

# Network Characteristics and the Value of Collaborative User-Generated Content

Sam Ransbotham, Gerald C. Kane

Carroll School of Management, Boston College, Chestnut Hill, MA 02467, sam.ransbotham@bc.edu gerald.kane@bc.edu

Nicholas H. Lurie

School of Business, University of Connecticut, Storrs, CT 06269, lurie@uconn.edu

User-generated content is increasingly created through the collaborative efforts of multiple individuals. In this article, we argue that the value of collaborative user-generated content is a function both of the direct efforts of its contributors and its embeddedness in the content-contributor network that creates it. An analysis of Wikipedia’s Medicine Wikiproject reveals a curvilinear relationship between the number of distinct contributors to user-generated content and viewership. A two-mode social network analysis demonstrates that the embeddedness of the content in the content-contributor network is positively related to viewership. Specifically, locally central content—characterized by greater intensity of work by contributors to multiple content sources—is associated with increased viewership. Globally central content—characterized by shorter paths to the other collaborative content in the overall network—also generates greater viewership. However, within these overall effects, there is considerable heterogeneity in how network characteristics relate to viewership. In addition, network effects are stronger for newer collaborative user-generated content. These findings have implications for fostering collaborative user-generated content.

forthcoming in *Marketing Science*

*Key words:* user-generated content; information value; wiki; social network analysis

---

## 1. Introduction

There is considerable interest in the value of user-generated content and its antecedents. Research shows that product reviews, for instance, influence consumer search and product choice, enhance sales forecast quality, affect product sales, and drive viewership (Chevalier and Mayzlin 2006, Godes and Mayzlin 2004, Li and Hitt 2008). Other research shows that the relative influence of user-generated content depends on characteristics of the content, the creators of content, and their interactions (Berger and Milkman 2011, Constant et al. 1996, Weiss et al. 2008). Longer and two-sided reviews have greater influence on attitudes and behavior than shorter one-sided reviews (Schlosser 2007, Weiss et al. 2008), the valence of product ratings affects consumer choice (Duan et al. 2008, Godes and Mayzlin 2004), and negative and high variance early reviews can cause later reviewers to adjust their own ratings downwards (Moe and Trusov 2011, Schlosser 2005). In addition, the perceived similarity of creators to receivers, their behavior in response to requests for content, and their perceived expertise all affect the value of user-generated content (Forman et al. 2008, Weiss et al. 2008).

Although individuals create most user-generated content, an increasing amount emerges from groups of people working collectively. Examples include the wiki Web sites Wikia and Wikipedia, where contributors work together on articles; virtual worlds such as World of Warcraft, where participants create shared spaces and perform shared tasks; open-source software projects such as Sourceforge and Linux where volunteers work together to create marketable products; and citizen journalism Web sites like CNN's iReport, where amateur reporters create content that drives advertising viewership. These examples all involve collaborative user-generated content, which differs from individually created content through concurrent editing of the same content, the need to reach consensus about what to include and exclude, and final output that often varies substantially from the original contributions made by individuals. Despite its growing importance, there is limited research on the value and antecedents of collaborative user-generated content.

In this article, we argue that the value of collaborative user-generated content is a function both of the direct efforts of its contributors and its embeddedness in the content-contributor network that creates it. To gain insights into the content-contributor network, we conduct a two-mode social network analysis (SNA), an insightful approach for studying collaborative environments (Wasserman and Faust 1994, Grewal et al. 2006). Typical SNA applications in marketing study a single mode of interactions, such as consumers interacting with other consumers (Frenzen and Nakamoto 1993, Frenzen and Davis 1990) or companies interacting with other companies (Iacobucci and Hopkins 1992). Prior research has examined network relationships among consumers (Brown and

Reingen 1987, Frenzen and Davis 1990, Manchanda et al. 2008) and among customers, producers, and collaborators in business-to-business settings (Frels et al. 2003, Rindfleisch and Moorman 2001). Researchers have examined how social networks might promote user-generated content (e.g., through voting or by providing links to this content; Elsner et al. 2009, Oestreicher-Singer et al. 2009).

