcome* of this potential inequity by examining its bias, size and changes over time.

Lastly, we ground our study in the long history of social scientists studying divergent groups, especially majority versus minority group behaviors in the physical world (see review in [42]). Among others, a lasting interest is to understand the phenomena that the numerical minority group tend to exhibit stronger ingroup bias and intergroup discrimination, i.e., stronger affiliation with group members and discriminative behaviors towards the majority group. There are two typical arguments explaining the phenomena [29]. The first concerns with the group salience associated with its smaller size, which engenders a stronger sense of social identity and concerns to maintain it positively. The second centers on the "insecurity" and perceived threats associated with categorization in a relatively disadvantaged or vulnerable group.

The ingroup bias results in "echo chamber effect" — opinions reinforced through interactions with like-minded others [46]. While often warned to be a potential danger to social stability by leading to opinion polarization and social fragmentation, Sunstein pointed out the benefit of enclave deliberation for the minority and marginalized groups, as "it may be the only way to ensure that those views are developed and eventually heard." [46] Less optimistically, other research suggests that opinion minorities may have a hard time surviving because of social compliance [16], which leads to compliance with the majority view, and "spiral of silence" [36], that the fear of isolation may force one to remain silent about the minority view. As a result, the minority group may fall weaker in spite of their actual distribution in a society. However, failed to observe spiral of silence in many empirical studies (e.g., [41]), researchers pointed to the moderating effect of attitude certainty and the existence of "hardcore minority", i.e., people holding strong attitude who would speak out regardless of the climate. In this study, we will explore how the engagement of participating opinion minorities evolved over time.

DATASET OVERVIEW

By using Twitter Streaming API, we collected all publicly available tweets containing "#snowden" (case insensitive). The collection started on July 6, 2013, one month after Snowden's first public appearance, and ended on April 25, ranging for 42 weeks. Including both original tweets and retweets, the dataset included 1.06 million English tweets. We used the Python NLTK package to exclude non-English tweets as we chose to focus on the English-speaking opinion space.

The dataset contained meta-data indicating whether a tweet is a retweet and if so, the ID of the original tweet. The meta-data is automatically generated when user used the "retweet" button. There are other ways people retweet, e.g., by manually copying and adding "RT", "retweet", etc. However, we counted tweets that contain these markers and only 7.9% of them are not shared through the retweet button. This number is significantly lower than what was reported in earlier studies (e.g., [45]), which were conducted not long after Twitter introduced the retweet button. It shows users' adaption to platform-provided features. Given that we are interested in the effect of the retweet button, we focus on automatically generated retweets only, and treat hand-copied ones as original tweets — to some extent, they are similar to tweets posted to share information. Excluding retweets, the dataset includes 440k original tweets. We also used the meta-data information to track how many times an original tweet was retweeted.

Figure 1 shows the volume of tweets by week. The trend is generally consistent with Google Trends data showing the number of times "Snowden" was searched. There was a burst of attention following his first media appearance, and it decreased rapidly in the first couple of months. From October 2013 onwards tweeting activities were generally stable. How-

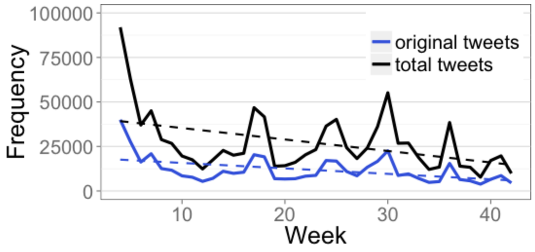

Figure 1. Total and original tweets volume by week

ever, there were “local peaks” following significant events such as Snowden’s Christmas speech, SXSW speech, etc., showing that there were continuous topical interests.

USER CLASSIFICATION

Inferring Twitter users’ opinions is a known challenging task [17, 43]. After experimenting with different approaches, we chose to follow [18] by using a variation of a Raghavan’s label propagation algorithm with manually labeled “seeds”. The method is based on the robust observation that retweeting is an endorsing activity [9] and one’s position can be inferred based on whom or what one retweets [12, 23, 47, 48].

For the retweeting network, we treat each user as a node, and each time a user retweeting another user as an out-going edge. The algorithm works in iterations: in each iteration, we start with a pro-Snowden cluster and an anti-Snowden cluster of nodes. Then we iterate every node in the dataset to assign or re-assign it to the cluster that it has more out-going edges to. When they are equal (but not 0), we assign it to the minority group to compensate for the smaller group size. We use the result of the current iteration to seed the next iteration, and we use a hand-labeled sample with confidently known positions to seed the initial iteration. We always fix the cluster of these known sample, and give outgoing edges to them a slightly higher weight by setting weight for edges to non-seeds at 0.8.

To obtain a sample of seeds, we crowdsourced opinion labels for 3% of random sample of daily original tweets. We chose to label tweets instead of users because it was easier and more suitable for crowdsourcing tasks. Random sampling also made it more likely to include active users, which are better positioned for seeding the network clustering algorithm. We had three Turkers to label each tweet. They were asked to only label a tweet if they could confidently judge its position, otherwise leave it in the ”unknown” category. If the three selected different labels thus did not reach a majority vote, we recruited 2 more Turkers to label the same tweet. We also had a researcher verify labels for tweets that did not get high agreement. Eventually, 23.5% of sampled tweets were confidently judged with positions. With this ”conservative” labeling schema targeting high precision, we classified authors of tweets labeled as pro or anti to be pro-Snowden or anti-Snowden seeds. Only 0.6% of them appeared in both groups, and we manually examined their tweets to decide their positions, if possible. We ended up with a sample of 519 anti-Snowden and 1,922 pro-Snowden seeds.

