JOURNAL OF COMMUNICATION

Journal of Communication ISSN 0021-9916

ORIGINAL ARTICLE

Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data

Elanor Colleoni[1], Alessandro Rozza[2], & Adam Arvidsson[1]

1 Department of Social and Political Sciences and Center for Digital Ethnography, University of Milan, Milan, Italy
2 Department of Science and Technology, University of Naples-Parthenope, Naples, Italy

This paper investigates political homophily on Twitter. Using a combination of machine learning and social network analysis we classify users as Democrats or as Republicans based on the political content shared. We then investigate political homophily both in the network of reciprocated and nonreciprocated ties. We find that structures of political homophily differ strongly between Democrats and Republicans. In general, Democrats exhibit higher levels of political homophily. But Republicans who follow official Republican accounts exhibit higher levels of homophily than Democrats. In addition, levels of homophily are higher in the network of reciprocated followers than in the nonreciprocated network. We suggest that research on political homophily on the Internet should take the political culture and practices of users seriously.

doi:10.1111/jcom.12084

In this paper, we analyze the structure of political homophily in Twitter. While there is a growing attention to this question in the literature, previous studies have focused on single cases. We use a Big Data approach that combines machine learning-based analysis of textual content with social network analysis to explore the complete network of 2009 Twitter users. This systematic approach allows us to make more generalizable statements about the nature of Twitter as a medium of political communication.

The question of political homophily is important because it concerns the ability of digital media to support the formation of a public sphere, where a diversity of opinion and information can interact, or, conversely, to function as an echo chamber that reinforces established perspectives and opinions. Both scenarios are well established in the research on political communication on the Internet.

Within Internet studies the public sphere scenario has been most prominently defined by Peter Dahlgren. Following Habermas' (1962/1989) work, Dahlgren (2005) defines a public sphere as "a constellation of communicative spaces in society that

Corresponding author: Elanor Colleoni; e-mail: elanor@inventati.org

permit the circulation of information, ideas, debates, ideally in an unfettered manner, and also the formation of political will" (p. 148). A public sphere should allow public dialog and reasoning through the advancement of claims and information that lead to deliberation. In line with these suggestions, research has focused on the potential of the Internet to reinvigorate the public sphere (Holt, 2004), suggesting that the Internet has increased citizens' exposure to political discussion and confrontation. In particular, Brundidge (2010) has found evidence that through inadvertent exposure, Internet use contributes to increase the heterogeneity of political discussion networks. According to Brundidge (2010), this increased exposure is due to the fact that the Internet facilitates people's inadvertent exposure to political difference, even if they are unlikely to seek out such political difference on their own. This exposure is facilitated through "(a) less than perfect online selective exposure strategies, (b) nonavoidance of encounters with political difference, and (c) weakened social boundaries between far flung geographic locations, between one discursive space and the next (blurred and porous boundaries creating increased interspatiality), between political and apolitical spaces of communication, and between the private and the public spheres" (p. 687). This is coherent with Wojcieszak and Mutz (2009)'s findings, which support the idea that exposure to heterogeneous networks and political views often happens accidentally and in places not directly devoted to political discussion, but where political and nonpolitical discussions co-occur. In other words, the tendency to support a public sphere where diverse opinions and information interact can be understood as part of the technical bias of the Internet as a medium of communication.

Other authors contest this public sphere scenario, suggesting that instead of reinforcing democratic discussion, the Internet reinforces prior political views due to selective exposure to political content (Bimber & Davis, 2003; Davis, 1999; Galston, 2003; Mutz & Martin, 2001; Noveck, 2000; Sunstein, 2001; Wilhelm, 1998). In this case, the Internet functions as an echo chamber where political orientation is reaffirmed (Sunstein, 2001). Recent research supports the idea that people tend to search for and evaluate opinions of others that are not too divergent from themselves (Kushin & Kitchener, 2009; Stroud, 2010). Bimber and Davis (2003) have analyzed the political attitude of visitors to the Gore and Bush websites in 2000 and found that they shared the political outlook advanced by the website they visited. They argued that the opportunity given by new media to select information and interactions has resulted in users' tendency to prefer partisan information. Mutz and Martin (2001) achieved similar results. They found that people tend to prefer information that reinforces partisan sources over that which includes different voices. Focusing on weblogs, Adamic and Glance (2005) also found evidence of balkanization and self-segregation, with political blogs primarily linking to those sharing the same political ideology.

