Examining a Twitter-Based Discourse Community of Composition Scholars

Brian Larson
PhD Student, Rhetoric & Scientific & Technical Communication
University of Minnesota
@Rhetoricked / www.Rhetoricked.com

Introduction

This paper considers the possibility of a theoretically motivated empirical means for detecting and delineating a discourse community for purposes of studying writing practices on Twitter. It is, in a sense, a reaction to some research I've seen splashed up on screens, mostly at conferences, with lots of pretty network graph visualizations like Figure 1.

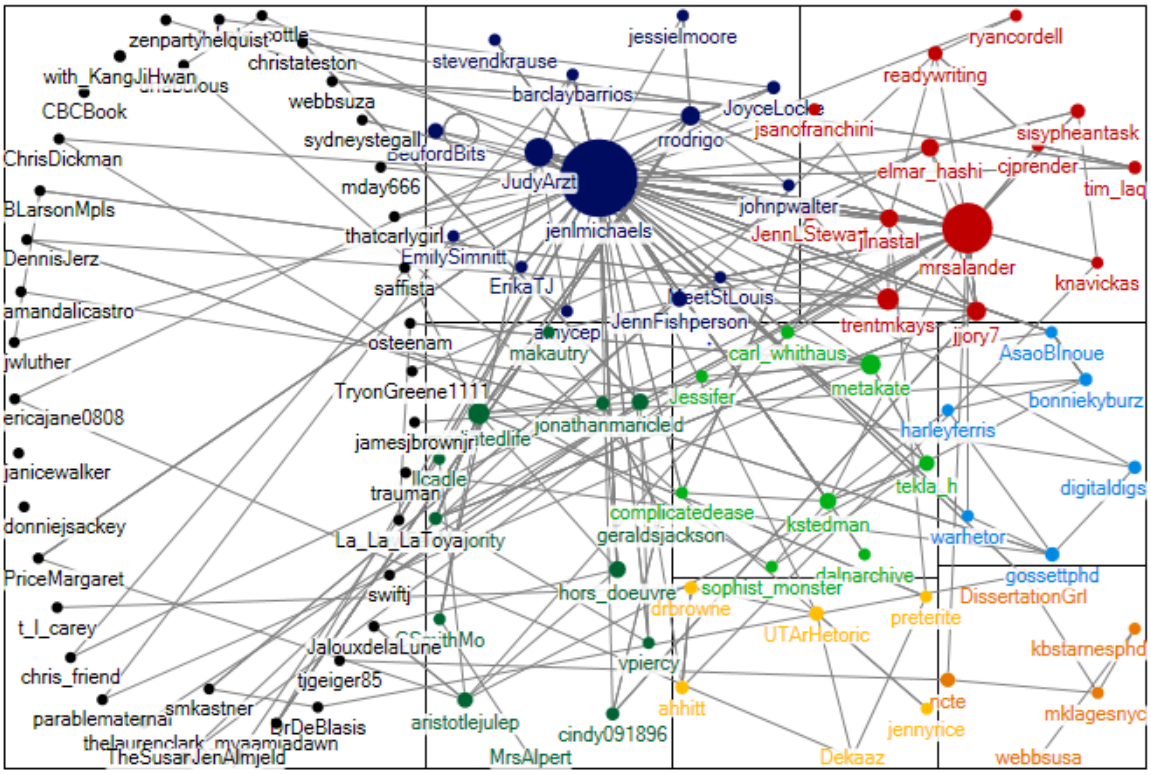

Figure 1 Candidate communities identified among users using CCCCs hashtag

This paper makes a first attempt at linking such graph visualizations to a theoretical construct of interest in writing studies: the discourse community or community of practice. But it also suggests

that making sense of these types of graphs will require richer qualitative research methods than have sometimes been proffered in the context of these visualizations.