We used the labeled sample to seed the above mentioned algorithm, which reached stability at the 56th iteration. We focused on users who actively participated in the discussion, i.e., those who produced at least one original tweet. We were able to identify 27,119 pro-Snowden and 2,400 anti-Snowden users. Their tweets accounted for 69.1% of the whole original tweets dataset. The cumulative distributions of the number of tweets of the two groups are fairly similar, with a few more extremely active users in the pro-Snowden group.

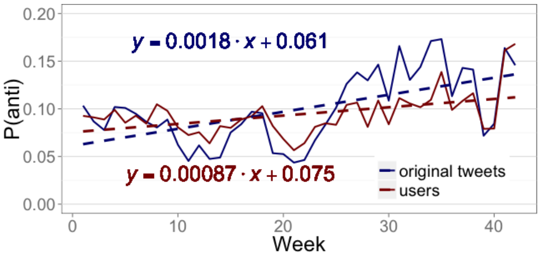

Figure 2. Percentage of tweets and users from anti group by week

For evaluation, we randomly sampled 200 pro-Snowden users and 100 anti-Snowden users from our identified users. We pooled them together and retrieved all their tweets, then had researchers to rate whether each person leans towards pro or anti-Snowden, or whether they could not judge because either the person had too few tweets or was only sharing factual information. Table 1 presents the evaluation results. In general, the classification method gave satisfactory results, with around 90% accuracy for each category.

Top Influencer

In some analyses, we explore the role that opinion leaders play — people who exert very high influence within each group. Cha et al. [13] concluded that the number of retweets a user has received can reflect a user’s topical influence. We therefore chose to identify the *top influencers* based on the total number of retweets they received. For each opinion group, we identify users above the 50% of the cumulative distribution of number of retweets received to be the top influencers — this was to account for the fact that the retweet distribution of pro-Snowden group has a heavier tail. 14 anti-Snowden users and 44 pro-Snowden users were identified to be top influencers. Note that this method may leave room for improvements. However, having tried more sophisticated methods, e.g., also considering mentioning, the results are generally correlated and do not change the conclusions. Hence we present the results by applying the simplest method.

RESULTS: ORIGINAL MESSAGES

In this section, we focus on the production of original messages by excluding retweets. We are interested in the pro- and anti-Snowden groups’ activity levels (RQ1), differences in communication patterns (RQ2), how they interacted with opinion-similar and opinion-different others (RQ3), and how these activities changed over the 10-month period.

Change of Activity Levels

We start by seeking answers to RQ1 — to understand the changes of the activity levels of the anti-Snowden group. We are interested in knowing whether the minority group sustained their presence in the opinion space. We look at the relative size of *tweets volume* and *participating users* (who posted

Group	agree	not agree	cannot judge
pro	90.7%	1.3%	8.0%
anti	89.0%	4.0%	7.0%

Table 1. User classification evaluation results

original messages) from the anti-Snowden group. The overall activity change has been discussed in the data overview section (Figure 1). In Figure 2, we present the relative proportion of tweets from the anti-Snowden group among all the identified users. Results are calculated by the unit of week, i.e., we count how many users posted tweets within the week. Varying from week to week, there were 5% to 45% of users among all the identified users participating.

Following [20], we study temporal changes by the linear fitted *trend line* with weekly data points, which allows us to compare the temporal changes of groups by testing the difference between their slopes (with t-test for the coefficient difference divided by the pooled standard error). Both the proportion of tweets volume ($t(40) = 3.41, p < 0.01$) and the size of users from the anti group ($t(40) = 2.62, p = 0.01$) show slopes significantly higher than 0, and the difference between the two is significant ($t(80) = 1.94, p = 0.05$). It suggests that over time the anti-Snowden group's tweeting activities were better sustained than the pro-Snowden group. Also, on average, the anti-Snowden users became more active than those in the pro group in producing original tweets, suggested by the higher proportion of tweets than proportion of users. This is an interesting observation suggesting that, instead of silencing themselves, the minority group increased their presence in the opinion space (of original messages). In the following sections, we will attempt to understand the reasons.

For all the temporal changes we study in the paper, we need to examine an alternative explanation that the changes could be caused by a significant change in the population of the two groups, instead of group behavioral differences. To do so, we look at the distribution of members joining week (the first time one tweeted #Snowden). A K-S test comparing the distributions of the two groups showed no significant difference (p=0.58). Also, we looked at the top active users above 50% of the cumulative distribution of the total number of tweets for each group, 92.3% in the anti and 95.6% in the pro-group were already tweeting in the first half period, and 93.6% in anti and 93.9% in the pro-group were still doing so in the latter half, suggesting that highly active users were generally persistent. Therefore, we can reasonably conclude that the temporal changes we observe should not be due to significant differences in the population changes between the groups.

Communication Patterns

In RQ2 we are interested in whether the two groups exhibited differences in their use of the platform, and thus produce contents with different characteristics. Finding such differences would have interesting implications for conducting content based analysis on controversial social media discussions.

Twitter provides widely used "markers" to inquire about communication patterns at the aggregate level. For events related discussion, Bruns et al. [10, 11] suggest that using two metrics—the percentage of original tweets containing URL, and the percentage of retweets among all tweets — one can classify if the discussions focus more on information sharing (high URL and retweets percentage), or more on making original commentary (low URL and retweets percentage). By sampling Twitter discussions from 40 hashtags related to political, social and entertainment events, they demonstrated that they distinctively fell into one of the two clusters [10].