However, SNA can also be used to study networks with two distinct types of nodes, also known as a two-mode network. Two-mode SNA has been used to study project teams and shared members, actors and the films they have worked on, and faculty and the courses they teach (c.f., Borgatti and Everett 1997). Researchers have also used this approach to examine how project leaders' involvement in other opensource software projects affects project value (Grewal et al. 2006, Oh and Jeon 2007). In this research, we analyze how shared contributors who work on multiple content sources connect the sources of user-generated content into a two-mode content-contributor network.

We test our theoretical predictions by analyzing Wikipedia's Medicine Wikiproject and examining how characteristics of the content-contributor network affect the market value of collaborative user-generated content. We assess market value through viewership since viewership is a primary determinant of the revenue that advertisers obtain from user-generated content. Our results provide good support for the hypothesized relationships. In particular, we find that the number of contributors is curvilinearly related to content value. We also find that the embeddedness of the content in the content-contributor network is positively related to content value. Finally, we find these effects are stronger for newer rather than older content. Out-of-sample analyses show that our models have good predictive validity not only within the same topic but also for completely different topics (i.e., the auto and fashion Wikiprojects).

By explicitly accounting for the network of content and content creators, we add to research that has treated these independently. By identifying the characteristics that distinguish collaborative user-generated content from content created by individuals, we point the way for future research on this growing phenomenon. We contribute methodologically by integrating two-mode SNA with hierarchical linear and Bayesian modeling and applying these approaches to large-scale data sets. We also contribute methodologically by examining network dynamics over time, which researchers have noted is an often-overlooked area of network research (Borgatti et al. 2009).

The methodology we use to form a two-mode network can be extended into many other domains of interest to marketing researchers outside of the specific case we demonstrate. For example, the approach we use can be used to study the extent to which consumers are connected through the brands they own and what this implies for brand choice (Berger and Heath 2007), how the

embeddedness of salespeople within the sales-customer network affects customer lifetime value, and how connections among reviewers and the restaurants they review affect the quality of reviews that are posted. Our approach also demonstrates a way to analyze databases that are much larger than those traditionally analyzed by marketing researchers (Naik et al. 2008). Beyond adding to prior research focused on content created by individuals (Chevalier and Mayzlin 2006, Godes and Mayzlin 2004, Moe and Trusov 2011), our results have practical implications for marketing practitioners who seek to encourage content creation by groups as well as individuals (Kozinets et al. 2008, Li and Bernoff 2008).

## 2. Theoretical Development

We explore three aspects of collaborative user-generated content that explain its perceived value. First, because the knowledge and effort provided by users are the primary inputs for developing user-generated content, the number of contributors available to a collaborative project may be an important predictor of the value of user-generated content. Although attracting a sufficient number of contributors to sustain collaboration is important, it is also possible to attract so many contributors that collaboration is impaired. Second, because contributors can apply content knowledge and collaboration skills acquired on one project to others on which they work, the embeddedness of user-generated content in the network of content and contributors is likely to be an important predictor of its value. Third, because collaborative user-generated content is likely to stabilize as it matures, content age should be an important moderator of the influence of these antecedents on content value.

### 2.1. Number of Contributors

Attracting a sufficient number of contributors is important for collaborative user-generated content. More contributors increase the effort and energy dedicated to creating content and provide a broader array of knowledge and abilities for content creation. This should increase the value of collaborative user-generated content. Research on prediction markets, virtual teams, and social networks suggests that the quality of aggregate information, number of ideas generated, and likelihood of a valuable answer increases with the number of participants (Constant et al. 1996, Foutz and Jank 2010, Martins et al. 2004).