The concept of a discourse community or community of practice is central to several important theories in writing studies; and the ability to select for study some meaningful subset of the millions of 'tweets' appearing on Twitter each day is essential for almost any empirical study of writing practices on Twitter. The concepts of "community" and its analogs play a vital role in theories and methods relating to writing studies. In genre theory, for example, Swales (1990) posits genres as "properties of discourse communities," which are "sociorhetorical networks that form in order to work towards sets of common goals" (p. 9). Berkenkotter and Huckin (1994) echo Swales when they describe "community ownership" as a theoretical principle undergirding genre. Though Russell (1997), describing activity system theory, does not refer expressly to discourse communities or communities of practice, he implies the need for the boundedness (or at least boundedness-for-now) of "subject," which he defines as "agent(s) whose behavior... the analyst is focusing on" (p. 510).

Empirical research in writing from which the researcher intends to generalize implicates concepts from statistics: "populations" and "samples" (MacNealy 1998). If a researcher wishes to generalize about a group of people, a population—let's say, "All folks who do x"—she must first define the population. If the population is too large to be studied *in toto*, she may select a random or representative sample of the population's members, and then generalize her results to the population, subject to certain limitations. A population described as "all folks who do x" may be a superset of communities, each of which consists of folks who do some "x," but who have other things in common, as we shall see.

Twitter poses challenges for the researcher seeking to define a community, activity system, or population for study. Twitter is an Internet micro-blogging service that allows a user to post short (140-character) messages visible to other users who follow her (Myers 2010). The messages may include links to other content on the web. Among the problems for researchers seeking to study writing practices on Twitter is the number of active Twitter subscribers, estimated at *more than 200 million* in December 2012 ("Twitter active users," 2012) and the number of the messages they post—or 'tweets'—*more than half a billion a day (that's about 5500 per second)* by October 2012 ("Report: Twitter hits," 2012).

Generalizing about or describing the writing practices of all Twitter users is thus a bit like generalizing or describing the activities of all the world's people; it's probably either impractical or invalid. Researchers are accustomed to describing the writing practices of smaller populations with more defined boundary conditions. For example, researchers might study writing practices among adults within a portion of a state; of students within a single class; of employees working in a workplace; of researchers in a particular discipline; or even of a closed but more diffuse network like a Usenet group. Usually, the bounding conditions of these communities or groups consist of the geography, discipline, physical environment, or physical or virtual forums (or some combination of these) in which they function as writers. These bounding conditions are difficult, if not impossible, to ascertain for a community or population of Twitter users.

This paper considers the possibility of a theoretically motivated empirical means for detecting and delineating a discourse community for purposes of studying writing practices on Twitter. It explores concepts of community within writing and genre studies; it considers useful variables for analysis and offers very preliminary ideas about operationalizing those variables. It presents a data set collected in spring 2012 for preliminary analysis, and one simple visualization of

it in network graph terms. Finally, it proposes that additional qualitative data are necessary to make meaning from network graphs and their visualizations; I'll recommend some next steps for exploring the techniques and frameworks presented here.

Communities and the variables for describing them

This section discusses conceptions of community for writing studies, particularly genre theory, and from a sociological study of Twitter. It identifies variables that may be useful for identifying candidate groupings of Twitter users and describing the extent to which they might be characterized as "communities."

Swales (1990) describes six characteristics that he says define a discourse community: "a broadly agreed set of common public goals"; a means for members to communicate with each other; a focus on providing "information and feedback" within the group; genres that it uses "in the communicative furtherance of its aims"; a common vocabulary or lexis; and a "threshold level of members with a suitable degree of relevant content and discoursal expertise" (p. 24-27). Berkenkotter and Huckin refer somewhat approvingly to this conception of discourse community, but they warn that "asserting a relationship between the concept of genre and that of 'discourse community' is a slippery proposition because neither concept refers to a static entity" (p. 21). In any event, Swales' situates "discourse community" within a disciplinary community of practice, as he is studying genres of academic discourse, especially the "research article." Twitter users, on the other hand, constitute a cross section of a portion of humanity, communicating about a wide variety of activities.