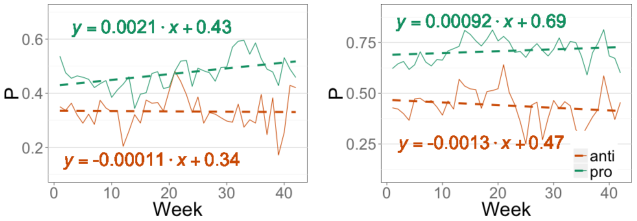

Figure 3. (left) Percentage of retweets among all tweets by week; (right) Percentage of original tweets containing URL by week

Figure 3 shows the percentage of retweets (left) and tweets with URL (right) by week for each group. We can make two main observations. First: the anti-Snowden group focused more on making original commentary, characterized by lower percentage of URL shared and lower percentage of retweets. This may resemble an "*audiencing*" experience where people watch and comment on unfolding events. [10] showed that tweets related to media events, such as political elections, where Twitter acts as a discussion back channel, often collectively exhibit such pattern. The pro-Snowden group appear to have centered more on sharing information, with higher percentage of URLs shared and a higher percentage of retweets. It may represent a "*gatewatching*" practice, with widespread desire to disseminate key information and broaden group influence, but limited interest in posting original comments. Second, we found that the differences became even more distinct over time, with the anti group becoming more commentary oriented, while the pro group focusing more on disseminating information. Comparison between the coefficients of the trend lines of the two groups is significant for retweets percentage ($t(80) = 2.19, p < 0.03$), and marginally significant for URL percentage ($t(80) = 1.80, p < 0.07$).

The above results imply that the anti-Snowden group became more active and also possibly more expressive over time. Imagine using a "pooled data" approach to study opinion changes on Snowden, one may draw conclusion that the climate has turned against Snowden — however, could the behavioral changes be driven by other reasons? We highlight one important factor to consider—differences in social interactions. Observing that 52.4% tweets from anti-Snowden group and 40.8% tweets from pro-Snowden group contained at least one mentioning, which signals inter-personal interactions, we note that changes in social interaction patterns likely contributed to changes in the overall presence of opinion groups. More importantly, we argue that group differences in social interactions may construct dissimilar social environments that could systematically drive group behavior changes. To shed light on such underlying mechanism, we study the differences in in-group and out-group interactions between anti and pro-Snowden groups in the next section.

In-group and Out-group Interactions

We will answer RQ3 by studying in-group (mentioning opinion-similar users) and out-group (mentioning opinion-different users) interactions. While mentioning can be used for different purposes, e.g., to reply, to initiate conversation, or to refer to another user, we do not differentiate but generally consider them as signaling interactions at the interpersonal level.

In Table 2 we present the number of total in- and out-group mentioning degree. For tweets that contain multiple mentions falling in one category, we count them as multiple instances. We also compare the ratio between the observed mentioning degree to the *expected degree* in the mentioning network, calculated as follow:

$$D[i- > j] = d_i \cdot \frac{U_j}{U_{anti} + U_{pro}} \qquad (1)$$

where $i, j = anti$ or pro. d_i is the total number of mentioning from group i to all the users we included in the analysis, and U_j is the total number of users in group j.

As shown in Table 2, we found that the minority group had more intergroup interactions than the majority group. However, the observed within-group interactions in anti-Snowden group is still significantly higher than the expected value, and the ratio was much higher than that of the pro-Snowden group. It suggests that both groups exhibited ingroup bias in their mentioning behaviors, with higher bias in the minority opinion group. However, due to the large size difference, the minority group still had more intergroup interactions than ingroup interactions.

We then examined the temporal changes of in-group and out-group interactions. We first looked at the average total mentioning degree by week (averaging by people who produced original tweets). As shown in Figure 4 (left), over time mentioning activities increased in the anti-Snowden group, but decreased in pro-Snowden group (the difference is significant at $t(80) = 2.10, p = 0.04$). We then examined, for those mentioning others, their average percentage of in-group mentioning and out-group mentioning. As shown in Figure 4 (middle), over time the in-group mentioning increased for anti-Snowden users, but decreased for pro-Snowden users, with the two trend lines significantly differed ($t(40) = 2.52, p = 0.01$). In contrast, as shown in Figure 4 (right), we did not observe such change for average out-group mentioning percentage, with no significant difference between the two groups ($t(40) = 0.03, p = 0.97$), and qualitatively we saw a decreasing trend in the later period for the anti-Snowden group. The results suggest that, compared to the pro-Snowden group, members of the anti group engaged in increasing inter-personal interactions over time, which also increasingly focused on like-minded others.

To further validate that the minority group had a higher "internalizing" tendency, we compare the change of mentioning network of the two groups. We construct a directed and unweighted network by drawing an edge from a user to another if the former ever mentioned the latter — this would exclude the possibility that the increasing in-group mentions was merely due to increasing interactions between some pairs. For both groups, we compare the mentioning networks for the first half period (1-21 week) and second-half period. In Table 3, we present the key network topology metrics.

	N		Ratio to Expected Value	
	→ Anti	→ Pro	→ Anti	→ Pro
Anti	4639	7797	4.59	0.68
Pro	5150	71691	0.82	1.02

Table 2. Total mentioning degree between and within groups

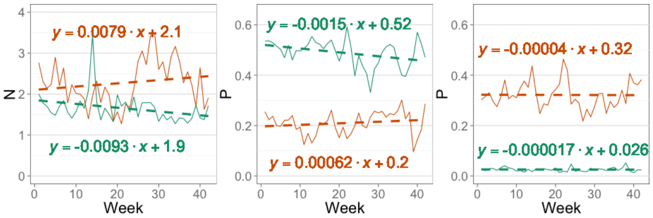

Figure 4. (left) Average mentioning degree; (middle) Average in-group mentioning percentage by person; (right) Average out-group mentioning by person for pro (green) and anti-Snowden (red) groups

For the anti-Snowden group, comparing the later to the earlier period, we found increases in the average degree, connectedness, reciprocity and clustering coefficient, and decreases in distance, all suggesting a more tightly-interconnected mentioning network over time. In contrast, for the pro-Snowden group, while the size of the mentioning network decreased to 59.4% in the latter half, compared to 73.9% for anti-Snowden group, the average distance increased, while all the other metrics decreased. It suggests that the mentioning network of anti-Snowden group became more inter-connected over time, but that of the pro-Snowden group became less so. Moreover, we found that the anti-Snowden mentioning network had much higher reciprocity, suggesting more reciprocal interactions, thus more bi-directional conversations.