The mechanism through which this fragmentation of political discourse operates is homophily, defined as the tendency of similar individuals to form ties with each other (for a review see McPherson, Smith-Lovin, & Cook, 2001). The most commonly cited explanations for this phenomenon are cognitive dissonance and selective exposure theories (Festinger, 1954). According to these theories, people experience

positive feelings when presented with information that confirms their opinions. When faced with divergent opinions, people tend to feel stressed and pressured to conform. Consequently, they will be more inclined to expose themselves to agreement and to information and discussion that reinforce their original view. It follows that the selective exposure process is due to a tendency of individuals to reduce their cognitive dissonance and, as a consequence, to create homogeneous groups, affiliating with individuals that are similar in certain attributes, such as beliefs, education, and social status (Lazarsfeld & Merton, 1954).

Homophily results in shared views of the world. When applied to the political domain, homophily produces shared political attitudes which can result in political polarization. "Homophily limits people's social worlds in a way that has powerful implications for the information they receive, the attitudes they form, and the interactions they experience" (McPherson et al., 2001, p. 23). Following this argument, the "echo chamber" effect is due to a tendency of individuals to create homogeneous groups and to affiliate with individuals that share their political view. This is not a trivial problem. As Scheufele, Hardy, Brossard, Waismel-Manor, and Nisbet (2006) have pointed out, the greater the network heterogeneity in which individuals are embedded, the bigger their desire for information on different topics. Furthermore, politically diverse personal networks increase awareness of oppositional viewpoints and political tolerance (Mutz, 1999). In contrast, the sole exposure to like-minded people seems to be associated with the adoption of more extreme positions (Sunstein, 2001; Mutz & Martin, 2001) leading to political polarization (Stroud, 2010).

The recent rise of social networking sites (SNSs; Boyd & Ellison, 2007) has given new relevance to the question of political homophily on the Internet. This is because SNSs like Facebook and Twitter enable high levels of interactivity and allow for diffused and real-time discussions with no geographical constraints (Rafaeli & Sudweeks, 1997). These factors, combined with the potential of SNSs to achieve high levels of diffusion, or "virality" of political (and other) content, mean that they have established themselves as important vehicles for political communication (Honeycutt & Herring, 2009; Williams & Gulati, 2009). At the same time, SNSs tend to foster both the public sphere scenario with low levels of homophily and the echo chamber scenario where homophily is high, as they tend to reinforce group cohesion as well as information diffusion (Boyd & Ellison, 2007; Kwak, Lee, Park, & Moon, 2010).

Twitter has achieved particular relevance as a medium of political communication. This is because posts are visible to every user by default (unless differently specified); content is easily sharable and can quickly spread in the network by using the retweet function; the system of hashtags and mentions allows the creation of publics around specific discussions without the need for group creation, and users can follow a particular account without asking the permission of its owner. This means that Twitter not only allows for the relations of reciprocated ties that are typical of Facebook; Twitter is also for following people who do not reciprocate one's following. Twitter is both a "social" and a "newsy" medium; to use Kwak et al.'s (2010) terms: It allows for the formation both of "symmetric social graphs" based on symmetric relationships and

of "nonsymmetric interest graphs" based on nonsymmetric relationships (Ravikant & Rifkin, 2010). This means that Twitter should theoretically be conducive to *both* the public sphere *and* the echo chamber scenario.

Indeed, research on homophily on Twitter so far has reached contrasting results. In their seminal work, Kwak et al. (2010) have investigated homophily by analyzing users' geographic location and popularity and found that users who have reciprocal relations of fewer than 2000 are likely to be geographically close. In general, however, they found low levels of homophily. By focusing instead on value homophily, Weng, Lim, Jiang, and He (2010) arrived at opposite conclusions. Using a topic model, they define homophily as the tendency to share similar content and therefore to exhibit similar interests. Based on a corpus of 1 million tweets from Singapore extracted between April 2008 and April 2009, they found evidence that friends on Twitter tend to share interests. Conover et al. (2011) made a first attempt to link homophily and political orientation. Based on a sample of 1,000 users they found evidence that political networks on Twitter are highly segregated, as users tend to retweet more from those users sharing the same political affiliation. Feller, Kuhnert, Sprenger, and Welpe (2011) achieved similar results. Using Adamic and Glance's (2005) approach, they looked at the conversations surrounding German political parties during the 2009 federal elections and found that political tweeters tend to be segregated according to shared political affiliation. Boutet, Kim, and Yoneki (2012) have also investigated users' political affiliation based on the mention/ retweet behavior and the segregation/contamination of retweets on Twitter during the 2010 U.K. general election. They found the graph to present a highly segregated partisan structure, and that party members were more likely to retweet material from their own party than material derived from other parties.

These studies are all based on restricted samples or singular events (like the 2010 U.K. elections). So far nobody has looked at how structures of political homophily play out on Twitter in general, by analyzing the entire network of users. In this paper, we set out to perform such a systematic analysis by comparing levels of political homophily among U.S. Republicans and Democrats based on the entire network of 2009 Twitter users.