At the same time, other research suggests that having too many contributors can also be problematic. After a certain point, the marginal cost of adding new members exceeds its marginal value. Consistent with the adage “too many cooks spoil the stew,” an excessive number of contributors negatively influences the value of user-generated content. As the number of contributors grows,

the marginal value of additional contributors decreases while the cognitive and coordination costs associated with contributions increases (Asvanund et al. 2004, Jones et al. 2004). In particular, those involved in the co-creation of content are likely to suffer from information overload as they try to make sense of and respond to others' contributions.

In consumer settings, increasing the amount of information that consumers are asked to evaluate slows processing speed, lowers choice quality, and reduces the likelihood that a choice is made (Iyengar and Lepper 2000, Lurie 2004). In computer-mediated environments, information overload can negatively affect a group's ability to organize information effectively (Hiltz and Turoff 1985). In collaborative online environments, such as those studied here, information overload lowers participation, reduces the likelihood that longer (and potentially more valuable; Schlosser 2007, Weiss et al. 2008) contributions are read, reduces contributor effort, and leads to shorter contributions as participants seek to reduce their cognitive load (Jones et al. 2004). The condition of "too many contributors" is increasingly common in new social media platforms that can attract thousands of users in a short time (Kane 2011).

This rationale suggests a curvilinear relationship between number of contributors and content value. The most valuable collaborative user-generated content is generated when enough contributors are attracted to sustain production but not so much that it creates information overload for contributors. Considerable empirical evidence supports such curvilinear relationships between number of contributors and outcomes in online collaborative groups (Asvanund et al. 2004, Butler 2001, Hansen and Haas 2001, Oh and Jeon 2007). Similar relationships have also been found in traditional organizations. For instance, moderate-sized firms are often more able to capitalize on new markets than small firms or large firms, the former lacking resources to innovate and the latter becoming too bureaucratic and rigid (Haveman 1993). In workgroups, new members introduce additional coordination cost and an increasing diversity of perspectives, making it more difficult for teams to reach consensus (Lovelace et al. 2001). For instance, software development teams need sufficient resources to accomplish their goals, but adding more members to a troubled or delayed project can compound delays by increasing coordination costs (Brooks 1975) as new members are added (Espinosa et al. 2007). Thus, we expect a curvilinear relationship between the number of contributors and the market value of collaborative user-generated content. These ideas lead to our first hypothesis:

*HYPOTHESIS 1. The market value of collaborative user-generated content has a curvilinear (inverted U) relationship with the number of contributors to it.*

## 2.2. Network Embeddedness

Although the energy and knowledge provided by the direct participation of contributors is an important resource for collaborative user-generated content, previous research points to the role of social capital in the development of intellectual capital (Adler and Kwon 2002, Nahapiet and Ghoshal 1998, Gu et al. 2008). Social capital is defined as “the sum of the actual and potential resources embedded within, available through, and derived from the network of relationships possessed by an individual or social unit” (Nahapiet and Ghoshal 1998, p. 243). Marketing researchers have noted the importance of social capital for generating customer solutions (Tuli et al. 2007) and developing effective governance relationships (Gu et al. 2008).

Researchers often refer to the role of social capital in production as *network embeddedness*, the degree to which a person or project is connected to other people or projects in the network (Granovetter 1985, Grewal et al. 2006, Gulati and Gargiulo 1999).<sup>1</sup> For collaborative user-generated content, embeddedness refers the extent to which a particular piece of content is connected to other pieces of content through the network of content creators.

Different types of network ties—such as proximities, relations, interactions, and flows—can mediate social capital (Borgatti et al. 2009). For example, work by collaborators on a common project is a salient mechanism for social capital in the creation of collaborative user-generated content (Monge and Contractor 2003) as it allows users to access network resources through both direct and indirect interactions with other users. A contributor may be exposed to valuable resources—such as relevant content or references, effective presentation styles, and how to manage conflict in the collaborative environment—even if they do not know the identity of the contributor from whom they acquired this knowledge. They may also learn the reputation of other contributors as effective or ineffective collaborators through work on other collaborative projects through simply observing their contributions and not interacting with them directly.