The above results suggest that the increasing activities of anti-Snowden group might be attributable to members who became increasingly active in interacting with like-minded others. To explore this, we looked at the top 50 users who were most active in the anti-Snowden group during the later half period (week 22-42), 80% of them were also among the top 50 with most mentioning, and 72% were among the top 50 with most in-group mentioning. The implication is that active members in the minority group might have become increasingly connected by seeking reinforcement in "echo chamber". We will further explore this question by examining the content of in-group and out-group interactions in the next section.

Content of In-Group and Out-Group Interactions

To further unpack the in-group and out-group interactions, we conducted qualitative analysis on the content of tweets containing in-group and out-group mentioning. For each of the four types of mentioning, anti → anti, anti → pro, pro → pro, pro → anti, we draw a random sample of 250 tweets from the pool. Researchers performed an iterative coding process

	Anti - 1	Anti - 2	Pro - 1	Pro - 2
# Nodes	**766**	566	**10539**	6260
Avg Degree	1.75	**1.86**	**2.52**	2.26
Avg Distance	**4.73**	4.18	6.10	**6.56**
Connectedness	0.11	**0.19**	**0.18**	0.12
Reciprocity	0.115	**0.118**	**0.043**	0.037
Clustering Coefficient	0.067	**0.081**	**0.062**	0.059

Table 3. Network topology metrics for mentioning networks over time (1st half vs. 2nd half). larger values in bold.

[34] to identify the *intention* of the mentioning. Four themes of mentioning intention emerged in the codes:

1) **Conversing**. It is typically part of a dialogue or to start a conversation. Consistent with previous findings [33, 49], we observed that a significant portion of in-group conversations were to show support for like-minded others, but out-group ones mainly expressed opposition, disapproval, or to question and provoke different-minded others, e.g.:
"@mygirls3333 EXACTLY! Rachel is FULL-OF-SHIT on that! She doesn't mention #Snowden since he ran to Russia!"
"@thetomtatum Actually, that is incorrect. The majority of the public supports #Snowden."

2) **Directing**. We observed an important use of mentioning is to direct information or facts to the targeted users [1]. These tweets were often to share factual updates about the Snowden issue or relevant external resources. Examples include:
"ICYMI (07/13) Alleged #Snowden Statement Clouded With Skepticism http://t.co/k4VdrBeZ6S cc @LibertyLynx @20committee @catfitz"

3) **Referencing**. It consists of tweets where users cited what the mentioned users have said. Different from directly using the retweet button, we observed that they often rewrote or adapted the original tweets by shortening or summarizing them, and also, more than half of them added additional comments. They were often used to call out opinion-different users to question, criticize, or mock their statements, or to support or endorse opinion-similar users'. For example:
"@YourAnonNews: Venezuela says it will shelter #Snowden: http://t.co/DhYb8F9UY9 #WeStandWithEdwardSnowden" While this can be seen as a form of sharing similar to retweeting, previous research [9] documented the conversational and relational aspect of adapting and commenting through manually sharing, with the intention to publicly agree or disagree, to start a conversation, to make visible one's presence as a listener, to signal friendship or loyalty, etc. Given the currently wider use of the retweet button, these motivations to manually reference could be especially noteworthy.

4) **Pointing**. Lastly, we found that mentioning is also used to point to a person when publicly addressing him or her. They are intended for the general audience but also making the mentioned person aware. For example:
"@LouiseMensch @grantshapps @MailOnline all gone quiet on #Snowden since Merkel concern. Thank god 4 the @guardian and #snowden for exposure"

We present the frequency of code occurrence for each category of mentioning in Table 3. The following conclusions could be drawn from the results:

Echo chamber and audiencing effects: The minority group were significantly more likely to engage in in-group conversations than the majority group ($\chi^2 = 29.67, p < 0.001$). This provides explanation for the previous observation that the mentioning network of the anti-group had higher reciprocity and became more inter-connected over time, as they were more likely to form meaningful social relationships through bi-directional conversations. It also suggests that these in-group conversations could have contributed to the increasing "audiencing" practices we observed earlier. Importantly, we consider it as evidence that the minority group were more likely to engage in exchanging reinforcing opinions, thus exhibiting stronger "echo chamber effect".

Cooperation effects: Directing was generally more likely to happen in in-group than out-group communications, and the in-group directing happened more in anti- than pro-Snowden group ($\chi^2 = 7.10, p < 0.01$). Qualitatively, in-group directing could be considered an cooperative behavior for people to share information or external resources that support the group position. A typical example of the anti-Snowden group directing supportive fact is:
"Americans support #Snowden? In latest poll 60% say he hurt US security; 52% want him charged. Some support. http://t.co/YI3Ptv4tDk @Skipease"
It suggests that the minority group adopted more cooperative strategy by sharing supportive resources with the alliances.