We chose to work with the 2009 network because it is an opportunity to work on the whole graph, which would be otherwise unfeasible to collect because of the prohibitive costs associated with the necessary computational and bandwidth resources. Working with the whole network allows us to perform an unbiased analysis of the Twitter network structure.

We chose to work with U.S. Democrats and Republicans because these political identities are both sufficiently popular and sufficiently distinct to allow for machine learning-based classification. Independents, while a larger political group than both Democrats and Republicans in the United States, are too indistinct as political identity to be successfully detected by an automatic classifier.

In addition, none of the previous studies have taken the dual nature of Twitter — as a social medium based on symmetric ties and as a newsy medium based on

nonsymmetric ties—seriously. Given that this distinction resembles two distinct homophilic mechanisms, namely, choice-produced homophily and induced homophily (McPherson et al., 2001), we believe that it might have implications for the structure of political homophily on Twitter, which so far have not been analyzed.[1]

In order to achieve this, we rely on machine learning techniques and social network analysis. We use machine learning to classify users as Democrat or as Republican supporters based on the content shared in Twitter. We then use network analysis to measure levels of homophily both in the nonsymmetric interest graph—where users follow other users without being themselves followed "back" and in the "symmetric social graph," where users follow other users that in turn follow them.

Methods

Machine learning is concerned with the development of algorithms that allow recognizing and extracting patterns from data and making intelligent decisions based on empirical data. In the last decades, machine learning has been widely applied in solving complex tasks in various areas of research, such as speech recognition, computer vision, and text mining. The strength of machine learning relies on its capability to automatically improve performance through experience. In our study, we make use of supervised classification to automatically detect political content in tweets and to discriminate Democratic and Republican political discourses.

A supervised learning algorithm uses a training set to infer a mathematical model that can be used for mapping new data. The training data consists of a corpus of text that has already been labeled according to the characteristic under investigation, in our case as political/nonpolitical and Democrat/Republican. The training set is then divided in two parts. The first part is used to "train" the classifying algorithm. The second part is used to test the algorithm by measuring its ability to correctly classify unseen labeled examples. Accuracy levels range from 0 to 1.

By classifying all the content posted according to its political orientation we are able to identify the general political orientation of the users and measure levels of political homophily in their network.

This section is organized as follows: In the first paragraph we describe the data employed and how it is linked together in order to create a suitable dataset for the analysis; in the second paragraph we explain the method used to predict the user political orientation; in the last paragraph we present the measure employed to compute political homophily.

Data

Twitter graph

Twitter graph was created by Kwak et al. (2010) and consists of all nodes and ties on Twitter in 2009. The nodes represent the users whilst the ties represent the relationships between them. The database contains over 40 million nodes and 1.47 billion ties.

Twitter content

Twitter content was created by Yang and Leskovec (2011) and consists of a representative sample of 467 million tweets from 20 million users covering a 7-month period from June 1 to December 31, 2009.

Political training set

The political training data (referred in the following as *PolTrainingSet*) is a set of online news feeds with a known mixture of political and nonpolitical news extracted from news blogs. The corpus consists of 59,757 political and 166,337 nonpolitical titles (from January 2008 to May 2013).

Democrat and Republican training set

This training set (hereafter *DemRepTrainingSet*) is based on the Twitter Content database. It is obtained by scraping all the political tweets of the users that follow Democrat or Republican accounts, assuming that they share the same political attitude of the accounts they follow. We do not consider users who are following both Democrat and Republican accounts. This way we identify 1,683 Democrat users and 8,868 Republican users (obtaining respectively 28,167 and 189,933 tweets). This result is consistent with the figures published by Democrat and Republican officials at the beginning of 2010. According to Democratic officials, 108 House Democrats (of a total of 255) have Twitter accounts. On the Republican side, officials confirm that there are 130 House Republicans (out of a total 178) on Twitter. According to TweetCongress, a website that monitors congress online activity, Republicans have more followers, they are more active, and they are more in sync with each other (Chittal, 2010).

Testing data

Merging the *Twitter Content* and *Twitter Graph* databases we obtain the entire 2009 ego-network of followers and followees for each user, as well as a representative sample of tweets. In the following we refer to this dataset as *TestingData*.

Predicting political orientation

Twitter users' communication streams and social network structures have been successfully used to detect user attributes like gender, ethnicity, and sexual orientation (Pennacchiotti & Popescu, 2011a). Applying machine learning techniques, Pennacchiotti and Popescu (2011a) used tweeting behavior, network structure, and the linguistic content of tweets to predict the political orientation and ethnicity of users. They showed that considering content shared by users and the information on the overall political orientation of their networks (i.e., the political orientation of content shared within the network) strongly increases the performance in predicting political orientation.

