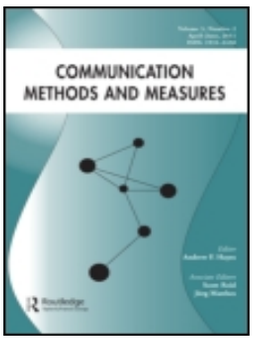


Routledge
Taylor & Francis Group

Communication Methods and Measures

ISSN: 1931-2458 (Print) 1931-2466 (Online) Journal homepage: https://www.tandfonline.com/loi/hcms20

Tweeting Apart: Applying Network Analysis to Detect Selective Exposure Clusters in Twitter

Itai Himelboim , Marc Smith & Ben Shneiderman

To cite this article: Itai Himelboim , Marc Smith & Ben Shneiderman (2013) Tweeting Apart: Applying Network Analysis to Detect Selective Exposure Clusters in Twitter, Communication Methods and Measures, 7:3-4, 195-223, DOI: 10.1080/19312458.2013.813922

To link to this article: https://doi.org/10.1080/19312458.2013.813922

Communication Methods and Measures, 7:195–223, 2013

ISSN: 1931-2458 print/1931-2466 online
DOI: 10.1080/19312458.2013.813922

Routledge
Taylor & Francis Group

Tweeting Apart: Applying Network Analysis to Detect Selective Exposure Clusters in Twitter

Itai Himelboim
University of Georgia

Marc Smith
Social Media Foundation

Ben Shneiderman
University of Maryland

Twitter users see content mostly from the other users they select to follow. Networks of connected users on Twitter define the set of content to which each user is exposed. We developed a Selective Exposure Cluster (SEC) method to study these connected networks and their discussion patterns in Twitter. To illustrate the SEC method, we collected networks of connections among users who talked about a shared topic: the U.S. President's State of the Union speech in 2012. Cluster analysis was applied to identify subgroups of users who were densely interconnected. These users followed users from their own cluster more than they connected to users in other clusters. In each cluster, the primary distributors of information—the hub users—were identified, along with the hashtags, hyperlinks, and top-mentioned usernames in the content of the messages. Each of these indicators was labeled in terms of its political orientation. An analysis of the resulting patterns of selective exposure suggests that users participate in fragmented interactions and form divided groups in which people tune in to a narrow segment of the wider range of politically oriented information sources. We discuss the strengths and weaknesses of the proposed Selective Exposure Cluster method.

Selective exposure describes individuals' selection of information that matches their beliefs, via interpersonal communication and news consumption (McPherson, Smith-Lovin, & Cook, 2001; Gergen, 2008). The rich literature about this phenomenon is grounded primarily in surveys, leaving a deficiency in a high external validity approach to data collection. We propose a network analysis of large social media data method which detects actual, rather than reported, patterns individuals' exposure to peers and information sources.

We apply social network analysis methods to detect patterns of selective exposure in political discussions in Twitter. We first collect topic-specific Twitter data and map the network of

Correspondence should be addressed to Itai Himelboim, Department of Telecommunications, Grady College of Journalism and Mass Communication, University of Georgia, 101L Journalism Building, 120 Hooper Street, Athens, GA 30602. E-mail: itai@uga.edu

users who participated in a discussion and the connections among them that are created by follow, mention, and reply relationships. We then extract several measurements as indicators of selective exposure. First, clusters—subgroups of interconnected users—are identified, using the Clauset-Newman-Moore algorithm. A modularity measurement is calculated to determine the extent to which these clusters are socially bounded, creating a structure of information silos where selective exposure can take place. Second, hubs are identified using in-degree centrality as a measurement of users' exposure to other users. We looked for ideological similarities among hubs in each cluster, which we interpreted as an indication of selective exposure. Third, the frequency of hyperlinks, @usernames, and #hashtags in content posted by users in each cluster is calculated. When two clusters within the same discussion have different frequently used hyperlinks, @usernames, and #hashtags, we interpret this as an indication of selective exposure. By measuring these types of self-similarity across network clusters we can show that users often selectively expose themselves by mostly following like-minded others on Twitter.

We used the free and open source NodeXL tool (Hansen, Shneiderman, & Smith, 2011), which allows users to collect social media network data that can be used to measure the selective exposure of users to one another's messages. NodeXL was adapted to measure selective exposure; this paper proposes ways to use the tool for this purpose, integrating network analysis and automated analysis of use of hashtags and hyperlinks. The approach requires the use of human content coders only to identify the political orientation of the most frequently mentioned hashtags and hyperlinks contributed in each discussion group. In the Methods section, we provide step-by-step instructions for implementing this method with a sample data set collected from users who tweeted about President Obama's 2012 State of the Union address. We report findings that illustrate the polarization of the populations in these discussions. We show that these groups have limited interconnection and make use of divergent content and resources in their discussion of a common topic.

This method makes a novel contribution to the study of selective exposure by quantifying the boundaries of social groups and measuring their similarity to one another. To implement this method, we collected a large log of data about user activity in Twitter. Logs offer some advantages over more commonly used methods such as surveys, experiments, and interviews. Surveys are a widely used tool for examining selective exposure (Chaffee, Saphir, Graf, Sandvig, & Hahn, 2001; Stroud, 2008), along with experiments (Knobloch-Westerwick, 2012) and interviews (Stromer-Galley, 2003). These methods are all based on self reports and are subject to a variety of limitations, including bias and recall (Babbie, 2012). Our method overcomes many of these potential drawbacks, capturing high fidelity details of social connections and shared content. Our approach uses network clusters as the primary unit of analysis, where the political orientation of the collections of individual users and hubs are used to describe each cluster. When examining selective exposure via surveys, in contrast, the individual is the unit of analysis. Therefore, individual-focused methods are limited in terms of their ability to capture data about indirect exposure to information (i.e., friend of a friend). Capturing the social boundaries of exposure to information allows us to evaluate each participant's potential for exposure beyond one's network cluster of directly connected social contacts. This is particularly significant on Twitter, where users often pass along others' content by repeating ("retweeting") it.

SELECTIVE EXPOSURE TO POLITICAL DISCOURSE

Exposure to cross-ideological opinions is considered to be socially beneficial, a perspective that can be traced to Mill (1956), Arendt (1968), Calhoun (1988), and Habermas (1989). Exposure to a range of opinions can help correct errors, encourage the consideration of alternative viewpoints, and expand public discourse to include a wider range of participants in the public sphere. The emergence of the Internet as a popular technology for conversations and information exchange first led to enthusiastic expectations regarding its democratic potential (Rheingold, 1993; Corrado & Firestone, 1996; Hauben & Hauben, 1997). Specifically, Shapiro (1999) expected the Internet to facilitate the distribution of information and perspectives that overcome traditional gatekeepers, such as major media outlets. McKenna and Bargh (2000) suggested that online interactions could facilitate exposure to opinions beyond one's immediate interpersonal social networks.

While the Internet has great potential for cross-ideological exposure, it also allows users to tune out individuals and information sources with whom they disagree (Gergen, 2008). Individuals tend to form new social network connections primarily with others who are often similar to them (McPherson et al., 2001), leading to fragmented interactions and divided groups that are increasingly homogeneous (Van Alstyne & Brynjolfsson, 1996). Focusing on self-exposure to information sources, Sunstein (2006) warned that the availability of a growing number of sources leads to a narrowing of the scope of news and views to which people choose to expose themselves. More recent examination of selective exposure behavior suggests that while people tend not to actively expose themselves to individuals and information sources with whom they disagree, neither will they actively avoid them (Holbert, Garrett, & Gleason, 2010; Garrett, Carnahan, & Lynch, 2011; Parmelee & Bichard, 2012).

Several studies have found indications of selective exposure on online spaces. Krebs (2004) examined sales of political books on Amazon.com and built a network analysis based on books that were related through people who bought them together. He found that book buyers were split into liberals and conservatives in terms of the books they consumed and recommended. Similarly, Adamic and Glance (2005) found political bloggers preferred sending hyperlinks to blogs with similar political orientations and to news sources with like-minded opinions. A study by Hargittai, Gallo, and Kane (2008) found similar patterns of selective exposure. Robertson, Vatrapu, and Medina (2009) found that users who posted links on Barack Obama, Hillary Clinton, and John McCain's Facebook walls linked primarily to websites that supported that candidate. Examining the political interaction on Twitter around major topics of the 2010 midterm elections, Himelboim, McCreery, and Smith (2011) found that users preferred following, mentioning, and replying to users with similar political views. Examining Korean politicians' use of Twitter, Choi, Park, and Park (2011) found that users preferred posting hyperlinks to sites associated with politicians within the same party.

In contrast, other studies provide evidence of cross-ideological exposure in discussion forums. McGeough (2010) found cross-ideological exposure in conversations on Amazon website forums. Kelly, Fisher, and Smith (2006) found indications that individuals often preferred discussing controversial political issues with users with whom they disagreed. Wojcieszak and Mutz (2009) surveyed participation in chat-rooms and message boards. They found that political discussions which took place in explicitly defined political spaces demonstrated low levels of exposure to

crosscutting political views while, in contrast, political discussions that took place in space dedicated to leisure and nonpolitical topics (e.g., hobbies) did exhibit cross-ideological exchange of opinions.

Detecting Selective Exposure: Current Methodological Approaches

In our review of existing studies of selective exposure we found surveys to be a popular methodological approach (e.g., see Chaffee et al., 2001; Stroud, 2008). Surveys, based on large probability samples, are high in external validity and have long been the only way to learn about patterns of selective exposure in large populations. Such studies often ask respondents to report their attitudes, political or otherwise, and their patterns of media consumption (e.g., Gil de Zúñiga, Correa, & Valenzuela, 2012; Hwang, 2010). When examining selective exposure, however, surveys have several drawbacks in terms of their validity. Surveys rely heavily on respondents' recall and their subjective perceptions. Studies indicate that these drawbacks undermine the authenticity of survey responses as a behavioral record (Eysenbach, 2008; Han et al., 2010). Furthermore, users are often motivated when seeking information when the matter is urgent or timely. This is not the case when users are asked to report about past information seeking. This results in less accurate recollection of past behavior. This phenomenon is commonly criticized as a lack of ecological validity (Han, 2008). To conclude, surveys inform us about the media sources that individuals recall and choose to report, not necessarily where they actually sought information.

Another limitation of surveys is their focus on direct exposure. Internet users often forward information from others, highlighting the importance of second and third degree information exposure. This phenomenon is difficult to capture via surveys since people usually do not know where their friends seek information.