The more embedded that collaborative user-generated content is in the content-contributor network, the greater access to the knowledge that has been combined and exchanged in other projects (Lin 1982). For collaborative user-generated content, the primary resource is information and knowledge, and social capital enhances the value of these resources by creating opportunities for the combination and exchange of existing knowledge to enhance its value (Nahapiet and Ghoshal 1998). Contributors can *transfer* the knowledge they acquire working on one project to the other

<sup>1</sup> Although network researchers have forwarded many different categories of embeddedness (Zukin and DiMaggio 1990), consistent with previous research (Grewal et al. 2006), we focus primarily on structural embeddedness—the structural properties of the network ties.

projects on which they work, combining and exchanging the transferred knowledge with knowledge contributed by others (Reagans and McEvily 2003). The more connected these contributors are to other collaborative environments, the better access they will have to the information resources available in the networks to improve the sources of user-generated content to which they contribute.

The access to information and knowledge resources in the network also allows contributors to *transform* the knowledge they acquire from working on other sources of user-generated content by combining it with their own experience and creating new knowledge (Carlile and Rebentisch 2003). This ability to transform existing information and create new knowledge can increase the value of acquired knowledge and that of contributors. Much in the same way that experts show superior task performance based on knowledge and experience (Alba and Hutchinson 1987), greater social capital allows contributors to more efficiently identify and transform valuable information into useful formats (Spence and Brucks 1997), provide more comprehensive information (Alba and Hutchinson 1987), and transfer relationships among content items in ways that makes content more informative (Gregan-Paxton and John 1997).

These ideas are consistent with current approaches to social capital that include connections to shared creations, such as relationships among software developers through projects on which they collectively work (Grewal et al. 2006, Oh and Jeon 2007, Singh et al. 2011). Network embeddedness is positively associated with workgroup performance (Oh et al. 2004) and production value (Grewal et al. 2006, Mallapragada et al. 2008). The importance of network embeddedness has been established in online (e.g., Grewal et al. 2006, Mallapragada et al. 2008) as well as offline (e.g., Uzzi 1997) settings. Thus, network embeddedness allows contributors to access information and knowledge resources available in other sources of collaborative user-generated content and apply those resources to improve the value of the content to which they contribute. Following this logic, we hypothesize that:

*HYPOTHESIS 2. The market value of collaborative user-generated content will be positively related to its embeddedness in the content-contributor network.*

### **2.3. Content Age**

The impact of these collaborative inputs on the value of user-generated content, however, may be different for older versus newer content. Unlike individually created content, such as consumer reviews, in which contributors are free to disagree, and for which there are no limits on the amount of content created, collaborative user-generated content often requires contributors to reach consensus and often places functional limits on content length (McAfee 2007). Collaborative work often

proceeds in stages, with early stages more chaotic and malleable and later stages more focused and reified (Tuckman 1965). Similar stages have also been found in online groups developing collaborative user-generated content (Ransbotham and Kane 2011), and researchers have argued that project stage should be considered when assessing the impact of antecedents to collaborative work (Hansen et al. 2005).

Developing newer collaborative content involves a number of costs not associated with more established content, such as determining project scope, collaborative norms, and informal leadership (Kuk 2006). Newer content also does not attract potential contributors as easily as more established content. People often join a collaborative project by first “lurking,” or observing collaboration to develop trust and learn the collaborative environment (Lave and Wenger 1991), which is difficult with newer content since there is less to observe. Contributors are also likely less willing to join a project that does not exhibit a strong likelihood for success for fear of wasting time and effort (Madey and Freeh 2004), and eventual success is more difficult to determine early in a project’s life. Thus newer collaborative projects face additional challenges while possessing fewer resources, so contributors and their associated social capital are particularly valuable early in the content creation process.<sup>2</sup>