Following Pennacchiotti and Popescu's suggestion, we exploit the linguistic dimension of the tweets. In order to predict the political orientation of users from the

Table 1 Tenfold Cross-Validation on the *PolTrainingSet* Using Passive–Aggressive

Accuracy	0.95897209076
Precision	0.934082795884
Recall	0.908904218262
F-measure	0.921312546649

content shared, we first need to distinguish between political and nonpolitical tweets. To do so, we extract representative features from the *PolTrainingSet* by employing {1,2,3,4,5}-grams of words (where *n*-gram is a contiguous sequence of *n* items from a given sequence of text, in this case words) and using term frequency-inverse document frequency (Manning, Raghavan, & Schütze, 2008). This algorithm computes the frequency of a term and divides it by the inverse document frequency, a measure of how the term is common across all documents. This measure is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

The representative features are then used to train a classification model. We choose to utilize a Passive–Aggressive classification algorithm, an effective classifier able to scale on a large corpus and to be updated over time (Crammer, Dekel, Keshet, Shalev-Shwartz, & Singer, 2006).

As shown in Table 1, the quality of the feature extraction and classification model is confirmed by the experimental results obtained through 10-fold cross-validation on the *PolTrainingSet*. These results are consistent with those achieved in related works (Monti et al., 2013). The 10-fold cross-validation consists of randomly partitioning the original sample into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model while the remaining subsamples are used as training data. The final result is achieved by performing all the possible permutations and averaging the measures computed on the validation sets.

Overall, we found that around 10% of the discourse is related to political issues. This result is consistent with Goel, Mason, and Watts (2010), who have found similar results analyzing Facebook users.

To separate Democratic tweets from the Republican, we apply a similar approach. We train a passive–aggressive algorithm on the features extracted by using the words contained in the *DemRepTrainingSet* and compute the term frequency-inverse document frequency.

As Table 2 shows, we reached an accuracy of 79% on the *DemRepTrainingSet* in 10-fold cross-validation. This result is consistent with the results achieved in related works (Pennacchiotti & Popescu, 2011b).

Considering the levels of accuracy, we are confident to apply a "chain" of classifiers, where the political tweets detected on the *TestingData* by the first classifier (trained on *PolTrainingSet*) represent the input for the DemRep classifier (trained on *DemRepTrainingSet*). The last step is then to evaluate the overall political orientation of a user by simply counting the number of tweets classified as Democratic and as

Table 2 10-fold Cross-Validation on the *DemRepTrainingSet* Using Passive–Aggressive

Accuracy	0.794763003728
Precision	0.781011371604
Recall	0.819281955639
F-measure	0.799638828396

Republican discourse normalized by the total number of political tweets per each user, and assigning the label according to the orientation of the majority of the political tweets posted by the user.

Measuring political homophily

Once all the users had been classified according to political orientation, we selected only those users who shared at least one political tweet by extracting only nodes that are Republican and Democrat. For each node, we select all those nodes connected by an outbound tie and classify them as Republicans, Democrats, or nonpolitical accounts. Based on this network (hereafter general graph), we compute the general level of political homophily, defined as the number of outbound ties directed to users who share political orientation, divided by the overall number of outbound ties (i.e., directed to users with similar political orientation plus directed to users with different political orientation). The homophily rate ranges from 0 to 1.

After having measured the homophily rate in the general graph, we measure the homophily rate of the nonsymmetric interest graph by extracting only those pairs of nodes that are connected by a tie that does not have a reversed arc (i.e., users who are followed by other users who do not follow them back). We measure the homophily rate of the symmetric social graph by extracting only those pairs of nodes connected by a tie that has a reversed arc (i.e., users who are followed by other users who, in turn, follow them back).

In order to assess the statistical significance of the homophily rates in the three graphs (general, nonsymmetric interest, and symmetric social graph), we compare the rates with a baseline homophily (Wasserman & Faust, 1994). We define the baseline homophily rate as the homophily that would be expected randomly in a graph with characteristics similar to the one under investigation. To do this, we generate another graph with the same structure as the graph under investigation and we randomly label the nodes according to the distribution estimated on the nodes of the original graphs. For instance, in the general Twitter graph, the Democrats' outbound ties are about 24 million, while the Republicans' are around 4 million. This means that a user has six times higher probability of following a Democrat, regardless of his/her desire to affiliate with politically similar others. By repeating the experiment 100 times, we can compute the probability distribution of random homophily rates, mean, and standard deviation. To estimate if the levels of homophily in our graphs are significantly different from the distribution of the random generated graphs, we simply computed

Table 3 Number of Users Following a Political Account and Number of Users Classified

	No. of Users Following a Political Account	No. of Users Classified
Node: Republican	8,868	72,302
Node: Democrat	1,683	782,371

two-tail *z*-tests (Sprinthall, 2003). The goal is to assess whether the average homophily rates in our graphs are statistically different from the random generated graphs and therefore significantly different than would be expected by chance.