Experiments are another common methodology used to examine specific conditions under which selective exposure occurs (e.g., Warner, 2010; Knobloch-Westerwick & Meng, 2011). Experiments, as discussed in the methodological literature at length (Durrheim, 2007; Rubin, Palmgreen, & Sypher, 1994; Babbie, 2012), have some advantages over surveys, including better causality inferences, better control of participants' exposure to stimuli, and, more broadly, higher internal validity. Using an experiment, however, a researcher is forced to limit the options of exposure, as the number of experimental stimuli is limited. This results in a low external validity, as, in reality, individuals can be exposed to a wider array of information sources. Another threat to external validity is the small sample that experiments often rely on (Durrheim).

Interviews (e.g., Stromer-Galley, 2003) can also be applied to the study of selective exposure and may offer better internal validity but their external validity is also very limited.

Our network analysis method addresses many of the limitations of these standard approaches to selective exposure. It can be applied to large social media data sets, providing researchers with the opportunity to address issues of external and internal validity. Furthermore, for experiments and surveys, an individual is the unit of analysis, limiting the examination of exposure to direct exposure. In contrast, a network analysis approach addresses indirect exposure as well as direct exposure.

APPLYING NETWORK ANALYSIS TO THE STUDY OF SELECTIVE EXPOSURE

Social network research is concerned with the consequences of different patterns and types of social ties on attitudes, beliefs, and behavior (Wasserman & Faust, 1999). In social media, users form social networks by forming a list of other users with whom they share a connection. On Twitter, users form relationships when they "follow" one another. One user follows another user by subscribing to their posted content; this content will then appear in the first user's Twitter feed. Users can also form relationships by replying to another user's posted message or mentioning a fellow user in a message by using their user name. These connections are also indications that one user has been exposed to another user's content. In aggregate, these connection activities create social networks (Hansen et al., 2011). Understanding the emergent patterns of social interactions on Twitter can inform our understanding of selective exposure in several ways, including documenting the form of clusters and the roles of key participants who act as hubs in the network. On Twitter, users often interact with different individuals about different topics. These networks may then vary depending on the topic. First, therefore, we explore the topical boundaries of these networks.

Topic Networks and Content Bounded User-Interactions

Twitter users use the service in a number of ways and discuss a variety of topics. Users often use the same Twitter account to exchange information and opinions about a wide variety of topics. In order to investigate selective exposure in a particular domain of inquiry, a subnetwork of Twitter activity is extracted. A topic network, therefore, is defined as the connections among a subgroup of users who posted about a specific topic, specifically a set of keywords, hashtags, or hyperlinks within a given time period. A topic network is therefore the contextual boundary of Twitter activity.

Clusters and the Boundaries of Personal-Exposure

Given the opportunity to interact freely, social actors often create subgroups in which connections internal to the group are more numerous than connections outside that subgroup (Granovetter, 1973; Watts & Strogatz, 1998). Theoretical work in social and other networks has involved considerable effort in describing the network patterns that characterize subgroups (Wasserman & Faust, 1999). A cluster is the concept used by network literature referring to a subgroup in a network in which nodes are more connected to one another than to nodes outside that subgroup (Carrington, Scott, & Wasserman, 2005). Clusters are also the mathematically identified subgroups in networks revealing the community structures found in many connected populations (Newman, 2004).

Researchers have deployed an array of metaphors to describe the idea of subgroups of individuals who selectively expose themselves to politically like-minded others and their information sources. A few examples include "enclave" (Sunstein, 2006), "filter bubbles" (Pariser, 2011), "gated communities" (Turow, 1997), "sphericules" (Gitlin, 1998), "monadic clusters" (Gergen, 2008), and "cyber-balkans" (van Alstyne & Brynjolfsson, 1996). Sunstein's term "enclave" is also a helpful metaphor here as it describes the result of filtering news selection based on political opinions (see also Warner, 2010). Most relevant to this study, however, are "echo chambers,"

a term that Garrett (2009) uses to describe political fragmentation and social polarization resulting from selective exposure to news sources that reinforce one's political opinions. These "echo chambers" take the form of network clusters, a term that will be used hereafter.

Network analysis of social media data hold another advantage over more traditional methodologies for examining selective exposure: It captures indirect exposure. Selective exposure has been traditionally examined in terms of direct exposure to individuals and news sources. Individuals, primarily via social media, are often exposed to information sources indirectly, as users share website hyperlinks and pass content to others (e.g., retweeting). Indirect exposure is therefore an inherent part of selective exposure. Using a cluster as a unit of analysis allows for the examination of both direct and indirect exposure. Within a cluster, some users are exposed directly to a news media organization, for example, by following it. Others may be exposed to that information source indirectly by following others who follow that source. Similarity of content and information sources within a cluster, as indicated in the research questions discussed next, is not only an indication of direct selective exposure, as many of the users within a cluster are directly connected to one another, but also indirect exposure, as users in a cluster are indirectly connected to those whom they do not follow directly.

Network analysis also allows us to identify exposure to the few very popular information sources (i.e., hubs) and exposure among all users. This allows us to compare very popular hubs and regular users. Network analysis provides us with measurements for the relative importance of individual participants as well as algorithms for creating clusters or groups of participants, as detailed in the Methods section. Hub users in a given cluster have the highest level of exposure to other users in that subgroup, and therefore similarities of hubs in clusters can be an indicator of selective exposure. For example, if hubs in one cluster are primarily liberal and in another cluster, primarily conservative, that would be an indication of selective exposure.

The study of selective exposure using network analysis of social media data, like other methodologies, is not free of limitations. Studying selective exposure, researchers are often limited to patterns of exposure to information sources within one social media space. Triangulation of data from different social media spaces may reveal similarities in patterns of selective exposure. However, as individuals often select different usernames across a large number of social media platforms, it remains difficult to capture the complete activity of any given user. A clear advantage of surveys is that the methodology is based on direct interaction with individuals, so one can simply ask about the complete media diet of respondents. Reliability can also be a challenge for the proposed method. Patterns of selective exposure of a group of users in a given point of time may or may not repeat beyond the window of time in which it was collected. To overcome this potential drawback, we propose capturing and analyzing data from several points of time. Similarities in selective exposure across several datasets can provide convergent validity. Based on the conceptual and methodological discussion above, we now turn to the research questions.

RESEARCH QUESTIONS

Social Media Clusters

We define *clusters* in network theoretical terms, using one of a small set of algorithms that group and separate nodes based on their levels of interconnection. Measuring the size, content, and level

of interconnection of these clusters can highlight the extent of selective exposure and provide empirical measures. Social media users are not required to form clusters and groups when using these tools, but, as discussed earlier, often these clusters are formed from network processes. These clusters may be separated from one another or primarily overlap; we discuss measures of division among clusters in the Method section. In order to examine selective exposure on Twitter, the first research question is:

RQ1: *Do users form distinct clusters when contributing to a given topic on Twitter?*

Twitter Hubs

User-interactions via online spaces form a skewed distribution of connections among users, where a few users attract a large and disproportionate number of social ties (Huberman, Romero, & Wu, 2009; Raban & Rabin, 2007). More broadly, this phenomenon is known as a power-law distribution of links among nodes (Newman, 2001). On Twitter, a small number of users called *hubs* are likely to attract a large number of followers, mentions, and replies. In contrast, many users are exposed only to a small number of other users. The formation of social network clusters on Twitter indicates that users form emergent groups of connected users who mutually expose themselves to one another's content. Similarities of the political leaning of hub users in a given cluster, as discussed earlier, can be an indicator of selective exposure in a discussion about a specific issue. Heterogeneity of hubs' political orientation in clusters, in contrast, will be an indication of a lack of selective exposure to political content. The second overarching question is therefore:

RQ2: *Do hubs in a cluster have a consistent ideological standpoint?*

Hashtags, Hyperlinks, and Top Mentioned Usernames and Selective Exposure

Another indication of selective exposure is the content that all users in a given cluster posted. Since tweets are limited to 140 characters, many users post hyperlinks and hashtags in their messages. The websites users select to post hyperlinks to in their tweets can also be used as an indicator of personal exposure. As discussed, clusters define the exposure boundaries of Twitter users. Users within a cluster are primarily exposed to content from other users within the same cluster. An indication of selective exposure is the similarity in the website hyperlinks used in tweets from people in a given cluster. Therefore, the next overarching research question is:

RQ3: *Do the hashtags, hyperlinks, and top mentioned usernames used in a cluster clearly identify with one side of a controversial issue?*

METHOD

This section details our method for examining selective exposure on Twitter. It provides step-by-step guidelines for executing the proposed method. Each step ends with a concrete example relevant to our case study, that is, the selective exposure in discussion on Twitter about the State of the Union Address from the President of the United States.

The free and open source NodeXL can be downloaded from http://nodexl.codeplex.com (see documentation in Hansen et al., 2011). NodeXL supports the exploration of social media with import features that extract network data from a range of data sources such as personal email collections on the desktop, Twitter, Flickr, YouTube, Facebook, Wikipedia, and hyperlinks. NodeXL can import other sources of data through UCINet, Pajek, text, CSV, spreadsheets, or GraphML files.

Next we first provide step-by-step instructions for Twitter topic-network data collection. The Selective Exposure Cluster method, detailed next, identifies major clusters (step 1) and their hubs (step 2), the most frequently posted hyperlinks (step 3), and hashtags and mentioned usernames (step 4) in the messages exchanged. Last, we suggest a few steps for visualizing the data (step 5).

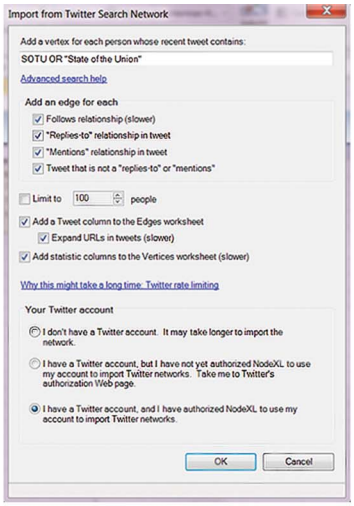
Data Collection – Topic Networks

First, one or more keywords or hashtags are selected that should capture the topical discussion of interest. For example, we used the search terms "SOTU" OR "State of the Union" to capture the conversation about the State of the Union speech. First, open a new NodeXL template workbook. In the NodeXL menu ribbon, in the "Data" menu, select "Import" and then "From Twitter Search Network." See Figure 1a.