Gatewatching effects: Referencing is generally more likely to happen in-group than out-group, and pro-Snowden group had significantly more in-group referencing than that of the anti group ($\chi^2 = 27.07, p < 0.001$). It suggests that the pro-Snowden group engaged more in sharing and publicly supporting other group members to signal their group affiliation. This is again consistent with the previous observation that the majority group focused more on the "gatewatching" practice.

The above results further supported the fact that the anti-Snowden group exhibited stronger "echo chamber effect" by engaging in reinforcing conversations and cooperatively sharing resource with like-minded others. It supports observations from a previous study where members of a minority political party were more likely to engage in conversations with their followers on Twitter [15]. We point out that these observations are consistent with phenomena frequently observed in the offline world [42] where the numerical minority groups tend to show stronger in-group bias. Such underlying mechanism could have contributed to the increasing presence and expressiveness, i.e., the "shared audiencing" practice, of the anti-Snowden group, which could have been critical for the marginalized group to sustain and develop [46].

RESULT: AMPLIFICATION

Having observed that the anti-Snowden group had an increasing presence regarding the production of original tweets, we now study whether and how the collective use of the retweet button distorted the presence of opinion groups by impacting the messages from majority and minority groups differently (RQ4). We introduce the concept of *amplification* to reflect the asymmetrical changes of message presence, visibility or

	anti-in	anti-out	pro-in	pro-out
Conversing	50%	56%	26%	59%
Directing	28%	10%	18%	14%
Referencing	20%	17%	42%	19%
Pointing	5%	18%	15%	9%

Table 4. Frequency of code occurrence in different mentioning tweets

[1] 13.8% of "directing" tweets overlapped with the "conversing" category, as they happened in the middle of conversation

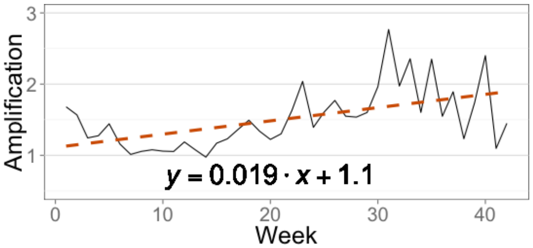

Figure 5. Amplification by week

influence for different opinion groups through the collective use of a design feature. To conceptualize amplification, we adopt a "black box" view and consider the distribution of original messages from different opinion groups as the *input distribution*, and the distribution of the ultimate presence of these messages, resulting from the use of the design feature, to be the *output distribution*. Depending on the design feature studied, the "distribution" could be defined as the total size, rating, weight etc. of the messages from different groups. We define amplification to be the ratio between the output distribution and the input distribution. By quantifying amplification, it allows one to measure the size of the distorting effect of the collective use of a design feature, and also, to monitor the effect over time.

In this section, we study the amplification effect of *retweeting*. In this context, the output distribution would be altered by the visibility of the messages as a result of collectively retweeting (e.g. a single tweet that is retweeted twice has lower visibility than a tweet that is retweeted five times). Using the meta-field of the ID of the retweeted tweet, we were able to track how many times each tweet was retweeted through the retweet button. We calculate the amplification effect with the following formula:

$$Amplification = \frac{N_{pro}/N_{anti}}{n_{pro}/n_{anti}} \quad (2)$$

where n_i is the number of original messages from group i, N_i is the total size after retweeting, calculated by $N_i = \sum_{j=1}^{n_i}(1 + N_{RT}(j))$, where $N_{RT}(j)$ is the number of times a tweet j got retweeted. If the amplification index is equal to 1, it means there is no amplification. If it is above (below) 1, it indicates amplification effect favoring the pro (anti) group.

Figure 5 shows the amplification effect by week. Two conclusions can be drawn: 1) not surprisingly, there was consistent amplification favoring the pro-Snowden (majority) group, as indicated by the amplification index being above 1 almost all time. 2) Interestingly, the amplification effect increased over time (positive coefficient significant at $t(80) = 4.05, p < 0.001$). In the later period, the index often reached more than 2, which means that *the collective use of the retweet button skewed the opinion space towards the pro-Snowden group* by more than twice of the distribution of original messages.

Here we can make some inferences about the causes of the increasing amplification effect. As the anti group shifted to more inter-personal communications, they became more disadvantaged for amplification. Not only because group members might have focused their energy on making original statements, but this type of commentary and conversational tweets are naturally less appealing for retweeting [45]. In contrast, for the pro-Snowden group, the kind of "gatewatching" practice is by nature an "amplifying" practice — by having a concentrated group of people producing original messages (low input), and having a large number of people retweeting these messages (high output).

While there could be many reasons for a group to engage in gatewatching practice, we attribute a key one to the existence of "super influencers" in the pro group, who were the target of "watching". They include well-known persons or organizations that are highly involved in the Snowden issue such as *wikileaks*, *Jesselyn Radack*, *Jacob Appelbaum*, etc., and dedicated activist accounts such as *free_snowden* and *NO2NSA*. In contrast, opinion leaders in the anti group are far less known, with a few political scholars, e.g. *Louise Mensch*, getting the most retweets. On the one hand, these super influencers are highly visible to attract retweeting by having a large number of followers and names that rise issue-specific attention. On the other hand, similar to what was observed in previous studies of activists, politicians and journalists on Twitter [26, 32, 35, 38, 50], we observed that many of them engaged in strategical tweeting behaviors that are more likely to prompt sharing, e.g., creating or citing external resources, calling attention for support, using political and philosophical slogans, framing issues, using humor, etc.. Examples include.:

"Edward Snowden should seek asylum in the only place truly beyond the reach of US law enforcement. Wall Street. #snowden" (wikileaks)

"#Snowden should win the Nobel Peace Prize. Dear @Nobelprize_org, please listen to the Internet We want Snowden! He's a hero! Retweet = √ ."(KimDotcom)