Once we have established the significance of the homophily rates in our graphs, we can directly compare the normalized homophily rates between the nonsymmetric interest and symmetric social graph, in order to assess which one exhibits higher levels of homophily.

Results

The goal of our analysis was to assess whether Twitter is enhancing discussion among users with different political views, or if its nature increases the exposure to like-minded people.

Our first result concerns user classification. Overall, we identified 72,302 Republicans and 782,371 Democrats. As can be seen in Table 3, while the number of users who follow the Republican official account is eight times higher than those who follow Democrat official accounts, the number of users who are identified as Democrats by the content shared rather than by their official affiliation is 10 times higher than the users classified as Republicans in this way. This indicates a different nature of political participation among Democrats and Republicans. Democrats are less likely to follow the official Democrat account, but the Democratic discourse is 10 times more present in the general discourse, suggesting that Democrats are more likely to express their ideas in the flow of the discussion. This result seems coherent with Wojcieszak and Mutz (2009), who found that political discussion does not primarily occur in political spaces, but in other networks where the political discussion comes up incidentally. They found that deliberation and political exposure to cross-cutting political views in chat rooms and message boards occurs primarily incidentally by talking about political topics or controversial public issues. On the other hand, Republicans are more likely to follow the official Republican account, but they are less likely to express their ideas in the flow of the discussion.

We then proceeded to measure the homophily rate in the general graph and compare it to a random graph. Given the properties of the original graph, the average random expected level of homophily was 0.77. The level of homophily in our graph is 0.80. By computing the *z*-test, we are able to say that the rate of political homophily observed in our graph is significantly higher than the rate expected by chance. However, as can be seen in Table 4, when looking at the distribution of outbound ties

Table 4 Outbound Ties Distribution Given the Node Political Orientation

Type of Graph	General		Nonsymmetric Interest		Symmetric Social	
Out Bound Ties	Republican	Democrat	Republican	Democrat	Republican	Democrat
Node: Republican	23.90%	76.09%	19.91%	80.08%	25.93%	74.06%
Node: Democrat	11.62%	88.37%	10.66%	89.33%	12.38%	87.61%

by political orientation, two strongly unbalanced and different distributions can be observed.

On average, Democrats create outbound ties in 88% of the cases with Democrats and in 12% of the cases with Republicans. On average, Republicans create outbound ties in 76% of the cases with Democrats and 24% of the cases with Republicans.

This result seems to indicate low levels of political homophily for the Republicans and high levels for the Democrats. Therefore, we decided to run additional tests for Democrats and Republicans, respectively. First, we computed the homophily rates and significance tests for those users classified as Democrats and as Republicans separately. We ran the experiment for the three graphs, that is, general, nonsymmetric interest graph, and symmetric social graph, and found that in all the three cases the homophily rate is significantly different from what would be expected by chance. Table 5 shows the observed and expected (in brackets) levels of homophily for Democrats and Republicans in the three graphs respectively.

This additional investigation has shown that Democrats have a significantly higher political homophily rate than expected by chance, while Republicans have a significantly lower political homophily rate than expected by chance. The result is consistent with the findings of Pennacchiotti and Popescu (2011b), who also classified user political orientation according to the content shared online. They found evidence that "Democrats tend to consistently have a large percentage of friends with the same affiliation. For Republicans, the political affiliation of the neighbors is more mixed (e.g., Republican Twitter users tend to have friends—and followers—with both probable Republican and Democrat affiliations)" (p. 435). Yet, the result seems inconsistent with other findings and with common sense. For instance, Boutyline and Willer (2013) found "consistent and robust evidence that conservatives are more homophilous than liberals" (p. 31). Messing and Westwood (2012) also found evidence that "the effect of social endorsements was strongest for partisans selecting

Table 5 Political Homophily Rates by Political Orientation and Type of Graph (Homophily ranges from 0 to 1)

Type of Graph	General	Nonsymmetric Interest	Symmetric Social
Democrats	0.88 (0.79)	0.89 (0.80)	0.87 (0.78)
Republicans	0.23 (0.63)	0.19 (0.68)	0.25 (0.61)

Table 6 Political Homophily Rate by Political Orientation and Type of Graph for Users Following Official Accounts (Homophily ranges from 0 to 1)

Type of graph	General	Nonsymmetric Interest	Symmetric Social
Democrats	0.44 (0.51)	0.41 (0.50)	0.53 (0.50)
Republicans	0.94 (0.91)	0.94 (0.91)	0.96 (0.93)

articles from ideologically misaligned sources, and stronger for Republicans than for Democrats" (p. 15). However, these two studies share a common methodological feature: The political orientation of the users is not inferred from the content shared. In the former study, Republicans and Democrats are classified accordingly to the Twitter official accounts they follow; in the latter users self-reported their political orientation.