Most Twitter content is publically available, which allows researchers to collect and analyze it (content defined by users as "private" is not publically available). NodeXL users can select which of the types of relationships between Twitter users they are interested in identifying: mentions, replies, and follows. The Tweet option allows users to capture tweets that were not associated with mentions or replies, and thereby increase the volume of messages in the data collection. Users may also limit the number of users from whom NodeXL will collect data. We recommend selecting the option to add a tweet column, so the content of messages will be captured as well. The "expand URL in Tweets" option is important, as many users post short versions of hyperlinks (e.g., bit.ly). When selected, NodeXL will convert these hyperlinks into their ultimate form.

The first time Twitter data are collected using NodeXL on a particular computer, it is necessary to authenticate NodeXL with Twitter by selecting the second option ("*I have a Twitter account, but I have not yet authorized NodeXL to use my account to import Twitter networks. Take me to Twitter's authorization Web page.*") from the "Twitter Account" section of the Search Import dialog box. This will take users to the Twitter website where users should approve the connection between Twitter and NodeXL by copying the numeric code Twitter provides. Paste this number in the newly opened NodeXL dialog box. After this process is performed it does not need to be repeated. Subsequent uses can click the third option ("*I have a Twitter account, and I have authorized NodeXL to use my account to import Twitter Networks.").* Data collection may take some time, even several hours (NodeXL does not indicate the time remaining for data downloading). When data appear in the Excel worksheets, data collection is completed. The workbook can now be saved to create an archive of the collection of tweets and the collections of connections among the participants. Twitter allows access only to recent data, up to a month or so, where the maximum number of users varies from 1,200 to 1,500 most recent users. Unfortunately, at this point, we cannot search past Twitter activity. Collecting data about a specific topic in different points of time will produce different data sets.

Two of the NodeXL worksheets are named "Edges" and "Vertices." The Edges worksheet contains the relationships among Twitter users who posted about the topic. For instance, the user



(a)

	A	B	N	O	P	Q	R	S	T	U	V
1			Other Columns								
2	Vertex 1	Vertex 2	Add Your Own Columns Here	Relationship	Relationship Date (UTC)	Tweet	Original URLs in Tweet	URLs in Tweet	Domains in Tweet	Hashtags in Tweet	Tweet Date (UTC)
3	_____bo	_____bo	ssman	Tweet	1/27/2012 2:22	fnna watch that State of The Union stuff for mr. TenBrook					1/27/2012 2:22
4	_callmewarc	_callmeward	_	Tweet	1/27/2012 3:14	I bet Obama got high as hell after the state of the union					1/27/2012 3:14
5	_carriep	whitneypitcher		Mentions	1/27/2012 4:03	RT @whitneypit	http://t.co/XG6aH	http://www.y	youtube.com	#sotu	1/27/2012 4:03
6	_carriep	atlasshrugs		Followed	1/27/2012 4:00						
7	_carriep	common_sense4u		Followed	1/27/2012 4:00						
8	_carriep	whitneypitcher		Followed	1/27/2012 4:00						
9	_carriep	cdnnow		Followed	1/27/2012 4:00						
10	_carriep	thorninaz		Followed	1/27/2012 4:00						
11	_luisantonio	_luisantonio		Tweet	1/27/2012 2:44	If Puerto Rico became a state, it'd simply become the next Rhode Island. #CNNDebate					1/27/2012 2:44

(b)

	A	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1													
2	Vertex	Followed	Followers	Tweets	Favorites	Description	Locatic	Web	Time Zone	Time Zone UTC Offset (Seconds)	Joined Twitter Date (UTC)	Custom Menu Item Text	Custom Menu Item Action
3	jjauthor	19739	20342	91946	461	Janie Johnson takes a stand for conservatism, patriotism	Nevada, U	http://w	Pacific Time (U	-28800	8/12/2010 14:41	Open Twitter Page	http://twitter.com
4	stevenerte	31884	33418	25153	6	Founder of LifeNews.com, the pro-life news service. Thi	Rocky Mc	http://w	Mountain Time	-25200	6/24/2009 20:52	Open Twitter Page	http://twitter.com
5	dbargen	8282	7937	23541	51	Conservative; libertarian leaning. Adoptive parent. Cons	Atlanta		Indiana (East)	-18000	3/7/2009 14:18	Open Twitter Page	http://twitter.com
6	katyinindy	73300	68426	76104	179	Cogito, ergo sum conservative. 'The price of freedom is	USA	http://k	Eastern Time (U	-18000	2/16/2009 3:08	Open Twitter Page	http://twitter.com
7	postpolitic	1659	36190	19949	10	Political coverage from @WashingtonPost. Get social w	Washingt	http://w	Eastern Time (U	-18000	4/9/2008 20:06	Open Twitter Page	http://twitter.com

(c)

FIGURE 1 Data collection and layout in NodeXL: (a) Importing Twitter Data; (b) Layout of data – Relationships; (c) Layout of data – Users (color figure available online).

@_carriep followed the user @cdnnow. If the edge is a reply, mention, or just a plain tweet, the full content of the Tweet appears in the Tweet column of the Edges worksheet in NodeXL. Information about the date and time of the edge's creation are included.

The Vertices worksheet in NodeXL provides information about each Twitter user whose message appears in the data set. This worksheet displays the self-description, activity statistics, and the hyperlink to their Twitter account. This worksheet also displays any network analysis metrics that are calculated about individuals in the graph (Figures 1b and 1c).

Illustration: A Case Study

We analyzed collections of messages that appeared on Twitter related to a selected topic. In order to capture these Twitter messages, we collected data by searching for selected keywords and hashtags using the NodeXL tool. In this study we used the terms "State of the Union" and "SOTU," a frequently used acronym for the presidential address. The Boolean string "State of the Union" OR "SOTU" was used to capture users who tweeted about either or both of these keywords. These queries resulted in a data set containing messages from roughly 1,000 users who tweeted a message containing these keywords in a time period close to the time of the query. The data included the content of the users' self-descriptions and the statistics Twitter maintains about their activity like the number of tweets, the time each user joined Twitter, the users' number of followers, and related information. We processed this data to generate information about the ways these users connected to one another through follow, mention, and reply relationships. A "follow" is created when one user adds another user to their list of people whose tweets should be displayed. A "reply" or "mention" connection is created when a user creates a message that contains the user name of another user. Replies are messages in which the user name appears in the very beginning of a Tweet, while in a mention the username appears anywhere else in the message. We collected data for the three days subsequent to the address, January 26–28, 2012, in order to capture the conversations triggered by the speech. This resulted in three Twitter network data sets that contain snapshots of the connections among people who mentioned the State of the Union on Twitter.

Selective Exposure Cluster Method

The proposed Selective Exposure Cluster (SEC) method captures selective exposure social networks measures. We explain how to calculate relevant network measurements using NodeXL next, followed by guidelines for their interpretation. In the NodeXL ribbon, find the Analysis section. Open the Groups drop down menu and select Group by Cluster. Select the Clauset-Newman-Moore clustering algorithm. Also check the "Put all neighborless vertices into one group" box. See Step 2, below, for an explanation of this clustering option. Click OK.

Network and content attributes can provide indicators of selective exposure in a topic space. The size and level of connection between clusters, the self-descriptions of hub and top mentioned users in each cluster, and the domain names and hashtags frequently used within each cluster can be contrasted.

To generate network metrics, select the Graph Metrics option from the Analysis section and select the following metrics for calculation:

Overall graph metrics: calculates statistics about the entire network, including Modularity, a measure of the quality of the algorithmically defined clusters. The values for these calculated metrics appear in the Overall Metrics worksheet.

Vertex In-Degree: we identify a hub based on the number of users who followed, mentioned, and replied to them. This measure captures this social activity, indicating how prominent each user is in terms of the exposure to users' posted content. Measures of each user's network properties are displayed in the Vertices worksheet.

Group metrics: calculates a range of cluster-specific measurements, including the number of users and relationships in each cluster. This information is used to identify the top clusters. These results appear in the Groups worksheet.

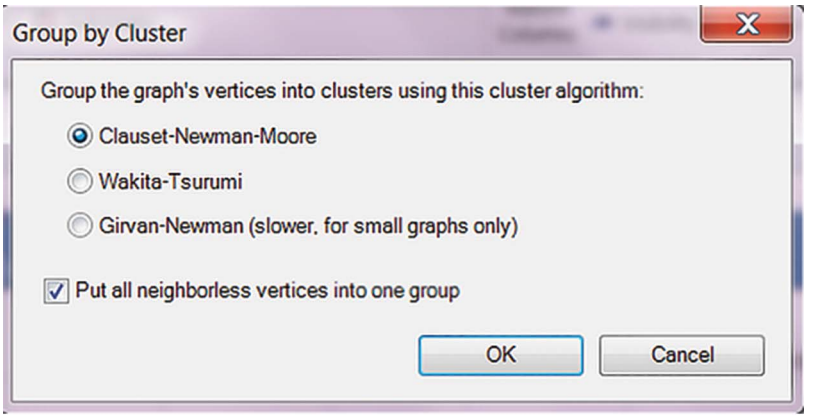
Twitter search network top items: this calculates the most frequently used hashtags, hyperlinks, and mentioned usernames in the tweets in the dataset. Results will appear in the Twitter Search Ntwrk Top Items worksheet (Ntwrk is an abbreviation of Network). Next, selective exposure is measured via measurements for clusters, hubs, hyperlinks, hashtags, and most mentioned users. We propose five steps for the analysis. We provide links to both the raw and the analyzed datasets (in graphs 4–6) and we encourage the readers to follow these steps with the dataset (downloaded via the hyperlinks) as an exercise (Figures 2a and 2b).