As content matures, however, the relative value of these resources diminishes. Content often reaches a stable equilibrium as it is refined through the ongoing work of its contributors, making it less adaptive to adapt to changes in underlying knowledge (March 1991). This is particularly acute in online settings where the collaborative platform can preserve and synthesize the contributions of prior members (Kane and Alavi 2007). The knowledge and experience accumulated thorough previous collaboration can create competency traps that make it more difficult for established groups to capitalize on innovations and new knowledge (Levinthal and March 1993). The value of social capital may also diminish over time. Once a network provides access to the most valuable knowledge and resources, the proportion of redundant or undesirable knowledge increases, making it more difficult to search for and find remaining valuable resources through this network (Aral and Van Alstyne forthcoming). Similarly, network embeddedness can reinforce established perspectives and norms, making groups less receptive to new information available for development processes over time (Uzzi 1997).

These arguments suggest that the impact of the network on content market value should be stronger for newer than for older content. In particular, when user-generated content is older and

<sup>2</sup> Although enterprise-level communities, such as Wikipedia or Sourceforge, provide valuable resources that help mitigate these start-up costs associated with individual user-generated projects, they cannot eliminate these costs entirely.

more difficult to change, the impact of having an optimal number of contributors (i.e., not too many and not too few) should have less impact on content value. Similarly, the importance of social capital, and the resultant ability to transform and transfer knowledge from other content, should have a lesser effect on the value of older, and therefore less malleable, collaborative user-generated content.

*HYPOTHESIS 3. The impact of (a) the number of contributors and (b) embeddedness on the market value of collaborative user-generated content declines with content age.*

### **3. Research Setting and Method**

We use two-mode Social Network Analysis (Wasserman and Faust 1994, Faust 1997) to examine how network characteristics affect the value of collaborative user-generated content. Like Grewal et al. (2006), we use this approach to examine how network embeddedness affects content creation. However, since collaborative user-generated content sources often do not have a formal project leader, we examine the previous involvement of all contributors to a particular source of user-generated content. This two-mode network approach is consistent with the idea that collaboration on shared projects provides access to network resources that enhance the creation of collaborative user-generated content.

#### **3.1. Research Setting**

We examine the relationships among 16,068 Wikipedia articles in the Medicine Wikiproject (i.e., all articles in this project during the study period) and the contributors to these articles. Drawn from the Hawaiian word meaning “quick,” a wiki is a Web site that anyone can edit. Wikipedia, established in 2001, uses a wiki platform to host an open-source encyclopedia. Users of the English version of Wikipedia have generated more than 3 million separate articles, and an additional 13 million articles are available in the 270 other languages in which Wikipedia is published. Although anyone can contribute to any article on Wikipedia, most contributions are made by a core group of individuals. In a Wikiproject, a group of contributors commits to develop, maintain, and organize articles related to a focal topic. The hundreds of Wikiprojects on Wikipedia are dedicated to a wide range of topics, from the mainstream to the obscure. Considerable research has investigated collaboration on Wikipedia (Denning et al. 2005, Kittur and Kraut 2008, Kriplean et al. 2008) and even conceptualized Wikipedia as a network (Brandes et al. 2009, Capocci et al. 2006, Zlatić et al. 2006), though most studies examine the topical network (i.e., articles and internal links); not the relationship between content contributors and the market value of information created by the

network. We also assess the model performance by estimating viewership in two additional large samples—the fashion and auto Wikiprojects—to confirm the robustness of our findings.

We focus on a single Wikiproject, because traditional sampling methods cannot be used for SNA (Wasserman and Faust 1994) and a network analysis of 16 million articles over time is computationally intractable. A Wikiproject provides clearly defined boundaries and norms for the network, permitting analysis. It also allows a comparison of the relative market value of the content, because content has vastly different viewership in different Wikiprojects. Moreover, studying articles dedicated to a particular Wikiproject limits the impact of potentially confounding factors. Because of their common subject matter, these articles are more likely to share contributors such that we obtain a relatively smaller, clearly defined, cluster of articles and contributors than we would with a wider, unconstrained, sample of Wikipedia articles.