Therefore, we decided to run a second set of experiments considering only Democrats and Republicans who follow Twitter official party accounts. Table 6 shows the observed and expected (in brackets) levels of political homophily of Democrats and Republicans who follow official party accounts in the three graphs respectively. By considering only those users who follow official party accounts, the results are reversed and consistent with previous findings. Democrats have significantly lower political homophily rates than expected by chance, while Republicans have significantly higher political homophily rates than expected by chance.

To summarize, if we focus on the political discourse (i.e., users classified according to the political content shared on Twitter), Democrat "thinkers" (i.e., users classified as Democrats who do not follow official Democrat accounts) privilege to a great extent to associate with other Democrat "thinkers." On the contrary, Republican "thinkers" associate to a very limited extent with other Republican "thinkers" (i.e., significantly below the random levels of political homophily). If we focus on formal affiliation (i.e., users following official party accounts), Republican "activists" (i.e., Republican users who follow official Republican accounts) privilege to a great extent to associate with other Republicans. On the contrary, Democrat "activists" associate to a very limited extent with other Democrats.

Are levels of homophily different in Twitter's symmetric "social graph" and in its nonsymmetric "interest graph"? In order to address this question, we have divided the normalized level of homophily of the nonsymmetric interest graph by the level of the symmetric social graph. We find that the symmetric social graph exhibits 16.1% higher levels of political homophily. This outcome confirms expectations from previous research (i.e., Baldassari & Bearman, 2006; Huber & Malhotra, 2013) as well as common sense expectations that reciprocal social relationships are more similar than nonreciprocal expressions of interest. This suggests that the double nature of Twitter as a "newsy" and a "social" medium implies that the platform is conducive to two distinct modalities of political communication: a more communitarian echo chamber-like model based on reciprocal ties, similar to Weng et al.'s (2010) results,

and a more open-ended public sphere-like mode of participation similar to what Kwak et al. (2010) have found.

Discussion

This article contributes to the discussion of nature of political participation on the Internet by empirically exploring the whole corpus of Twitter traffic from 2009. We have investigated whether Twitter is conducive to a public sphere-like scenario where users are exposed to diverse content or whether the platform is conducive to an echo chamber-like scenario where established partisan positions tend to be reinforced. The answer is that this depends.

It depends on how we analyze Twitter. If we look at Twitter as a social medium we see higher levels of homophily and a more echo chamber-like structure of communication. But if we instead focus on Twitter as a news medium, looking at information diffusion regardless of social ties, we see lower levels of homophily and a more public sphere-like scenario. This confirms results from political sociology that suggest higher levels of homogeneity in interpersonal networks (Baldassari & Bearman, 2006), as well as common sense expectations.

But results also depend on factors that are quite unrelated to the technical affordances of Twitter as a communication platform. If we had focused on Democrats alone we would have seen higher levels of homophily about among the democratic public, but lower levels of homophily among activists (defined as followers of official accounts). Had we instead focused on Republicans alone, the results would have been inverted. Without venturing too far into political sociology we might suggest that a possible explanation for these results lies with differences in the structure of Democratic and Republican political culture. While there is growing evidence of increasing political polarization in the United States (Bishop, 2009; Fiorina, Abrams, & Pope, 2005; Iyengar, Sood, & Lelkes, 2012), this polarization is more pronounced at the level of political activists and cadres than at the level of the general public (Fischer & Mattson, 2009; Heaney, Masket, Miller, & Strolovitch, 2012). Because a notably higher percentage of Republican tweeters are also followers of official Republican accounts (12.2 vs. 0.2% for Democrats), Republican political participation seems to be much more strongly organized and Republican tweeters much more likely to exhibit a more "activist" like mode of participation. Indeed levels of homophily among Republican followers of official accounts are much higher than among Democrats (0.94 vs. 0.44 for the general graph, and similar results for the other graphs), suggesting that these adhere more strongly to the "activist" mode of participation and are integrated in a more tightly organized "political machine." Higher levels of homophily among the Democratic public, on the other hand, might be explained by the identity-focused nature of democratic political discourse. "Liberals" have been identified as the "second largest" subculture in the United States with strong internal value consistency (Ray & Anderson, 2000). This means that people who identify as Democrats, without being actively involved in politics, might select to follow and associate with people

who confirm their views to a higher extent than Republicans. At the same time, a larger proportion of Republicans are more actively involved in politics (they follow official accounts) and when they are, they exhibit significantly stronger levels of partisan integration.