The definition of community in this broader sense is contested and evolving. According to Gruzd et al. (2011), it can be "a set of people who share sociability, support, and a sense of identity" (p. 1295); or "a spatially compact set of people with a high frequency of interaction,

interconnections, and a sense of solidarity" (p. 1296); or in a nod to the Internet, the same definition less the requirement for spatial compactness. Gruzd and his colleagues explored the Twitter interactions of one of them, Wellman, to determine whether it was possible to characterize his network of followers, "sources" (folks he followed but who did not follow him), and "mutuals" (folks he followed and was followed by) to detect a community (p. 1296). They looked for evidence of characteristics urged by three different conceptions of community. The first, drawn from Anderson's *Imagined Communities* (1983), called for a common language; "temporality" or "the presence of 'homogeneous' time, in which a community is 'moving' through history together by sharing 'a consciousness of a shared temporal dimension in which they co-exist'" (p. 1303); and the decline in prominence of "high centers," entities that "'society is naturally organized around and under'" (p. 1303). The second centered on Jones' (1997) conception of the "virtual settlement," typified by interactivity among members, a variety of communicators, a "common public place where members can meet and interact"; and "sustained membership over time" (p. 1307). Supplementing these conceptions was a third from McMillan and Chavis (1986), which required a "sense of community." The sense of community arises when putative members of a putative community feel that they are members of the community, members have influence within the community, the community meets some member needs, and members share an emotional connection.

Based on these concepts, when examining Twitter data for evidence of communities and their characteristics, I have considered operationalizing the variables in Table 1. Because of the sheer volume of tweets and twitterati, however, I believe it's necessary to prioritize the empirical study, using the "Candidate" variables to identify candidate communities, then using the

"Descriptive" variables to characterize them more fully. I've also identified some variables that I assume or argue are not necessary or useful as "Excluded."

Table 1 Variables considered for detecting and describing Twitter communities

Variable	Type	Description
Temporality	Candidate	"[T]he presence of 'homogeneous' time, in which a community is 'moving' through history together by sharing 'a consciousness of a shared temporal dimension in which they co-exist'" (Gruzd et al., p. 1303; Anderson)
Interactivity among members	Candidate	The frequency with which, and extent to which, members of a candidate community interact (Gruzd et al.; Jones)
Variety of communicators	Candidate	I interpret this as requiring communication to be initiated by a substantial percentage of the members of a candidate community (Gruzd et al.; Jones). It perhaps addresses Swales' call for a threshold level of members.
Common interest	Candidate	Though Swales offered "common public goals," I think the broader conception of "common interest" makes sense in Twitter.
Common language	Descriptive	Language practices that distinguish the candidate community from the rest of Twitter (Gruzd et al.; Anderson). Also addresses Swales "common lexis."
Membership feelings	Descriptive	Extent to which putative members of a candidate community feel that they are members of a community (Gruzd et al.; McMillan & Chavis)
Member influence	Descriptive	Extent to which members have influence within the candidate community (Gruzd et al.; McMillan & Chavis). This partially satisfies Swales' call for "information and feedback," as well.
Utility	Descriptive	Extent to which the candidate community meets some member needs (Gruzd et al.; McMillan & Chavis). This partially satisfies Swales' call for "information and feedback," as well.
Emotional connection	Descriptive	Extent to which members of a candidate community share emotional connection (Gruzd et al.; McMillan & Chavis)
High centers	Excluded	The entities that "'society is naturally organized around and under'" (Gruzd et al., p. 1303; Anderson)
Common place	Excluded	I assume this for Twitter, as Twitter itself functions as the common forum for the communities that exist in it (Gruzd et al.; Jones)
Sustained membership	Excluded	Gruzd et al. evaluated this by looking at Wellman's Twitter network at both ends of a six-month period (Gruzd et al.; Jones). I *assume* it with regard to the study of a temporally localized event, such as the CCCCs.