Selective Exposure Cluster Method, Step 1: Clusters

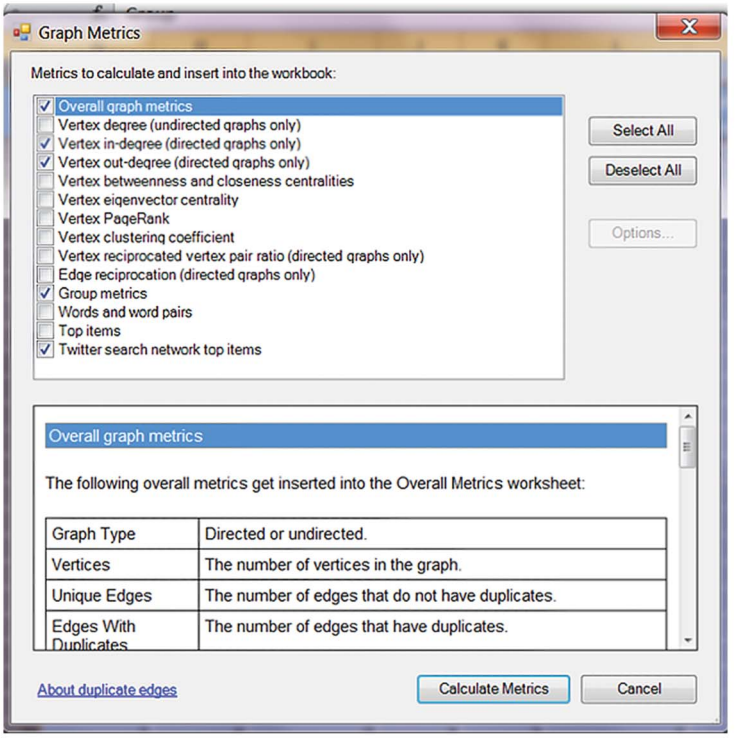
Identifying clusters is a required first step, because each of the following selective exposure indicators are identified and evaluated within their respective clusters. We identified clusters of relatively more connected groups of users in the topic-networks using the Clauset-Newman-Moore algorithm (Clauset, Newman, & Moore, 2004), which is included in the NodeXL software. We selected this algorithm for its ability to analyze large network data sets and efficiently find subgroups. Two other options are available: the Girvan-Newman and Wakita-Tsurumi algorithms. Both are slower for the analysis of such large data sets.

We use Newman's (2004) measurement of modularity to measure the quality of the divisions imposed on the network. Modularity measures the extent to which the divisions among clusters is a good one, in the sense that there are many links within clusters and only a few between. Modularity values range between zero and one. The higher the modularity value, the more distinct or separate the clusters are; that is, the clusters are less connected to one another. However, beyond a modularity value of .6, networks show little or no increase in the division of clusters. Newman therefore suggests 0.6 as a threshold for detecting meaningfully distinct clusters. More recent work by Wang (2012) compared modularity measures across different clustering algorithms and found .4 to be a sufficient threshold for accepting the results of a clustering algorithm. Wang showed that sensitivity to divisions among clusters reaches a plateau for values higher than .4. For the purpose of this study, we suggest .6 as a threshold for high modularity fitness and .4 as a medium level. In other words, a modularity value smaller than .4 is considered low separation among clusters, between .4 and .6 we consider medium, and .6 and above is high. In a study of selective exposure, a more divided (high modularity) network indicates more selective exposure: users within a cluster are more exposed to one another and less exposed to content from users outside that cluster.

Clustering algorithms often divide a network into a few major clusters and several small ones. We ranked clusters by the number of users assigned to each. This creates a "Scree Plot," where a



(a)



(b)

FIGURE 2 (a) Cluster analysis; (b) Network analysis (color figure available online).

few clusters contain most of the users in the network. These clusters accounted for the majority of connected users and relationships, as detailed in the Findings section.

Selective Exposure Cluster Method, Step 2: Hubs

Through a process of preferential attachment, a few individuals in a network often attract a large and disproportionate number of connections from other users (Newman, 2001). Mathematical models of network growth describe a process of preferential attachment by which the likelihood of the addition of a new connection to a node is dependent on the number of connections it already has (Newman, 2001). This is a variation of the "rich get richer" model (sometimes called the Matthew Effect in sociology), which reflects the reality in many Twitter communities, that is, a small number of users, called hubs, have a much larger than average number of connections. On Twitter, preferential attachment means that very few users have many more followers, mentions, and replies than most other users. These hub users are often the core of the cluster in which they are located. Because of the skewed distribution of connections, hub users are likely to capture a significant fraction of the attention of other users in a cluster. We identified the curve in the distribution of connections in each cluster.

Illustration: A Case Study

We collected tweets about the State of the Union data, and the resulting networks were analyzed following the steps described. We identified clusters and modularity values for the whole network and calculated in-degree for each participant in the data set. A user's in-degree measures the number of follows, mentions, and replies that user attracted in the network. We ranked users by their in-degree metric in descending order and selected the five users who had the highest in-degree. These top five users were selected by the number of their connections with other users. Selecting top users in each cluster, rather than in the entire network, allowed us to identify hubs in clusters without regard to the size of hubs in other clusters, thus finding the local leaders in each cluster. In some cases there are fewer than five hub users in a cluster, for example in the cluster surrounding President Obama's Twitter account, we only selected the Obama account as a hub, as all other users attracted very small number of followers.

We assigned a political orientation to each hub, labeling each as conservative, liberal, neutral, or unclear. We based this categorization on political orientation statements in hubs' self-description text, including posted hyperlinks. The political orientation of each hub was coded by both the first author and another coder (Scott's Pi = .97).

Selective Exposure Cluster Method, Step 3: Hyperlink Analysis

Users in all clusters often posted messages containing hyperlinks to external resources on the web. The political orientation of these resources is another indicator of the general orientation of the users in that cluster. NodeXL captures the hyperlinks used in tweets created by users in each cluster. The tool analyzes the hyperlinks by counting mentions of both complete web addresses and mentions of the pared-down domain name. NodeXL reports the frequency of mentions of each domain for the network as a whole and for each cluster. In our analysis of State of the Union related tweets, we classified each domain name (e.g., a newspaper, cable TV show, a blog) based

on its political orientation as conservative, liberal, neutral, or unclear. We classified hyperlinks based on the content of the web page. All hubs were coded by the first author and another coder (Scott's Pi = .96), and the most used domains in each major cluster are reported.

Selective Exposure Cluster Method, Step 4: Analysis of Hashtags and Mentioned Usernames

Messages from Twitter users often contain hashtags as well as the mention of other users by their @username. The use of specific hashtags and the mention of specific users in the content of tweets is another indication of the political orientation of the users in a cluster. In the State of the Union data sets, we calculated the frequencies of the mention of users and hashtags and ranked them in descending order. We identified the most frequently mentioned hashtags and users based on frequency count, then classified each hashtag and each mentioned user based on their political orientation as conservative, liberal, neutral, or unclear. We based our classification on the most mentioned users' self-description text, and we analyzed hyperlinks and hashtags based on their content. The first author and another coder classified top mentioned users and hashtags (Scott's Pi = .94).

Selective Exposure Cluster Method, Step 5: Visualization

Mapping and visualizing the social network created by users and their exposure relationship on Twitter is a helpful way to convey patterns of selective exposure. NodeXL allows for the visualization and customization of network data. Click "Show Graph" either in the NodeXL ribbon or on the Document Action window. The graph will appear and may look somewhat unorganized. A few simple steps will help with laying out the data in a more meaningful way. In the edges spreadsheet, click the "Relationships" drop menu and deselect the "Tweet" value. This step removes "self-loops" from the graph. Open the Algorithm dropdown menu and select layout options at the bottom. In the dialogue box, select the option "Lay out each of the graph's groups in its own box and sort the boxes by group size." Click OK and then "Refresh Graph." The new layout highlights the clusters, making the links within and between clusters visible. If you would like to add information that describes each cluster's content (e.g., the most frequently used hashtags), select the Groups spreadsheet. The column "Labels" allows you to add text for each cluster. Add the text and click Refresh Graph. Many customization options are available on NodeXL, and we encourage readers to explore them.

RESULTS

We built a data set containing a collection of tweets related to the 2012 State of the Union Address. We collected the maximum amount of data the Twitter API (application programming interface) allowed, which generated data sets ranging from 1,200 to 1,500 of the most recent users who posted messages containing the given keyword or hashtag. The January 26 data set included 1,244 users, 619 of them connected by 4,304 unique relationships (i.e., follows, mentions, or replies). The January 27 data set included 1,309 users, 928 of them connected by 4,897 unique relationships. The January 28 data set included 1,276 users, 652 of them connected

by 3,979 unique relationships. We mapped each data set based on the relationships among its users and applied network analysis to identify clusters and hubs as discussed next.

Clusters

RQ1: *Do users form distinct clusters when contributing to a given topic on Twitter?*

We used the Clauset-Newman-Moore algorithm to analyze and cluster the networks into subgroups. In the January 26 network, we discovered two large clusters—one with 188 users linked by 1,904 connections and the other with 175 users linked by 1,036 connections. These two major clusters accounted for 58.6% of all connected users and 63.8% of all relationships in the network. The modularity value for these clusters was .51, suggesting a moderate level of separation among clusters.

The January 27 network revealed three major clusters—one with 312 users and 559 relationships, a second with 223 users and 1,888 relationships, and the third with 180 users and 1,059 relationships. These three major clusters accounted for 77.1% of all connected users and 71.6% of all unique relationships in the network. The modularity value was .45, suggesting a moderate level of separation among clusters.

The January 28 network had two large clusters—one with 169 users connected by 1,680 relationships and the second with 103 users and 530 relationships. These two major clusters accounted for 41.7% of all connected users and 55.5% of all unique relationships in the network. The modularity value was .47, suggesting a moderate level of separation among clusters. Another way to illustrate the separation among clusters is by comparing the percentage of connections within clusters to the percentage of connections between them (see Figure 3). In response to research question one, measurements of modularity and inter-cluster links, then, indicate that users form medium to high levels of cluster separation when contributing to a given topic on Twitter.

Hubs

RQ2: *Do hubs in a cluster have a consistent ideological standpoint?*

To identify hub users, measures of each user's connections and position in the network were calculated. These measures include the number of relationships directed toward a user (i.e., its in-degree centrality). We selected the top five hubs for each cluster.