This is a tentative scenario that must remain open for further research. Our point here is that a systematic Big Data analysis such as ours reveals that the widely different forms of political participation that have been revealed by previous case study based research might very well coexist on Twitter. This implies that the question if Twitter, or any other SNS for that matter, is by itself conducive to this or that form of political (or other) participation might be wrongly put. We would suggest a different perspective where the focus of analysis moves away from the single medium or platform itself (whether this be "Internet," "Twitter," or something else) to looking at how affordances and features of such platforms interact with the culture and practices of users. Such a turn away from "virtual" to "digital" methods (Marres & Lezaun, 2011; Rogers, 2012) — away from treating the Internet or SNS as a separate reality and towards a focus on the Internet as one among many aspects of social reality in general — might open up interesting and fruitful avenues for Big Data Analysis.

Note

1 That is, to the extent that "some observed prevalence of homophilous ties can be attributed to individual, psychological preferences, it should be called choice homophily, and to the extent that it can be shown to arise as a consequence of the homogeneity of structural opportunities for interaction, as in […] friendship circles, it should be labeled induced homophily" (Kossinets & Watts, 2009, p. 407).

References

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. *Proceedings of the 3rd International Workshop on Link Discovery*. New York, NY: ACM.

Baldassari, D., & Bearman, P. (2006). *Dynamics of political participation.* ISERP Working Paper 06–07. New York, NY: Columbia University.

Bimber, B., & Davis, R. (2003). *Campaigning online: The Internet in U.S. elections*. Oxford: Oxford University Press.

Bishop, B. (2009). *The big sort: Why the clustering of like-minded America is tearing us apart*. Boston, MA: Houghton Mifflin.

Boutet, A., Kim, H., & Yoneki, E. (2012). What's in Twitter: I know what parties are popular and who you are supporting now! *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*, pp.132–129. New York, NY: ACM.

Boutyline A., & Willer, R. (2013). *The social structure of political echo chambers: Ideology and political homophily in online communication networks*. Working paper, University of Berkeley.

Boyd, D., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, **13**(1), 210–230. doi:10.1111/j.1083-6101.2007.00393.x.

Brundidge, J. (2010). Encountering "Difference" in the contemporary public sphere: The contribution of the Internet to the heterogeneity of political discussion networks. *Journal of Communication*, **60**, 680–700. doi:10.1111/j.1460-2466.2010.01509.x.

Chittal, N. (2010). Twitter reality: The Republicans are crushing the Democrats when it comes to tweeting. Retrieved from http://www.alternet.org/story/147822/twitter_reality%3A_the_republicans_are_crushing_the_democrats_when_it_comes_to_tweeting

Conover, M.D., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on Twitter. *Fifth International AAAI Conference on Weblogs and Social Media.*

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, S. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, **7**, 551–585.

Dahlgren, P. (2005). The Internet, public spheres, and political communication: Dispersion and deliberation. *Political Communication*, **22**(2), 147–162. doi:10.1080/10584600590933160.

Davis, R. (1999). *The web of politics: The Internet's impact on the American political system*. New York, NY: Oxford University Press.

Feller, A., Kuhnert, M., Sprenger T.O., & Welpe, I. (2011). Divided they tweet: The network structure of political microbloggers and discussion topics. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 11).*

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, **7**(2), 117–140.

Fiorina, M., Abrams, S., & Pope, J. (2005). *Culture war?* New York, NY: Pearson Longman.

Fischer, C., & Mattson, G. (2009). Is America fragmenting? *Annual Review of Sociology*, **35**, 435–455. doi:10.1146/annurev-soc-070308-115909.

Galston, W. A. (2003). If political fragmentation is the problem, is the Internet the solution? In D. M. Anderson & M. Cornfield (Eds.), *The civic web: Online politics and Democratic values* (pp. 35–44). Lanham, MD: Rowman & Littlefield.

Goel, S., Mason, W., & Watts, D. J. (2010). Real and perceived attitude agreement in networks. *Journal of Personality and Social Psychology*, **99**(4), 611–621. doi:10.1037/a0020697.

Habermas, J. (1962/1989). *The structural transformation of the public sphere*. Cambridge, MA: MIT Press.

Heaney, M., Masket, S., Miller, J., & Strolovitch, D. (2012). Polarized networks: The organizational affiliations of national party convention delegates. *American Behavioral Scientist*, **56**(12), 1654–1676. doi:10.1177/0002764212463354.

Holt, R. (2004). *Dialogue on the Internet: Language, civic identity, and computer–mediated communication*. Westport, CT: Praeger.