Gruzd et al. assumed temporality in Twitter because of its focus on current events. I'm uncomfortable with the notion that a mere focus on current events is sufficient to give rise to a shared consciousness of a shared temporal dimension. I think some other means of identifying a shared consciousness is warranted.

Having identified variables that may prove useful for identifying and characterizing candidate communities, I now need to propose some ways of operationalizing them. For this project, I'll focus on the Candidate variables and two of the Descriptive variables, Member Interest and Utility. I propose to use network analysis tools taking advantage of social network theory.

Network analysis tools and social network theory analysis

This section provides an overview of the use of network graphs to represent relations among participants in a social activity. It then considers some of the ways that graph concepts might be used to operationalize some of the variables discussed in the previous section

One way to visualize a social network is with a sociogram or network graph (Hansen et al. 2010). "Social scientists commonly use graph theory and network concepts to operationalize theoretical statements about structural regularities in social systems" (Holland and Leinhardt 1976, p.1). In the simplest form of a graph, "nodes" or "vertices" represent individuals, and "edges" or "arcs" represent relationships between the individuals. In visual representations, a vertex is represented as a point on the graph, and an edge is a line between vertices. Figure 2 represents a simple network involving six people.

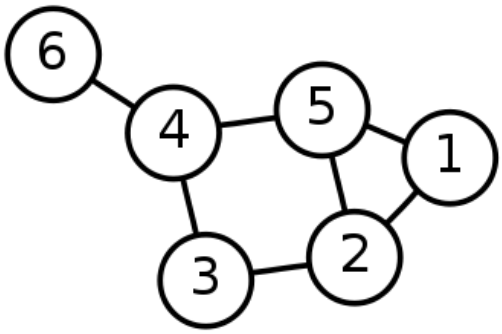

Figure 2 Example of network graph. Source: Wikimedia commons.

In Figure 2, the circles numbered 1 through 6 are the nodes or vertices of the graph. The lines between them are the edges of the graph, representing the relations among the people. We'll assume for this example that the relationship represented is friendship and that friendship is always reciprocated. There are thus no edge lines with arrows (called "arcs") representing one-way relationships. In this example, Person 6 is friends only with Person 4; Person 4 is friends with Persons 6, 5, and 3; etc. Note that the drawing in Figure 2 is not the graph, but simply a visualization of the graph. This particular graph could be described just as accurately as a set of unordered pairs, where each pair represents two friends: {(1,2), (1,5), (2,3), (2,5), (3,4), (4,5), (4,6)}. A graph where each node shares an edge with each other node is called a "complete" graph. Figure 2 is not a complete graph because, for example, Person 6 is not connected to any person other than Person 4. People can be described in terms of the "geodesic distance," the shortest path or the smallest number of steps to traverse, between them. Here, the geodesic distance between Person 1 and Person 6 is 3; between Person 1 and Person 2, it is 1. Graphs and nodes have characteristics that are useful for describing networks and the roles of their members. We'll explore a few of them that may be useful for this project.

We can discuss graphs in terms of their density, diameter, and number of connected components. Density is a ratio of the total number of edges observed among the nodes to the total number possible (in a complete graph). In the example of Figure 2, there are 7 edges out of a

maximum possible of 15 edges; the density is thus 7/15 or approximately 0.467. A graph's diameter is the longest observed geodesic distance in the graph; in other words, it is the longest shortest distance between two nodes. In Figure 2, the diameter is 3, the distance between Person 1 and Person 6. Connected components are "clusters of vertices that are connected to each other but separate from other vertices in the graph" (Hansen et al. 2010, 5.3.3).