"My artcl: How Obama misled public when he said protected legal channels exist that #Snowden could've used http://t.co/CuGqsDoImU via @Salon"(JesselynRadack)

We quantitatively compare the activities of the top influencers and the group's retweeting of them between the two groups (the method to identify them was discussed in the "user classification" section). We found that, indeed, the top influencers in the pro group were far more likely to be retweeted, with an average retweeting rate of 36.44 per tweet, compared to 4.20 for the top influencers in the anti group. By analyzing the temporal changes of the percentage of tweets contributed by the top influencers (Figure 6 (left)), we found that the contributions of pro-Snowden opinion leaders were steady, and slightly increased over time, in contrast with the decreasing contribution of the anti-group opinion leaders (the difference is significant at $t(80) = 3.28, p = 0.002$). Meanwhile, there was an increasing tendency for the pro group to focus on retweeting the top influencers, as reflected by the increasing proportion of retweets directed to the top influencers (Figure 6(right), the difference is significant at $t(80) = 1.92, p = 0.05$). Together, it suggests that the increasing amplification could be at least partially attributed to increasing gatewatching for opinion leaders in the pro-Snowden group.

To summarize, the collective use of *the retweet button amplified the voices of the majority group*, and the *amplification increased over time*. It indicates that, while the production of original tweets dropped faster for the pro-Snowden group,

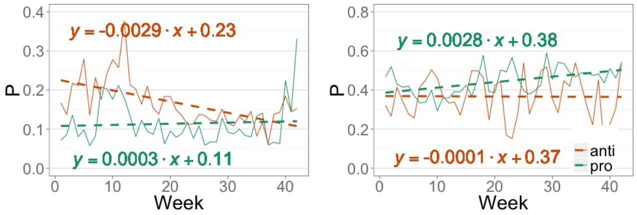

Figure 6. (left) Percentage of tweets from top influencers (right) Percentage of retweets directed to top influencers' tweets

their retweeting activities were better sustained, suggesting that the pro-group members shifted their attention from making original comments to retweeting existing tweets, especially retweeting those from opinion leaders. An important implication is that even though the minority group increased their presence in creating original messages, their ultimate presence on the platform was still weakened. This sends a critically warning that *pooling social media data without disentangling the distortion effect of design features could lead to wrong conclusions about online opinion spaces*.

SUMMARY AND DISCUSSION

We present a case study to demonstrate that behavioral differences between opinion groups may distort online opinion spaces, which is important to consider for studies using social media data. We highlight that such group behavioral differences could be driven by the social environment and platform design. By studying the Twitter discussions around #Snowden, we found that considering only the original messages, the minority group increased its presence in the opinion space over time (RQ1) and engaged in more commentary making (RQ2). Such patterns were driven by their increasing in-group interactions, including having conversations and directing resources to the peers (RQ3). In contrast, the pro-Snowden group engaged in more gatewatching practices — by collectively disseminating messages produced by the group members, especially the group leaders. As a result, the retweeting activities collectively amplified the presence of the majority group, and the amplification increased over time as we found that the majority group further shifted from producing original tweets to retweeting others (RQ4). We highlight the role of the retweet button design, which essentially enabled the amplification.

Majority and Minority Groups in Online Opinion Spaces

While our observational data may be limited in drawing causality, we highlight the consistency between our results and the large literature on majority and minority group behaviors observed in the offline world [42]. The stronger ingroup bias exhibited by the minority group was often explained as "insecure" reactions to perceived environmental threats and heightened needs for strengthening its group identity [29]. In the online opinion space, this insecurity could also translate into minorities' active engagement in expressing and defending their views, as we observed in our study. Moreover, social media creates a unique "resource" perspective—existing supportive information created by in-group members — that fundamentally differs for majority and minority groups. Immersed in a resource rich and salient environment, members of the dominant majority group may experience a decrease in their motivation to create original messages.

Lastly, we want to discuss potential differences regarding entry barriers to join an online opinion space. The increasing presence of opinion minorities (w.r.t. original messages) is not necessarily an rejection of the idea of "spiral of silence". Given its prediction that the entry barrier to start expressing one's opinion would be higher for opinion minorities, it implies that, on average, the minority group who are in the discussions might have higher issue involvement to begin with, thus higher motivation to defend their positions [25]. With lower entry barriers for the majority opinion group, some with lower issue involvement might be willing to join the discussion in the beginning, but drop their engagement over time or shift to less demanding activities such as retweeting.

Implications for Studying Online Opinion Spaces

A main goal of the study is to underline the problem of studying public opinions with "pooled" data from social media without considering the behavioral differences of opinion groups and their evolution. Our findings represent a warning to our research community that the behavioral and motivational differences between opinion group members can lead to differences in expressiveness, thus creating biases in the presence of opinions on social media [24, 39]. Our study highlights that the social environment could systematically drive such differences. By identifying key group features and their impact on individual and group behaviors, we can potentially predict how different opinion groups behave and change, and use that knowledge to inform the development of better methods for studies using social media data which can take the potential biases into account. Our study suggests that we may resort to behavioral and social theories for such purpose, but it is also necessary to conduct empirical studies that can consistently validate these predictions. Research on individual differences in the selectivity of seeking agreeable social opinions online [30, 31] would also be helpful for advancing our understanding and prediction of behavioral differences of opinion groups.