Honeycutt, C., & Herring, S. (2009). *Beyond microblogging: Conversation and collaboration via Twitter*. *Proceedings of the 42nd Hawaii International Conference on System Sciences HICSS '09*. Los Alamitos, CA: IEEE Press.

Huber, G., & Malhotra, N. (2013). *Dimensions of political homophily: Isolating choice homophily along political characteristics*. American Political Science Association annual meeting, New Orleans, LA.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, **76**(3), 405–431. doi:10.1093/poq/nfs038.

Kossinets, G., & Watts, D. (2009). Origins of homophily in evolving networks. *American Journal of Sociology*, **115**(2), 405–450.

Kushin, M., & Kitchener, K. (2009). Getting political on social network sites: Exploring online political discourse on Facebook. *First Monday*, **14**(11), 1–16.

Kwak, H., Lee, C., Park, H., & Moon, S., (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World wide web*, April 26–30, 2010, Raleigh, NC. 10.1145/1772690.1772751.

Lazarsfeld, P. F., & Merton, R. (1954). Friendship as a social process: A substantive and methodological analysis. In B. Morroe, T. Abel, & C. Page (Eds.), *Freedom and control in modern society* (pp. 18–66). New York, NY: Van Nostrand.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. London: Cambridge University Press.

Marres, N., & Lezaun, J. (2011). Materials and devices of the public: An introduction. *Economy and Society*, **40**(4), 489–509. doi:10.1080/03085147.2011.602293.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, **27**, 415–444.

Messing, S., & Westwood, S. (2012). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research*, 1–22. doi:10.1177/0093650212466406(pre-printed online).

Monti, C., Rozza, A., Zappella, G., Zignani, M., Arvidsson, A. & Colleoni, E., (2013). Modelling political disaffection in Twitter. *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM 2013)*, August 8–11, 2013, Chicago, IL.

Mutz, D. C., & Martin, P. S. (2001). Facilitating communication across lines of political difference: The role of mass media. *American Political Science Review*, **95**(1), 97–114.

Noveck, B. S. (2000). Paradoxical partners: Electronic communication and electronic democracy. *Democratization*, **7**(1), 18–35. doi:10.1080/13510340008403643.

Pennacchiotti, M., & Popescu, A. (2011a). A machine learning approach to Twitter user classification. *Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM 2011)*.

Pennacchiotti, M., & Popescu, A. (2011b). Democrats, Republicans and Starbucks afficionados: User classification in Twitter. *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (KDD-WISDOM)*, August 21–24, 2011, San Diego, CA.

Rafaeli, S., & Sudweeks, F. (1997). Networked interactivity. *Journal of Computer-Mediated Communication*, **2**(4). doi:10.1111/j.1083-6101.1997.tb00201.x.

Ravikant, N., & Rifkin, A. (2010). *Why Twitter is massively undervalued compared to Facebook*. Retrieved from http://techcrunch.com/2010/10/16/why-twitter-ismassively-undervalued-compared-to-facebook/

Ray, P., & Anderson, R. (2000). *The cultural creatives*. New York, NY: Three Rivers Press.

Rogers, R. (2012). *Digital methods*. Cambridge, MA: MIT Press.

Scheufele, D. A., Hardy, B. W., Brossard, D., Waismel-Manor, I. S., & Nisbet, E. (2006). Democracy based on difference: Examining the links between structural heterogeneity, heterogeneity of discussion networks, and democratic citizenship. *Journal of Communication*, **56**, 728–753. doi:10.1111/j.1460-2466.2006.00317.x.

Sprinthall, R. C. (2003). *Basic statistical analysis* (*Seventh edition* ed.). Boston, MA: Pearson Education Group.

Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication*, **60**(3), 556–576. doi:10.1111/j.1460-2466.2010.01497.x.

Sunstein, C. R. (2001). *Republic.com*. Princeton, NJ: Princeton University Press.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, England: Cambridge University Press.

Weng, J., Lim, E., Jiang, J., & He, Q. (2010). TwitterRank: Finding topic-sensitive influential twitterers. *Proceedings of the Third ACM International Conference on Web Search and Data Mining WSDM '10.*

Wilhelm, A. (1998). Virtual sounding boards: How deliberative is online political discussion. *Information, Communication and Society*, **1**(3), 313–338. doi:10.1080/13691189809358972.

Williams, C.B., & Gulati, G. J. (2009). Explaining Facebook support in 2008 Congressional Election Cycle. *Working Papers 26, Political Networks Paper Archive*, Retrieved from http://opensiuc.lib.siu.edu/pn_wp/26.

Wojcieszak, M., & Mutz, D. (2009). Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of Communication*, **59**, 40–56. doi:10.1111/j.1460-2466.2008.01403.x.

Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, February 9–12, 2011, Hong Kong, China. 10.1145/1935826.1935863.