Nodes are often characterized according to their centrality in the network, and according to at least three different measures: degree, closeness centrality, and betweenness centrality. Degree is simply the number of edges emanating from a node. In Figure 2, the degree of Person 5 is 3 (connected to Persons 1, 2, and 4); the degree of Person 6 is 1 (connected to Person 4). Degree is a measure of the quantity of a node's connections, but not their quality (Hansen et al. 2010, 3.5.2). Closeness centrality is the average distance between the vertex and every other vertex in the graph; lower numbers represent greater centrality. For example, in Figure 2, Person 1's closeness centrality is 1.8, Person 3's is 1.6, and Person 5's is 1.4. Betweenness centrality is a measure of how often a vertex lies on the shortest route between two other vertices. So, in Figure 2, Person 1 has a betweenness centrality of 0, as no other person needs to connect through Person 1; Person 5 has a betweenness centrality of 2, because she lies on the shortest route between Person 1 and Persons 4 and 6; Person 4 has the highest betweenness centrality at 4, as she lies on the only route between Person 6 and the other four people.

Another potentially important measure of a node's potential position within a network or community is its clustering coefficient. This is a measure of the density of a sub-network consisting of the person's connections (Hansen 2010, 3.5.2). Thus, if a person's friends are all friends of each other, she has a high clustering coefficient. In the example in Figure 2, Person 1 has a clustering coefficient of 1, because her two friends, Persons 2 and 5, are friends of each other. Person 3 has

two friends who are not friends of each other, and thus has a clustering coefficient of 0. Person 5 falls in the middle: of the three possible relations between Persons 1, 2, and 4, only one exists, making Person 5's clustering coefficient 0.33.

Based on these possible network metrics, I propose to operationalize the variables described above using the measures set out in Table 2.

Table 2 Candidate bases for operationalizing variables

Variable	Operationalized
Temporality	A sample bounded by time, with a beginning and end date.
Common interest	A common Twitter hashtag.
Interactivity among members	Density of edges representing @-replies and retweets among candidate group members; measured by clustering coefficient of the candidate community compared to randomly generated clustering coefficient and compared to edges between members and non-members.
Member influence	Density of edges representing @-replies and retweets among candidate group members.

Ultimately, this selection of approaches to operationalizing variables will need to be justified and validated based on some empirical evidence (more on that below). But first, I'd like to introduce a data set that may support such an analysis.

The CCCC 2012 Twitter data set

Between March 9 and March 23, 2012, Jen Michaels, a graduate student at Ohio State University, captured an archive of tweets using Twitter hashtags applicable to the Conference on College Composition and Communication Conference 2012 (the CCCCs) using the Twitter Archiving Google Spreadsheet (Hawksey, n.d.). The resulting archive consisted of more than 5000 tweets by nearly 600 different Twitter subscribers. I will not refer to these subscribers as "people," since some of the accounts appear to be, institutional, including "CengageEnglish" and "ncte."