Findings indicate that primarily grassroots conservatively leaning hubs, such as bloggers and pundits (e.g., @Jjauthor who "Takes a stand for conservatism, patriotism & optimism"), appeared together in clusters in all data sets (January 26, 27, and 28), creating clusters with a consistent conservative ideological standpoint. In the January 26 and 27 data sets, liberally leaning hubs were institutional in nature (e.g., @OFA_N, Obama for America in North Carolina; @EricBoehlert of the progressive Media Matters). These hubs appeared in clusters together with news media (e.g., @Postpolitics of the *Washington Post*), creating clusters of liberal institutions and news media information sources. The January 28 data set produced one cluster, which included conservative grassroots hubs, similar to conservative clusters in earlier data sets. In the second cluster, hubs were grassroots liberal bloggers (e.g., @defocus, who describes herself as "Political junkie on the left"), unlike earlier clusters that mixed institutional liberal sources and news media. The January

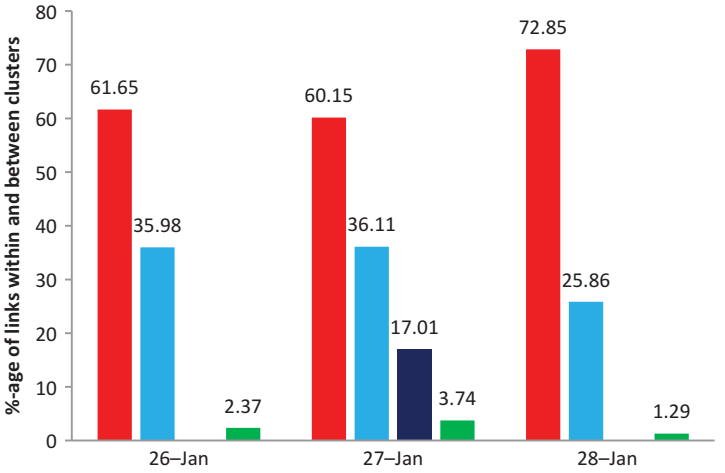


FIGURE 3 Inter-cluster and intra-cluster Twitter relationships (in percentages). Bar chart of the percentage of links within each of the major clusters and links that cross cluster boundaries. Percentages are based on the total number of links that started in each cluster that appeared within the main clusters and across them. Colors are assigned to illustrate the political orientation of hubs in each cluster (red for conservative, light blue for mix of traditional media and liberal hubs, dark blue for the Obama cluster in 1/27, and green for relationships that crossed cluster boundaries) (color figure available online).

27 data set was the only network that formed a third cluster surrounding President Obama's Twitter account. The President's account was the only hub in that cluster. Table 1 details all hubs in clusters.

NodeXL also allows researchers to visualize their networks. Figure 4 maps users who tweeted about the State of the Union and their relationships (follow, mention, and replies) in the January 26 data set. This social network illustrates the two major clusters that were formed, the interconnectedness within each cluster, and the scarcity of cross-cluster communication. Figure 5 shows a three-cluster network, the conservative cluster (top-right), the liberal/news cluster (bottom-left), and the Obama cluster (top-left). This visualization illustrates not only the separation of clusters, but also the more frequent connection between the Obama cluster and the liberal/news cluster than between any other two clusters. Figure 6 shows the January 26 data set, with two major clusters. Readers can find links to the raw and analyzed NodeXL datasets next to each of the figures. The links to the analyzed data sets also provide images in color and higher resolution.

Hashtags, Hyperlinks, and Mentioned Usernames Analysis

RQ3: *Do the hashtags, hyperlinks, and top mentioned usernames used in a cluster clearly identify with one side of a controversial issue?*

TABLE 1
Hubs by Cluster

Date	*Cluster*	*Twitter User*	*Brief Description*
1–26	Cluster 1	@Jjauthor	"Takes a stand for conservatism, patriotism & optimism"
		@stevenertelt	"Founder of LifeNews.com, the prolife news service"
		@Dbargen	"Conservative; libertarian leaning"
		@Katyinindy	"Cogito, ergo sum conservative"
		@atlasshrugs	"Obama's War on America and Stop the Islamization of America."
	Cluster 2	@Postpolitics	Washington Post
		@Thelastword	MSNBC news commentator Lawrence O'Donnell
		@OFA_FL	Obama for America in Florida
		@OFA_NC	Obama for America in North Carolina
		@johnfmoore	"Founder and CEO of @GovInTheLab, trying to bridge the divide between citizens, politicians, and municipal employees. Politics, open government, human rights."
1–27	Cluster 1	@BarackObama	President Obama
	Cluster 2	@Heritage	A conservative think tank
		@JonahNRO	Jonah Goldberg editor, columnist and the author of "Liberal Fascism,"
		@coutpost	Conservative Outpost a conservative blog,
		@amandacarpenter	Amanda Carpenter a conservative blogger, author, and commentator.
		@mastadonarmy	Mastadon Army ". . . put the best candidate in the GOP seat."
	Cluster 3	@markknoller	CBS News White House Correspondent Mark Knoller
		@EricBoehlert	Eric Boehlert, a Senior Fellow at the progressive Media Matters for America watchdog
		@NorahODonnell	Norah O'Donnell, CBS News Chief White House correspondent NO'Donnell
		@WHLive	The White House Live account
		@CollegeDems	College Democrats
1–28	Cluster 1	@jjauthor	Janie Johnson, "standing for conservatism"
		@exposeliberals	Expose Liberals, a "Fiscally conservative, socially conservative / libertarian"
		@coutpost	Conservative Outpost, a conservative blog.
		@obama_games	Big Guns, ". . . our government is out of control with all this wasteful spending,"
		@IndyEnigm	Marty Smith, stands for "Conservative principles" and "Libertarian ideals"
	Cluster 2	@defocus	Bonnie Lesley, a "Political junkie on the left"
		@LeftsideAnnie	A "godless liberal"
		@sunshineejc	Wishes to make sense of "Globalization's effect on domestic economy"
		@A_ThinkingGirl	"Liberal perspective, occupy the vote. Obama2012"
		@UnshackleUS	Captain Clarion, "pragmatic progressive"

*All quotations are from users' self descriptions.

In the January 26 data set, we identified two major clusters and ranked the hyperlinks mentioned in each by frequency of mention. In a cluster containing conservative hubs, the most frequently mentioned domain names primarily linked to article-sharing and search sites (typepad.com with 10 uses; google.com with 8; youtube.com with 7). The conservative ronpaul2012.com appeared in 9 tweets and blogspot.com in 6. The most mentioned users were @barackobama

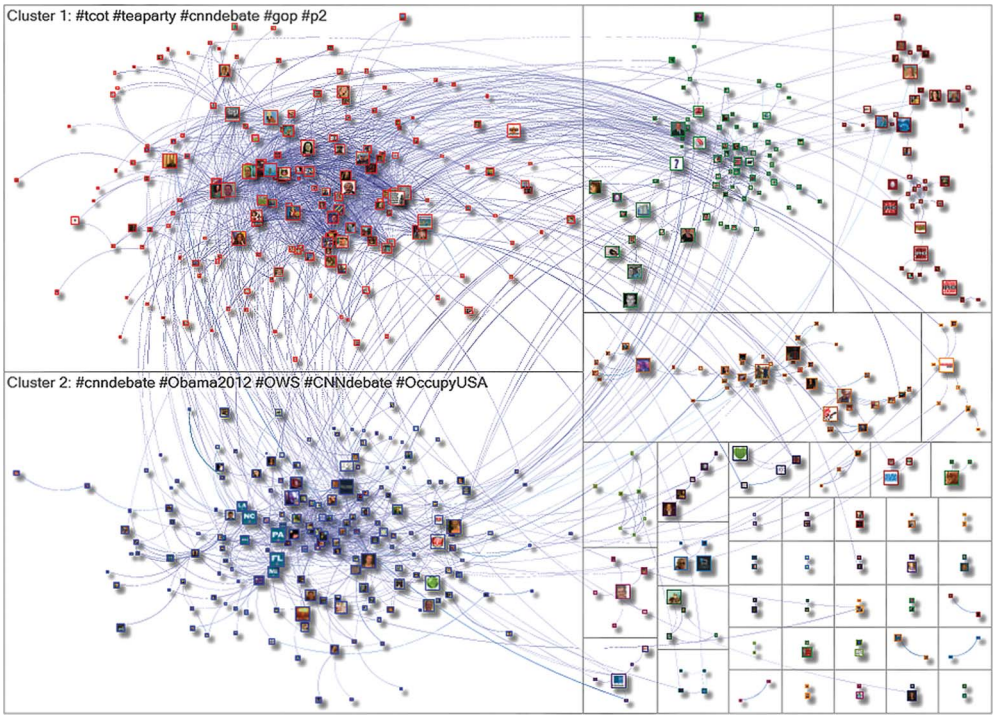


FIGURE 4 The State of the Union social network, January 26, 2012. Twitter users are represented by their profile image. Their relationships are represented by arrows. This graph illustrates the two major clusters that emerged: at the top-left the cluster where hubs indicate a conservative leaning and in the bottom-left, the cluster where hubs associate themselves with traditional media or liberal opinions.
We made the datasets public via the *NodeXL Graph Gallery*. Follow the "Download the Graph Data as a NodeXL Workbook" link at the bottom of the page. We recommend saving the file in a folder, before opening it. The raw data set can be downloaded from http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=2820
The analyzed data set can be downloaded from http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=2819
To open a file, go to the Data section in the NodeXL menu and open the Import dropdown menu. Select "From NodeXL Workbook Created on Another Computer" and browse to the location of the file you just downloaded from the NodeXL Graph Gallery (color figure available online).

(11), @realdonaldtrump (7), @ronpaul (7), and @blacksheeprpt (7). The most frequently used hashtags were #tcot ("Top Conservatives on Twitter"; 42 uses), #teaparty (14), #cnndebate (9), #gop (8), #p2, ("Progressives 2.0" – A liberal hashtag; 7), #gop2012 (7), and #RonPaul (7).