We also highlight a less studied problem, i.e. biases created by the unequal effect of platform design features for different opinion groups. There are two important implications. First, we emphasize that, when analyzing social media data, it is important to consider whether, and how to account for effects enabled by design features, e.g., whether to include retweets, if so, whether to give retweets the same weight as the original tweets. These are practical questions that can lead to different conclusions. Second, we argue that by understanding amplification — the unequal effect of design features—itself will enable us to better understand the opinion space. As our study illustrated, the retweeting button is a key component in the Twitter eco-system. Without studying how retweeting behaviors differ and change, we would have only formed an incomplete understanding of the pro and anti-Snowden groups' activities and evolvement.

Implications for Social Media Design: Towards Assessing Amplification as a Black Box

Recently, the HCI community has studied the biases and inequality created by technologies in order to inform more ethical designs and uses of technologies [28, 22]. We also contribute to this emerging area of interest by highlighting the amplification effect of social media design features for different opinion groups. Besides Twitter, numerous social media platforms provide various design features. It is necessary to study whether, and how they would put an opinion or ideological group at an advantage or disadvantage. In this section, we will discuss an attempt to assess such amplification effects.

We conceptualized the amplification effect by adopting a "black box" view — by seeing a platform as a black box, the original messages produced by opinion groups as the "input", and the outcome through the collective use of a design feature as the "output". An important reason for us to encourage such a view is because it could potentially allow one to probe the amplification effect without opening the black box, e.g., without collecting a complete dataset and conducting full-range analysis. Theoretically, one can assess a black box effect with a small sample of input and observe its output. A very similar idea has been explored in evaluating algorithm biases [21], which uses carefully selected samples to identify what kind of input is more or less likely to have biased results, in order to assess or even reverse engineer algorithm bias.

As a first step towards exploring this idea, we propose a human computation based method to perform a quick assessment of amplification effects based on a small sample, consisting of the following steps:

1. Draw a random sample from the input, i.e., the original messages. Depending on the targeted platform, this can be tweets, forum posts, reviews, etc.
2. Crowdsource labels for the sample to identify the opinion/position of each message, where possible. Depending on the topic, the labels can be dichotomous or have more dimensions.
3. Use the labeled data as the input, and use its eventual presence on the platform as output. Depending on the design feature, the output can be re-defined. For example, on Reddit the output could be the total up/down voting scores, which may determines the ranking results. By comparing the output and input distributions of messages with different opinion labels, one could assess the amplification effect.

This method can give a quick assessment of the amplification effect of a design feature for a specific opinion space. By performing the assessment periodically, one may monitor the amplification effect to potentially identify changes of group dynamics within the platform, and better understand where and when to "open the black box" to further explore the underlying causes.

Here we briefly explore the feasibility of this method with the sample we have at hand. To seed the user classification algorithm, we collected a random sample of 3% daily tweets and crowdsourced their opinion labels, where 23.5% of them were identified. We also have the data on how many times each of them were retweeted. Following the proposed method, we calculate the amplification index with these labeled samples. As shown in Figure 7, consistent with what we found with the complete dataset (see Figure 5), this small sample was able to demonstrate an amplification effect favoring the pro-Snowden group which increased over time.

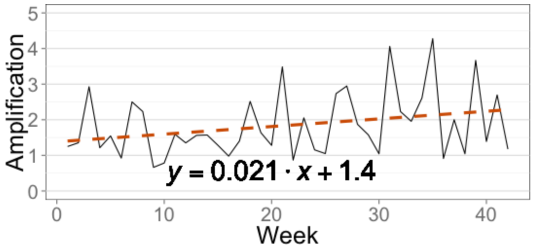

Figure 7. Amplification assessed from sampled tweets

This coarse method is a preliminary proposal and future research is needed for validation and clarifying procedural details. For example, an important question is what sample size would be sufficient. While we demonstrated its feasibility with a 3% sample, it is possible to be done with fewer, especially if the discussion contains higher proportion of opinionated messages. More sophisticated sampling method could also be applied to reduce the required sample size.

Limitations

First and foremost, we acknowledge that our results are observational and have limited explanation power for causality. However, we argue that our focus is to explore whether differences in the *observed* group behaviors may skew an online opinion space. While our results are consistent with theoretical predictions, we refrain from over-generalizing the phenomenon and acknowledge that future research is needed to claim general patterns. We also acknowledge that we did not discuss the external causes, e.g., events and general opinion shift. However, from labeling tweets, we observed little evidence that Twitter users shifted sides, neither was there evident opinion shift reflected in public polls within the time frame. Also, by using 42 weeks of data, our conclusions are likely not impacted by any single event.

Another limitation is the user classification, where we excluded isolated users in the retweeting network. However, we argue that the focus of the paper is to study the behavioral differences of, rather than to conclude, population statistics. Also, the "connected users" are more likely to be serious participants — they are on average 6.2 times more active than the isolated users, so the exclusion of the latter is not likely to change the conclusions. Moreover, we want to point out the general challenges in classifying user positions in the case of highly uneven opinion groups. The sparsity of the minority group's activities resulted in inferior results with content based machine learning method [43], and made other common methods such as methods tapping into retweeting of popular tweets [23] or the use of common hashtags [8] impractical, as the ones for the anti group were in the long tail and thus hard to identify comprehensively.

ACKNOWLEDGEMENT

We would like to thank Arnim Bleier for his help in collecting the dataset used in this study.

REFERENCES

1. 2013. Post-ABC poll: NSA surveillance and Edward Snowden. (24 July 2013). `http://www.washingtonpost.com`.
2. 2013. Public Split over Impact of NSA Leak, But Most Want Snowden Prosecuted. (17 June 2013). `http://www.people-press.org`.

3. 2014. How Twitter Reacted To The Snowden Interview. (19 May 2014). `http://www.nbcnews.com`.

4. 2014. Poll Results: Snowden. (28 March 2014). `https://today.yougov.com`.

5. Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*. ACM, 36–43.

6. Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy* (1992), 992–1026.

7. Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.

8. Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind egyptian political polarization on twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 700–711.

9. Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 1–10.

10. Axel Bruns and Stefan Stieglitz. 2012. Quantitative approaches to comparing communication patterns on Twitter. *Journal of Technology in Human Services* 30, 3-4 (2012), 160–185.

11. Axel Bruns and Stefan Stieglitz. 2013. Towards more systematic Twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology* 16, 2 (2013), 91–108.

12. Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira Jr, and Virgílio Almeida. 2011. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 150–158.

13. Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM* 10, 10-17 (2010), 30.

14. Lu Chen, Wenbo Wang, and Amit P Sheth. 2012. Are twitter users equal in predicting elections? A study of user groups in predicting 2012 US Republican presidential primaries. In *Social informatics*. Springer, 379–392.

15. Christian Christensen. 2013. Wave-riding and hashtag-jumping: Twitter, minority third parties and the 2012 US elections. *Information, Communication & Society* 16, 5 (2013), 646–666.

16. Robert B Cialdini and Melanie R Trost. 1998. Social influence: Social norms, conformity and compliance. (1998).

17. Raviv Cohen and Derek Ruths. 2013. Classifying Political Orientation on Twitter: It's Not Easy!. In *ICWSM*.

18. Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter.. In *ICWSM*.

19. Michael D Conover, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2012. Partisan asymmetries in online political activity. *EPJ Data Science* 1, 1 (2012), 1–19.

20. Munmun De Choudhury, Andres Monroy-Hernandez, and Gloria Mark. 2014. Narco emotions: affect and desensitization in social media during the mexican drug war. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3563–3572.

21. Nicholas Diakopoulos. 2014. Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism* ahead-of-print (2014), 1–18.

22. Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasnt really that close to [her]: Reasoning about invisible algorithms in the news feed. In *Proceedings of the 33rd Annual SIGCHI Conference on Human Factors in Computing Systems*. 153–162.

23. Huiji Gao, Jalal Mahmud, Jilin Chen, Jeffrey Nichols, and Michelle Zhou. 2014. Modeling user attitude toward controversial topics in online social media. In *Eighth International AAAI Conference on Weblogs and Social Media*.

24. Daniel Gayo-Avello. 2013. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review* (2013), 0894439313493979.

25. William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill. 2009. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin* 135, 4 (2009), 555.

26. Avery E Holton and Seth C Lewis. 2011. Journalists, social media, and the use of humor on Twitter. *Electronic Journal of Communication* 21, 1/2 (2011).

27. Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology* 60, 11 (2009), 2169–2188.

28. Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3819–3828.

29. Geoffrey J Leonardelli and Marilynn B Brewer. 2001. Minority and majority discrimination: When and why. *Journal of Experimental Social Psychology* 37, 6 (2001), 468–485.

30. Q Vera Liao and Wai-Tat Fu. 2013. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2359–2368.

31. Q Vera Liao and Wai-Tat Fu. 2014. Can you hear me now?: mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 184–196.

32. Simon Lindgren and Ragnar Lundström. 2011. Pirate culture and hacktivist mobilization: The cultural and social protocols of# WikiLeaks on Twitter. *New Media & Society* 13, 6 (2011), 999–1018.

33. Zhe Liu and Ingmar Weber. 2014. Is twitter a public sphere for online conflicts? A cross-ideological and cross-hierarchical look. In *Social Informatics*. Springer, 336–347.

34. Joseph A Maxwell. 2012. *Qualitative research design: An interactive approach: An interactive approach*. Sage.

35. Eni Mustafaraj, Samantha Finn, Carolyn Whitlock, and Panagiotis Takis Metaxas. 2011. Vocal minority versus silent majority: Discovering the opionions of the long tail. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 103–110.

36. Elisabeth Noelle-Neumann. 1974. The spiral of silence a theory of public opinion. *Journal of communication* 24, 2 (1974), 43–51.

37. Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* 11 (2010), 122–129.

38. Thomas Poell and Erik Borra. 2012. Twitter, YouTube, and Flickr as platforms of alternative journalism: The social media account of the 2010 Toronto G20 protests. *Journalism* 13, 6 (2012), 695–713.

39. Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014), 1063–1064.

40. Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.

41. Dietram A Scheufle and Patricia Moy. 2000. Twenty-five years of the spiral of silence: A conceptual review and empirical outlook. *International journal of public opinion research* 12, 1 (2000), 3–28.

42. Bernd Simon, Birgit Aufderheide, and Claudia Kampmeier. 2001. The Social Psychology of Minority-Majority Relations. *Blackwell handbook of social psychology: Intergroup processes* (2001), 303–323.

43. Laura M Smith, Linhong Zhu, Kristina Lerman, and Zornitsa Kozareva. 2013. The role of social media in the discussion of controversial topics. In *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 236–243.

44. Stefan Stieglitz and Nina Krüger. 2011. Analysis of sentiments in corporate Twitter communication–A case study on an issue of Toyota. *Analysis* 1 (2011), 1–2011.

45. Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*. IEEE, 177–184.

46. Cass R Sunstein. 2002. The law of group polarization. *Journal of political philosophy* 10, 2 (2002), 175–195.

47. Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 290–297.

48. Michael J Welch, Uri Schonfeld, Dan He, and Junghoo Cho. 2011. Topical semantics of twitter links. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 327–336.

49. Sarita Yardi and Danah Boyd. 2010. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society* 30, 5 (2010), 316–327.

50. William Lafi Youmans and Jillian C York. 2012. Social media and the activist toolkit: User agreements, corporate interests, and the information infrastructure of modern social movements. *Journal of Communication* 62, 2 (2012), 315–329.