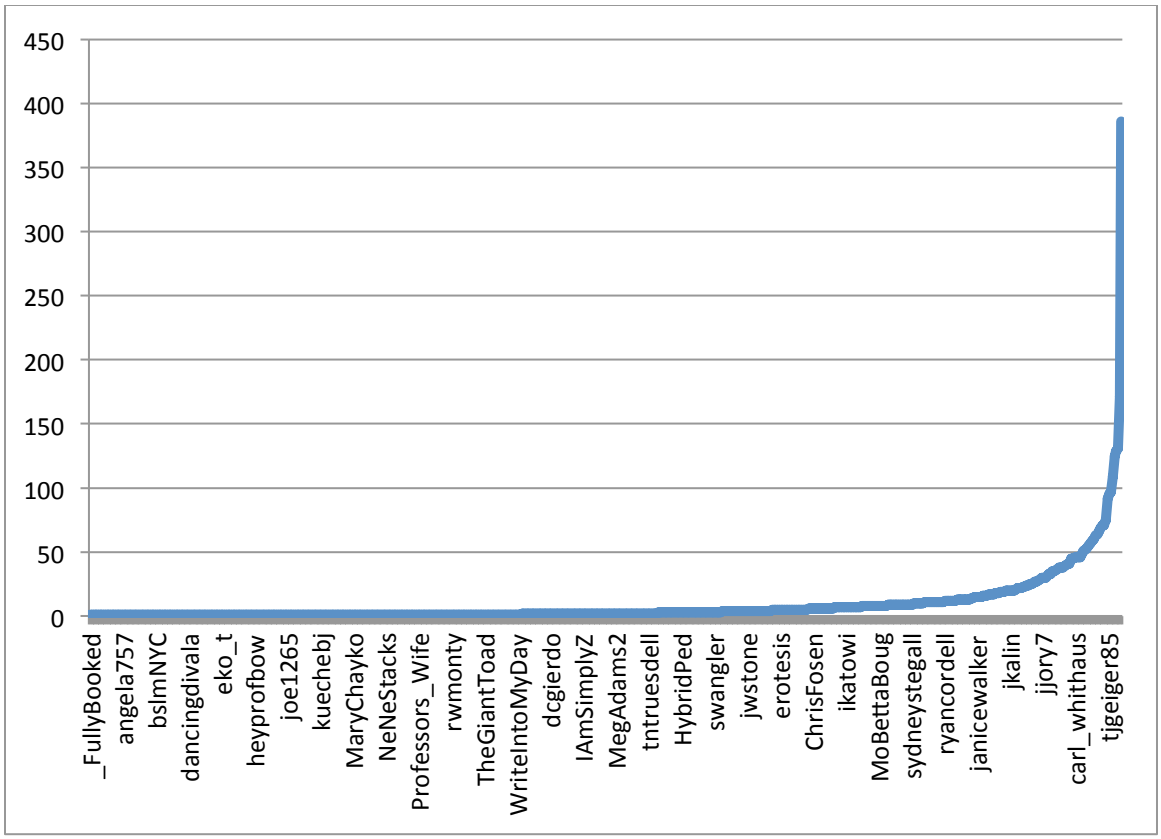

Figure 3 Distribution of tweets among subscribers using CCCCs hashtags

The distribution of tweets among the accounts, shown in Figure 3, was extremely skewed: in fact, of approximately 600 subscribers, only 115 or so tweeted more than 10 times using the hashtags, and fewer than 60 tweeted more than 25 times using them. High-volume tweeters included six accounts that tweeted 100 or more times.

To begin to visualize any and characterize candidate communities in this data set, I used NodeXL (Hansen et al. 2010) to generate a graph, considering only subscribers who had received retweets or @-replies. After NodeXL automatically generated clustering coefficients for the nodes, it generated a network diagram showing candidate communities within the broader hashtag community (see Figure 1).

Figure 1 shows the results of NodeXL's efforts to automatically classify the Twitter accounts into groups or candidate communities. The rectangles drawn around subsets of nodes

represent NodeXL's best guess as to the boundaries of each candidate community. In this case, the size of each node represents its degree, the number of times this account sent or received @-replies or retweets. It's clear that two members of the CCCCs hashtag 'community,' *jenlmichaels* and *mrsalander,* have high degree ratings and also have connections broadly across all candidate communities, not just within their 'home' groupings. Within the candidate groups, members appear to have relatively few connections to each other, so it's possible that the relevant clustering coefficients, though significant enough to cause NodeXL to group the nodes together, might not represent a real sense of membership in any of these groups. This leads to the next section, and its discussion of next steps.

Next steps

For this work to lead to a valuable contribution to the discipline, I believe that systematic qualitative examination of the Twitter account holders and their accounts is necessary to see if divisions into communities motivated by graph theory are borne out according to the other variables identified above.