A second major cluster containing hubs associated with traditional media and liberal-leaning users also included more generic article-sharing and search websites (youtube.com, twitter.com

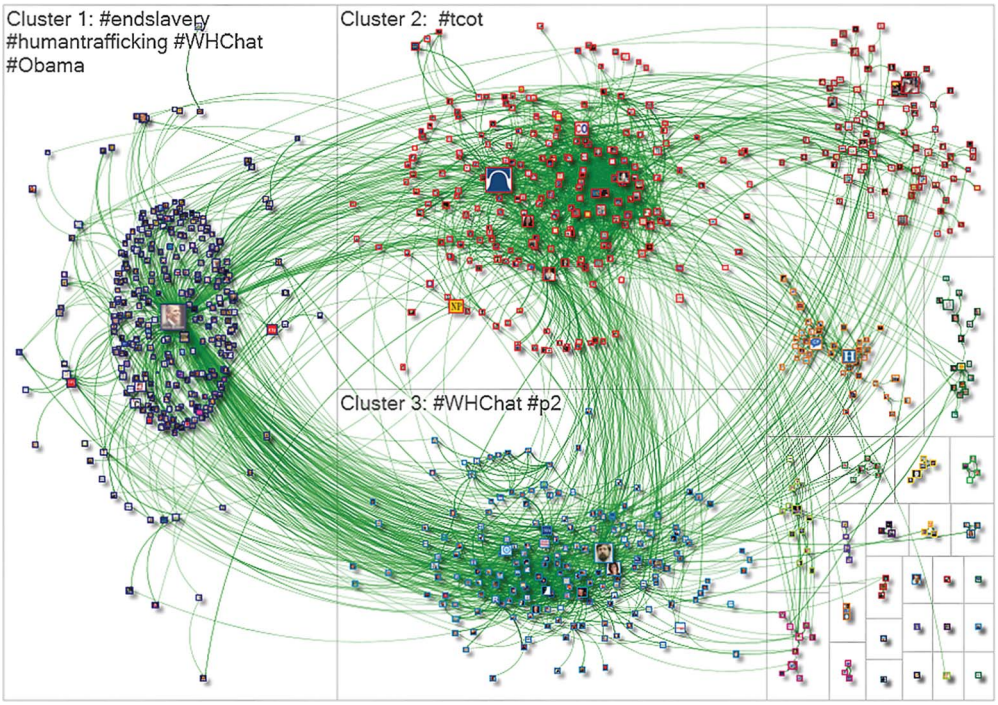


FIGURE 5 The State of the Union social network, January 27, 2012. Twitter users are represented by their profile image. Their relationships are represented by arrows. This graph illustrates the three major clusters that emerged: at the top-left the cluster with the twitter user account for President Barak Obama is the dominant hub; hubs with a conservative leaning are in the cluster in the top-right; in the bottom-left cluster the hubs associate themselves with traditional media or liberal opinions. We made the datasets public via the *NodeXL Graph Gallery*. Follow the "Download the Graph Data as a NodeXL Workbook" link at the bottom of the page. We recommend saving the file in a folder, before opening it.
The raw data set can be downloaded from http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=2818
The analyzed data set can be downloaded from http://www.nodexlgraphgallery.org/Pages/Graph.aspx?graphID=2782
To open a file, go to the Data section in the NodeXL menu and open the Import dropdown menu. Select "From NodeXL Workbook Created on Another Computer" and browse to the location of the file you just downloaded from the NodeXL Graph Gallery (color figure available online).

and google.com, with 11, 9, and 5 uses), as well as liberal-leaning sites (barackobama.com with 13 uses), mainstream media (time.com, 5), and huffingtonpost.com (4). The top mentioned users were @barackobama (10), @samyoungman (the account for Sam Youngman, a political campaign correspondent for Reuters; 7), @yougottavote (self-described as progressive; 5), @annybush (self-described as "Former Republican turned Die-Hard Liberal"; 4), and @ richardwolffedc

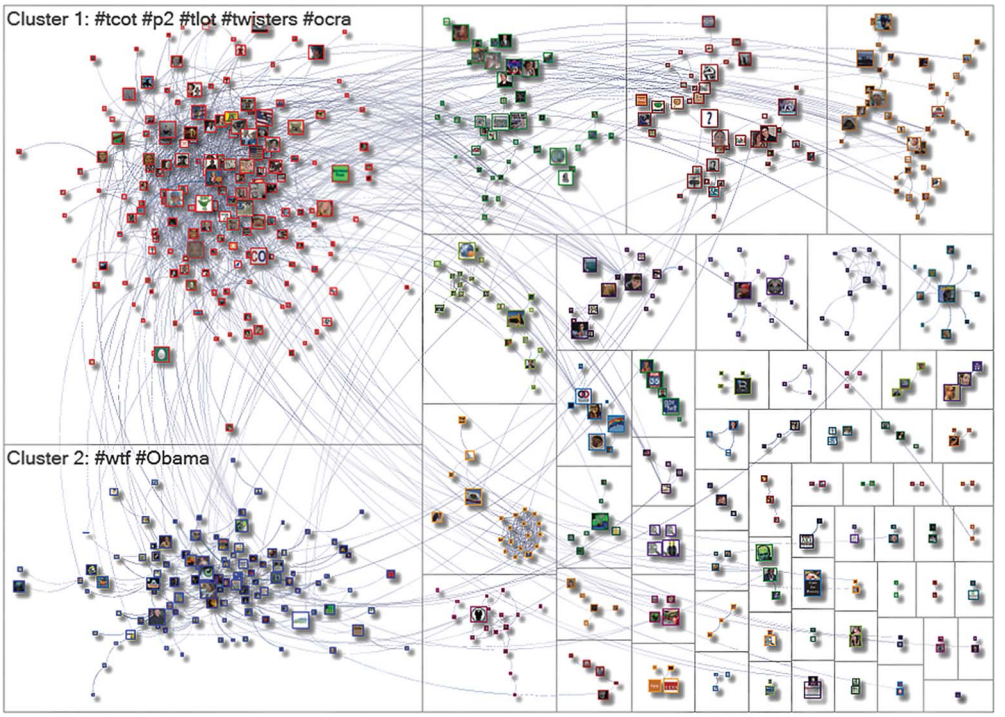


FIGURE 6 The State of the Union social network, January 28, 2012. Twitter users are represented by their profile image. Their relationships are represented by arrows. This graph illustrates the two major clusters that emerged: at the top-left the cluster when hubs indicate a conservative leaning and in the cluster and in the bottom-left, when hubs associate themselves with liberal opinions. We made the datasets public via the *NodeXL Graph Gallery*. Follow the "Download the Graph Data as a NodeXL Workbook" link at the bottom of the page. We recommend saving the file in a folder, before opening it.
The raw data set can be downloaded from http://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=2817
The analyzed data set can be downloaded from http://www.nodexlgraphgallery.org/Pages/Graph.aspx?graphID=2808
To open a file, go to the Data section in the NodeXL menu and open the Import dropdown menu. Select "From NodeXL Workbook Created on Another Computer" and browse to the location of the file you just downloaded from the NodeXL Graph Gallery (color figure available online).

(the account for Richard Wolffe, an MSNBC political news analyst; 4). The most frequently used hashtags were #cnndebate (10), #Obama2012 (8), #OWS (6), #CNNdebate (6), and #tcot (6).

In the January 27 data set, we identified three major clusters and ranked the hyperlinks mentioned in each by frequency of mention. In one cluster, which included the Twitter account for President Obama along with other news and liberal political hubs, the most frequently mentioned

domain names were ijm.org (international justice mission) with 30 hyperlinks, barackobama.com (26 references), youtube.com (20) and mediabistro.com (a links-sharing site; 11). The conservative nationalreview.com and the newspaper nytimes.com had 4 references each. In this cluster the most mentioned users were @barackobama (76 mentions), @ijmhq (international justice mission headquarters; 25 mentions), @ijmcampaigns (25 mentions), and @iswanthehill (the account for Ian Swanson, the Associate Editor for the *Washington Journal*; 15 mentions). Less commonly mentioned was the conservative representative @ericboehlert (7 mentions). The most used hashtags were #endslavery, a nonprofit organization that raises awareness about human trafficking issues (28), and #humantrafficking (28). Less frequently used hashtags included #WHChat (8) and #Obama (7).

The second major cluster included hub users who were conservative bloggers. In this cluster the most frequently mentioned URL domains were the conservative nationalreview.com (which appeared in 35 hyperlinks), washingtonpost.com (15), the conservative blog commonamericanjournal.com (12), the blogging host site blogspot.com (9), google.com (8), and the conservative think-tank heritage foundation herit.ag (8). The top mentioned users were @jonahnro (the account for Jonah Goldberg, a writer for the Tyranny Blog for the conservative National Review; 30 mentions), @barackobama (13), @amandacarpenter (the account for Amanda Carpenter, the author of *Pickpocket: How big-government regulations, bureaucracy, taxes, and out-of-control spending rob the taxpayer*; 13 mentions), @heritage (the account for the Heritage Foundation, a conservative think-tank; 9 mentions), @realdonaldtrump (the official Twitter profile for Donald Trump; 8 mentions), and @beesnguns (the account for a writer for the blog Common American Journal that intends to "inspire our fellow Patriots;" 8 mentions). The single most popular hashtag was #tcot (Top Conservatives on Twitter; 17 times).

The third cluster included hub users with a liberal orientation as well as traditional media outlets. Fewer hyperlinks appeared in this cluster. The most frequently mentioned domain name, nationalreview.com, was found in 15 hyperlinks. Others were even less frequently mentioned: google.com (10), twitter.com (5), marketwatch.com (a business news aggregator; 4), publicpolicypolling.com (a polling website; 4), usa.gov (4), and ncpssm.org (National Committee to Preserve Social Security and Medicare; 4). The most frequently mentioned hashtags were #WHChat (White House Chat; with 31 uses), #ObamaDerangementSyndrome (15), and the liberal hashtag #p2 (12). The most mentioned users were @ericboehlert (the account for Eric Boehlert, a Senior Fellow at Media Matters for America; 37), @jonahnro (the account for Jonah Goldberg, a writer for the conservative National Review; 15), @markknoller (the account for Mark Knoller, a CBS News White House Correspondent; 14), @childerbrant (the account for Chris Hilderbrant, a civil rights activist; 12), and @barackobama (9).

In the January 28 data set, we identified two major clusters. One cluster contained hub users identified as conservative leaning. The top domain names in hyperlinks mentioned in this cluster were sites often used for user-generated content, including google.com (17), youtube.com (11), twitlonger.com (8), and blogspot.com (5). The most frequently mentioned users in this cluster were @barackobama (13), @cspanwj (the account for the C-SPAN morning call-in program; 10), @jjauthor (the account for Janie Johnson, who self-describes as someone who "takes a stand for conservatism, patriotism & optimism"; 8), @barry_o44 (a conservative Twitter user; 7), and @nicksreport (self-described as a "Reagan Republican"; 7). The top hashtags used in tweets from were #tcot (54), #p2 (16), #tlot (Top Libertarians on Twitter; 13), #twisters (representing female conservatives on Twitter; 12), and #ocra (the Organized Conservative Resistance Alliance; 11).

We identified the second largest cluster based on its liberal-leaning hubs. The top domain names used were youtube.com (9), epi.org (the liberal Economic Policy Institute; 6), nytimes.com (6), blogspot.com (5), and barackobama.com (4). The most mentioned users were @barackobama (17), @21stprincipal (self-described as "an occasional foray into liberal politics"; 13), @brian8473 (self-described liberal; 6), @march4teachers (a teacher; 6), and @eacheronthemic (an account that was removed from Twitter; 6). The most popular hashtags were #wtf (6) and #Obama (5).

In response to the third research question, hashtags, hyperlinks, and top mentioned users who are associated with right-wing political views appeared together in clusters. We found that their left-wing counterparts were present in mixed clusters that included hashtags, hyperlinks, and top mentioned users associated with traditional media. This mix may reflect a common perception by conservative individuals that mass media like the *New York Times* and the *Washington Post* hold liberal views. See Table 2.

DISCUSSION

This paper proposes a Selective Exposure Cluster method for the network analysis of large social media log data which can identify clusters that have common discussion topics. Network analysis of large data sets that capture actual self-exposure of users to one another's content addresses the limitations of other commonly used methods. While survey research has internal validity limitations that result from personal reporting, experiments face external validity issues caused by small samples and controlled settings. This network-based SEC method, therefore, is novel as it allows for the examination of actual activity of large groups of individuals. It also examines patterns of selective exposure in the naturally occurring boundaries of personal selection of information sources and social interactions beyond the social horizon of one's immediate contacts. We first evaluate the indicators used here to identify selective exposure and follow with a discussion of the advantages of the SEC method and its limitations. We conclude with a short discussion of the specific network analyzed here.

Evaluation of Selective Exposure Indicators

We used clusters, hubs, hashtags, hyperlinks, and top mentioned usernames as indicators for selective exposure. Cluster analysis produced sufficiently high modularity values for these networks to indicate a sufficient degree of cluster separation. A comparison of the links sent within and across the major clusters also supports the existence of distinct clusters. We also found hub users to be consistent indicators of selective exposure across the three data sets. Based on the limited data analyzed here, top-mentioned users and hashtags seem to be better indicators for selective exposure, as they also indicated consistency of exposure to content with references particularly to one side of the political spectrum. When examining websites that users posted hyperlinks to, two challenges emerged. First, the total numbers of domain names mentioned were often low. Second, many of the most frequently used sites were the popular user-generated content distribution platforms, such as YouTube. While an interesting finding in itself, its contribution to our selective exposure questions is more limited. Broadly, we did not use statistical analyses here for two main reasons. Whereas thousands of users formed the networks that we analyzed in this

TABLE 2
Hubs by Clusters¥§

Date	*Cluster**	*# of Users*	*# of unique links*	*Top hashtags*	*Top mentioned users*	*Top domains in hyperlinks*
1–26	1	188	1,904	#tcot (38), #teaparty (14), #cnndebate (9), #gop (8), #p2 (7)	@barackobama (11), @realdonaldtrump (7) @ronpaul (7), @blacksheeprpt (7)	typepad.com (10), ronpaul2012.com (9), google.com (8), youtube.com (7)
	2	175	1,036	#cnndebate (10), #Obama2012 (8), #OWS (6), #CNNdebate (6), #tcot (6)	@barackobama (10) @samyoungman (7) @yougottavote (5) @annybush (4) @ richardwolffedc (4)	barackobama.com (13), youtube.com (11) Twitter.com (9) Google.com (5), time.com (5)
1–27	1	312	559	#endslavery (28), #humantrafficking (28), #WHChat (8), #Obama (7)	@barackobama (76) @ijmhq (25) @ijmcampaigns (25) @iswanthehill (15)	ijm.org (30), barackobama.com (26) youtube.com (20) mediabistro.com (11)
	2	223	1,888	#tcot (17)	@jonahnro (30) @barackobama (13) @amandacarpenter (13) @heritage (9) @realdonaldtrump (8)	nationalreview.com (35) washingtonpost.com (15) commonamericanjournal.com (12) blogspot.com (9) google.com (8) herit.ag (8)
	3	180	1,059	#WHChat (31), #ObamaDerangement-Syndrome (15) #p2 (12)	@ericboehlert (37) @jonahnro (15) @markknoller (14) @childerbrant (12) @barackobama (9)	nationalreview.com (15), google.com (10) twitter.com (5) marketwatch.com (4) publicpolicypolling.com (4) usa.gov (4) ncpssm.org (4)
1–28	1	169	1,680	#tcot (54) #p2 (16) #tlot (13) #twisters 12)	@barackobama (13) @cspanwj (10) @jjauthor (8) @barry_o44 (7)	google.com (17) youtube.com (11) twitlonger.com (8) blogspot.com (5)

(*Continued*)

TABLE 2
(Continued)

Date	*Cluster**	*# of Users*	*# of unique links*	*Top hashtags*	*Top mentioned users*	*Top domains in hyperlinks*
				#ocra (11)	@nicksreport (7)	
	2	103	530	#wtf (6)	@barackobama (17)	youtube.com (9)
				#Obama (5)	@21stprincipal (13)	epi.org (6)
					@brian8473 (6)	nytimes.com (6)
					@march4teachers (6)	blogspot.com (5)
					@eacheronthemic (6)	barackobama.com (4)

¥ Top selective exposure indicators — hubs, mentions, hashtags, and domains in hyperlinks — were identified according to the drop of frequencies, when ordered in descending order, and are presented in that order in the table. Specific frequencies are reported in the Findings section. When no clear drop in frequencies appeared, we reported the top 5.

§ Frequencies in prentices

* Numbers were assigned to cluster for consistency with Table 1 and Graphs 4–6.

* In this cluster only @BarackObama emerged as a hub. All other users attracted very few followers.

study, our main unit of analysis is a network cluster. Consequently, the number of units—seven different clusters—is not large enough for a meaningful statistical analysis. Second, for many of the indicators, the descriptive analysis was clear-cut. Hubs in any given cluster were either all conservative or a mix of liberal and media. In other words, there was no indication of a mix of liberal and conservative hubs. That said, for the purposes of illustrating our SEC method, analysis of these data sets is sufficient. By applying this method to a larger number of topic-networks, which will generate a larger number of clusters, we will be able to test significant differences in exposure to politically oriented information sources on Twitter.

Strengths and Limitations of the SEC Method

The SEC method stands on two legs: the use of large volumes of log data and the application of network analysis. Large amounts of data about social media activities are increasingly available to researchers. NodeXL is not the only tool to collect data; however, its vital added value is capturing the relationships among the users who posted content on social media spaces—in this case, Twitter. Unlike surveys that capture personal-reported activities (i.e., Chaffee et al., 2001; Stroud, 2008), which can lead to a variety of potential validity issues, log data capture actual behavior and preferences in information selection. Experiments (e.g., Knobloch-Westerwick, 2012) capture individuals' actual activity, but for a small number of individuals, lead to low external validity. Network analysis identifies patterns of personal exposure of subgroups of users, and detects clusters—the social boundaries of information exposure and flow that users form when following, mentioning, and replying to one another on Twitter. These clusters are the unit of analysis of selective exposure, providing the indicators of selective exposure we discussed above. A cluster also captures users who are exposed indirectly to hubs, as users are closely interconnected to one another and often retweet messages. Expanding selective exposure beyond direct contact, then, is another novelty of this SEC method.

The proposed SEC method also has important limitations to consider. First, it identifies selective exposure but does not explain why it takes place. An experiment remains a better way to identify causal relationships. That said, future studies may improve our understanding of the conditions under which selective exposure occurs on Twitter and other social media via longitudinal topic-network analyses. Second, each data set captured a snapshot of interactions among roughly one thousand users who participated in each discussion. This poses a problem of external validity. The data are not a sample. The data set contains all users who participated in the discussion about a given topic and contributes to the method's external validity. However, snapshots of these crowds, like photographs, miss users who were active at different times. We address this concern at two levels. One indicator of selective exposure, hubs, are expected to take an active role in the discussion. Users who did not participate frequently enough to be captured in this data may not be as central to the conversation. We also propose capturing data sets across longer periods of time. In our case study, discussions were linked to a scheduled event, so we captured data for three sequential days around that event. To examine ongoing discussions, it may be necessary to sample messages for longer periods of time. Also, we did not examine the content of messages, exploring only their unique characteristics of Twitter messages, that is, hashtags, hyperlinks, and mentioned user names. Finally, an analysis of existing data logs does not allow us to interact with users to gain a more in-depth understanding of their motives, attitudes, and reasons for selecting information sources and conversation partners. A survey remains

a good methodology for these questions. That said, as NodeXL collects tweets as well, while this method analyzes only the connections, it would be a relatively simple matter to use other software (e.g., LIWC) to code different kinds of content and see whether that has any relationship to network structures.

Case Study: Selective Exposure in the State of the Union Twitter Discussion

Our study provides support for earlier concerns about a process of "echo chambers," where selective exposure to information sources leads to fragmented interactions and divided clusters in which people tune in to a narrow scope of news and views (van Alstyne & Brynjolfsson, 1996; Sunstein, 2006). The clusters created when Twitter users differentially link to people with shared viewpoints is an indication of selective exposure. Users can see messages from users in their own cluster more easily than from users in other clusters. In each of the three days that followed the State of the Union speech, we identified two clusters, which we interpret as "red" (conservative) and "blue" (liberal). In the "red" clusters, the major hubs were users who self-identified as conservatives and were not associated with major media organizations or other major institutions. The most frequently mentioned hashtags in these clusters had a conservative orientation (e.g., #tcot), and the most mentioned users self-identified as conservatives. Many of the most frequently used domain names in hyperlinks were associated with generic Internet platforms that allow users to share a wide range of content (e.g., YouTube). However, the other popular websites referenced in these tweets had a clear conservative political orientation (e.g., the National Review). The "blue" clusters were distinctly different, illustrating a group with exposure to a mix of liberal and mainstream media organizations, (e.g., CBS's Mark Knoller and Eric Boehlert from the progressive Media Matters for America watchdog). These observations support earlier studies suggesting that conservatives and Republicans showed greater tendency to seek opinion-reinforcing information than Democrats and liberals (Iyenger, Hahn, Krosnick, & Walker, 2008). The "blue" clusters, in contrast, contain a mix of hashtags, hyperlinks, and top-mentioned usernames in terms of their political leaning, mixing left-wing and mainstream media. The "red" clusters, in contrast, showed more right-wing homogeneity. One explanation could be a lack of trust of traditional news outlets by conservative users.