Of course, systematic exploration of the data makes considerable sense. Looking at the ways in which different network graph metrics tend to represent the relations among the network nodes should permit the researcher to enumerate a variety of possible approaches for detecting and characterizing communities within Twitter. This probably entails spending time analyzing and visualizing the CCCCs data, focusing in turn on the various measures of centrality and clustering. Another key consideration is thresholds of involvement: To what extent should Twitter accounts that tweet only once be included in the analysis? How about Twitter accounts that tweet at a disproportionately high level (such as *jenlmichaels* and *mrsalander*)? But systematic examination of Tweets and follower-followed relationships in a Twitter archive alone is not a sufficient basis for

making many, or perhaps any, theoretical claims in writing studies. It will be important to validate these findings by exploring some of the other community variables discussed above.

Understanding the extent to which the systematic, and theoretical, approach results in useful representations of candidate communities requires study of the Twitter and account holders themselves. How do the account holders characterize themselves in their Twitter profiles and in websites to which the profiles link? To what extent do members of candidate communities feel or believe that the candidate theoretical communities are real communities? To what extent to those who retweet and @-reply to each other feel that those actions are constitutive of a community among them? Answering these questions may require reaching out to the Twitter account holders with surveys or interviews (subject to IRB approval, of course).

Based on these steps, it may be possible to revise the proposed means for operationalizing the variables; the result could be a vocabulary and approach for describing the methods of studying communities of writers on Twitter that other researchers can use and enter into dialog with. Unfortunately, this kind of data collection is hindered by its many challenges, including IRB approvals, copyright and terms-of-use concerns, and willingness of persons unknown to the researcher to take part in qualitative study. Sorting through these issues will require time and energy, and I'm interested if others wish to collaborate.

Conclusion

I hope that I've illustrated some of the ways that we might consider operationalizing notions of "community" within the vast user-base of Twitter using social network theory while also identifying some of the additional qualitative research that would be necessary to make claims about the validity of this kind of research. Pretty pictures of graph visualizations do not by themselves

constitute new knowledge about Twitter user communities, but taken together with the proper next steps, they may be useful for writing research in that social media context.

Works cited

Anderson, B. (1983). *Imagined communities: Reflections on the origin and spread of nationalism.* Verso.

Berkenkotter, C., & Huckin, T. N. (1994). *Genre Knowledge in Disciplinary Communication: Cognition/culture/power* (pp. 1-25). Routledge.

Gruzd, A., Wellman, B., & Takhteyev, Y. (2011). Imagining Twitter as an imagined community. *American Behavioral Scientist, 55*(10), 1294 –1318.

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world* (1st ed.). Morgan Kaufmann. [Kindle edition.]

Holland, P. W., & Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology, 7*, 1–45.

Jones, Q. (1997). Virtual communities, virtual settlements and cyber-archaeology. *Journal of Computer Mediated Communication, 3*(3), n.p.

MacNealy, M. S. (1998). *Strategies for Empirical Research in Writing.* Longman.

McMillan, D. W., & Chavis, D. M. (1986). Sense of community: A definition and theory. *Journal of Community Psychology, 14*(1), 6-23.

Myers, G. (2010). *The Discourse of Blogs and Wikis.* Continuum Discourse Series. London: Continuum International Publishing Group.

Report: Twitter hits half a billion tweets a day. (n.d.). *CNET.* Retrieved March 11, 2013, from http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/

Russell, D. R. (1997). Rethinking genre in school and society: an activity theory analysis. *Written Communication, 14*(4), 504–554.

Swales, J. M. (1990). *Genre Analysis: English in academic and research settings.* Cambridge Applied Linguistics. Cambridge: Cambridge University Press.

Hawksey, M. n.d. *TAGS: Twitter Archiving Google Spreadsheet.* http://mashe.hawksey.info/twitter-archive-tagsv3/

Twitter active users pass 200 million. (2012, December 18). *The Guardian.* Retrieved March 11, 2013, from http://www.guardian.co.uk/technology/2012/dec/18/twitter-users-pass-200-million