Network analysis of Twitter discussions shows that while clusters were distinct, they were not completely isolated from one another. Modularity levels were in the medium range, and while percentages of links across clusters were low, they were not zero. This suggests that while users are primarily exposed to content shared by their cluster mates, there is a limited exposure to content from outside a cluster. In other words, while many users do not actively expose themselves to diverse sources of information, they may be indirectly exposed to content from diverse sources. This is encouraging, as earlier studies suggested that users' selective avoidance is not as prominent as selective exposure (Garrett, 2011; Garrett et al., 2011; Parmelee & Bichard, 2012).

CONCLUSION

We have provided a step-by-step method for detecting patterns of selective exposure using network analysis of social media data. The authors hope that researchers will use the proposed SEC method to complement existing methods to better explore selective exposure in their areas of

interest. For example, in political communication, researchers can continue to explore selective exposure to like-minded individuals and hubs. In health communication, exploring patterns of selective exposure to accurate versus inaccurate information can be valuable. In international communication, researchers can examine the selective exposure to news via Twitter based on countries of origin using the SEC method. This SEC method has advantages and weaknesses compared with other methods used in the field. We hope researchers can use it to expand our understanding of selective exposure in a wide range of areas of communication.

ACKNOWLEDGEMENTS

The authors wish to thank the *Pew Internet and American Life Project* and its director Lee Rainie for the support of this study.

REFERENCES

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. Retrieved from http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf

Arendt, H. (1968). Truth and politics. In H. Arendt (Ed.), *Between past and future: Eight exercises in political thought*. New York, NY: Viking Press.

Babbie, E. R. (2012). *The practice of social research*. Belmont, CA: Wadsworth.

Calhoun, C. (1988, Fall). Populist politics, communication media and large scale societal integration. *Sociological Theory*, *6*, 219–241. doi: 10.1111/j.1475-682X.1998.tb00474.x

Carrington, P. J., Scott, J., & Wasserman, S. (2005). *Models and methods in social network analysis*. New York, NY: Cambridge University Press.

Chaffee, S. H., Saphir, M. S., Graf, J., Sandvig, C., & Hahn, K. S. (2001). Attention to counter-attitudinal messages in a state election campaign. *Political Communication*, 18(3), 247–272. doi: 10.1080/10584600152400338

Choi, S., Park, J., & Park, H. (2011, February). *Twitter, a medium for social mobilizing?: An exploratory study on the use of twitaddons.com in South Korea*. Paper presented at the International Network for Social Network Analysis. St. Petersburg, FL.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*, 066111. doi: 10.1103/PhysRevE.70.066111

Corrado, A., & Firestone, C. M. (Eds.). (1996). *Elections in cyberspace: Toward a new era in American politics*. Washington, DC: Aspen Institute.

Durrheim, K. (2007). Research design. In M. T. Blanche, K. Durrheim, & D. Painter (Eds.), *Research in practice: Applied methods for the social sciences* (pp. 33–59). Cape Town, South Africa: University of Cape Town Press.

Eysenbach, G. (2008). Credibility of health information and digital media: New perspectives and implications for youth. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 123–154). The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: MIT Press.

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication*, *14*, 265–285. doi: 10.1111/j.1083-6101.2009.01440.x

Garrett, R. K., Carnahan, D., & Lynch, E. K. (2011). A turn toward avoidance? Selective exposure to online political information. *Political Behavior*, 35(1), 113–134. doi: 10.1007/s11109-011-9185-6

Gergen, K. J. (2008). Mobile communication and the transformation of the democratic process. In J. Katz (Ed.), *Handbook of mobile communication studies* (pp. 297–310). Cambridge, MA: MIT Press.

Gil de Zúñiga, H., Correa, T., & Valenzuela, S. (2012). Selective exposure to cable news and immigration in the U.S.: The relationship between FOX News, CNN, and attitudes toward Mexican immigrants. *Journal of Broadcasting & Electronic Media*, *56*(4): 597–615. doi: 10.1080/08838151.2012.732138

Gitlin, T. (1998). Public sphere or public sphericules? In T. Liebes & J. Curran (Eds.), *Media, ritual and identity* (pp. 168–174). London, UK: Routledge.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, *81*, 1287–1303. doi: 10.2307/2776392

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

Han J. Y. (2008). *Examining effective use of an interactive health communication system (IHCS)*. (Unpublished doctoral dissertation). University of Wisconsin–Madison.

Han, J. Y., Wise, M., Kim, E., Pingree, R., Hawkins, R., Pingree, S., . . . Gustafson, D. H. (2010). Factors associated with use of interactive cancer communication system: An application of the comprehensive model of information seeking (CMIS). *Journal of Computer-Mediated Communication*, *15*, 367–388. doi: 10.1111/j.1083-6101.2010.01508.x

Hansen, D. L., Shneiderman, B., & Smith, M. A. (2011). *Analyzing social media networks with NodeXL: Insights from a connected world*. Burlington, MA: Morgan Kaufmann.

Hargittai, E., Gallo, J., & Kane, M. (2008). Cross-ideological discussions among conservative and liberal bloggers. *Public Choice*, *134*(1–2), 67–86. doi: 10.1007/s11127-007-9201-x

Hauben, M., & Hauben, R. (1997). *Netcitizen*. London, UK: Wiley.

Hwang, Y. (2010). Selective exposure and selective perception of anti-tobacco campaign messages: The impacts of campaign exposure on selective perception. *Health Communication*, *25*(2): 182–190. doi: 10.1080/10410230903474027

Himelboim, I., McCreery, S., & Smith, M. (2011, May). *Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter*. Paper presented at the International Communication Association Annual Conference, Boston, MA.

Holbert, R. L., Garrett, R. K., & Gleason, L. S. (2010). A new era of minimal effects? A response to Bennett and Iyengar. *Journal of Communication*, *60*(1): 15–34. doi: 10.1111/j.1460-2466.2009.01470.x

Huberman, B. A., Romero, D. M., & Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday, 14*(1–5) [online]. Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063

Iyenger, S., Hahn, K. S., Krosnick, J. A., & Walker, J. (2008). Selective exposure to campaign communication: The role of anticipated agreement and issue public membership. *The Journal of Politics*, *70*(1): 186–200. doi: 10.2307/30218868

Kelly, J. D., Fisher, D., &. Smith, M. (2005). *Debate, division, and diversity: Political discourse networks in Usenet newsgroups*. Retrieved from http://www.coi.columbia.edu/pdf/kelly_fisher_smith_ddd.pdf

Kelly, W., Fisher, D., & Smith, M. (2006). Friends, foes, and fringe: Norms and structure in political discussion networks. *ACM Conference Proceeding Series*, *151*, 412–417.

Krebs, V. (2004). *The social life of books: Visualizing communities of interest via purchase patterns on the WWW*. Retrieved from http://orgnet.com/booknet.html

Knobloch-Westerwick, S., & Meng, J. (2011). Reinforcement of the political self through selective exposure to political messages. *Journal of Communication*, *61*(2): 349–368. doi: 10.1111/j.1460-2466.2011.01543.x

Knobloch-Westerwick, S. (2012). Selective exposure and reinforcement of attitudes and partisanship before a presidential election. *Journal of Communication*, *62*(4): 628–642. doi: 10.1111/j.1460-2466.2012.01651.x

McGeough, R. E. (2010, November). The market AS the forum: Amazon.com discussion forums as deliberative spaces. Paper presented at the Annual Convention of the National Communication Association, San Francisco, CA.

McKenna, K. Y. A., & Bargh, J. A. (2000). Plan 9 from cyberspace: The implications of the Internet for personality and social psychology. *Personality and Social Psychology Review*, *4*, 57–75. doi: 10.1207/S15327957PSPR0401_6

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–445. doi: 10.2307/2678628

Mill, J. S. (1956). *On liberty*. Indianapolis, IN: Bobbs–Merrill.

Newman, M. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, 64, 025102. doi: 10.1103/PhysRevE.64.025102

Newman, M. (2004). Detecting community structure in networks. *The European Physical Journal B*, *38*(2): 321–330. doi: 10.1140/epjb/e2004-00124-y

Parmelee, J. H., & Bichard, S. L. (2012). *Politics and the Twitter revolution: How tweets influence the relationship between political leaders and the public*. Lanham, MD: Lexington Books.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London, UK: Penguin.

Raban, D. R., & Rabin, E. (2007). The power of assuming normality. Paper presented at *EMCIS*, Valencia, Spain.

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier*. Reading, MA: Addison-Wesley.

Robertson, S. P., Vatrapu, R. K., & Medina, R. (2009). The social life of social networks: Facebook linkage patterns in the 2008 U.S. presidential election. *Proceedings of the 10th Annual International Conference on Digital Government Research*. Puebla, Mexico.

Rubin, R. B., Palmgreen, P., & Sypher, H. E. (1994). *Communication research measures: A sourcebook*. New York, NY: Guilford Press.

Shapiro, A. L. (1999). *The control revolution: How the Internet is putting individuals in charge and changing the world we know*. New York, NY: Public Affairs.

Stromer-Galley, J. (2003). Diversity of political conversation on the Internet: Users' perspectives. *Journal of Computer-Mediated Communication*, *3*(8). doi: 10.1111/j.1083-6101.2003.tb00215.x

Stroud, N. J. (2008). Media use and political predispositions: Revising the concept of selective exposure. *Political Behavior*, *30*(3): 341–366. doi: 10.2307/40213321

Sunstein, C. (2006). *Republic 2.0*. Princeton, NJ: Princeton University Press.

Turow, J. (1997). *Breaking up America: Advertisers and the new media world*. Chicago, IL: The University of Chicago Press.

van Alstyne, M., & Brynjolfsson, E. (1996). Electronic communities: Global village or cyberbalkans? In J. DeGross, S. Jarvenpaa, & A. Srinivasan (Eds.), *ICIS 1996* (pp. 80–98). doi: 10.2307/20110380

Wang, Y. (2012, March). Forcing a breakdown: *Establishing the limits of community detection algorithms*. Paper presented at the International Network for Social Network Analysis Sunbelt, 32nd Annual Conference, Redondo Beach, CA.

Warner, B. R. (2010). Segmenting the electorate: The effects of exposure to political extremism online. *Communication Studies*, 61(4), 430–444.

Wasserman, S., & Faust, K. (1999). *Social network analysis: Methods and applications*. New York, NY: Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*, 440–442. doi: 10.1038/30918

Wojcieszak, M. E., & Mutz, D. (2009). Online groups and political discourse. *Journal of Communication*, *59*, 40–56. doi: 10.1111/j.1460-2466.2008.01403.x