We focus on health and medical information, as it often represents an early and prominent use of online sources (Ferguson and Frydman 2004). A recent Pew study reveals that Internet users increasingly turn to user-generated health and medical information online, and nearly 60% of Internet users have relied on Wikipedia as a source of health information (Fox and Jones 2009). The healthcare industry also draws on user-generated content to promote lifestyle changes, encourage collaboration among physicians, develop collaborative patient support networks, and provide a valuable resources to patients and providers (Kane et al. 2009). Previous studies have affirmed the quality of medical information on Wikipedia (e.g., Clauson et al. 2008), which also has considerable economic value. Healthcare in the United States is a \$2.3 trillion industry, and by 2011, online pharmaceutical advertising expenditures are expected to reach \$2.2 billion, or 5% of Internet advertising (Phillips 2007). Wikipedia does not accept formal advertising, but other online providers of medical content (e.g., WebMD, HealthCentral) do, and these sites increasingly leverage user-generated content, such as blog communities and user forums.

### **3.2. Primary Data Collection**

We downloaded the full text history of 2,029,443 revisions of 16,068 articles by 40,479 unique contributors in the Medicine Wikiproject as of June 2009, which resulted in a 50 GB data set of raw data. We employed a 70-node Linux cluster to allow for simultaneous downloads and processing of these extensive data. For each contribution, we record the contributor’s identity, the changes made, a description of the change, and the time of the change.

To ensure that our analysis was based on the behavior of people, rather than computers, we excluded edits made by automated software programs (i.e., bots). Wikipedia’s site policy requires that all bots be approved and registered; we obtained a list of active and previously active bots

from Wikipedia. Bots on this list made 2% of the changes in our sample (37,237 revisions) and we excluded their edits from the analysis. A manual check of 75 random articles similarly revealed that 2.13% of the edits were bot activity. We also manually checked the userpages of the 100 most prolific contributors and found no unknown bots. Bot activity in other areas could be greater, as Wikipedia’s own statistics show that most automated edits occur in non-English-language Wikipedias and reflect particular types of edits (e.g., formatting dates, deleting placeholder articles called *stubs* that have not yet been developed).<sup>3</sup> Thus, though we may have missed some edits by unregistered bots in our sample, we excluded most automated activity.

From the remaining full-revision history, we constructed a 132,447-observation monthly panel. For each month, we built a two-mode affiliation network and linked articles through contributors. We represent the two-mode network as a 16,068 row (article) by 40,479 column (contributor) sparse matrix, in which the values in the matrix cells represent the number of contributions for the article-contributor pair. The 141,282 non-zero elements in the sparse matrix represent articles in the medicine Wikiproject and contributions to them. To measure local and global network centrality, we created a  $16,068 \times 16,068$  incidence matrix of contributors and content sources by multiplying the matrix by its transpose. An incidence matrix is a common way to represent two-mode networks (Faust 1997). Because we view user-generated content as composed of discrete content sources, connected by individuals who contribute to them, our incidence matrix treats content sources as nodes and contributors as ties.

### 3.3. Dependent Variable

We operationalize market value as the number of times a Wikipedia article is viewed in a given month. Viewership reflects the value that the market ascribes to particular content, and advertisers focus on content that delivers more viewers (Miller 2009). For each article, we collected the number of views each day from December 2007 until June 2009; these data are not available for the entire history of Wikipedia. We summarized the view counts by month; then scaled the monthly article views by the number of days in the month so that months with fewer than 31 days were comparable with months with 31 days. Article views are integer counts, but we transform the dependent variable by taking their natural log. A Shapiro-Francia test fails to reject the null hypothesis that the distribution is normal ( $W' = 0.9914$ ,  $p < 0.5$ ).

<sup>3</sup> See <http://stats.wikimedia.org/EN/BotActivityMatrix.htm>.

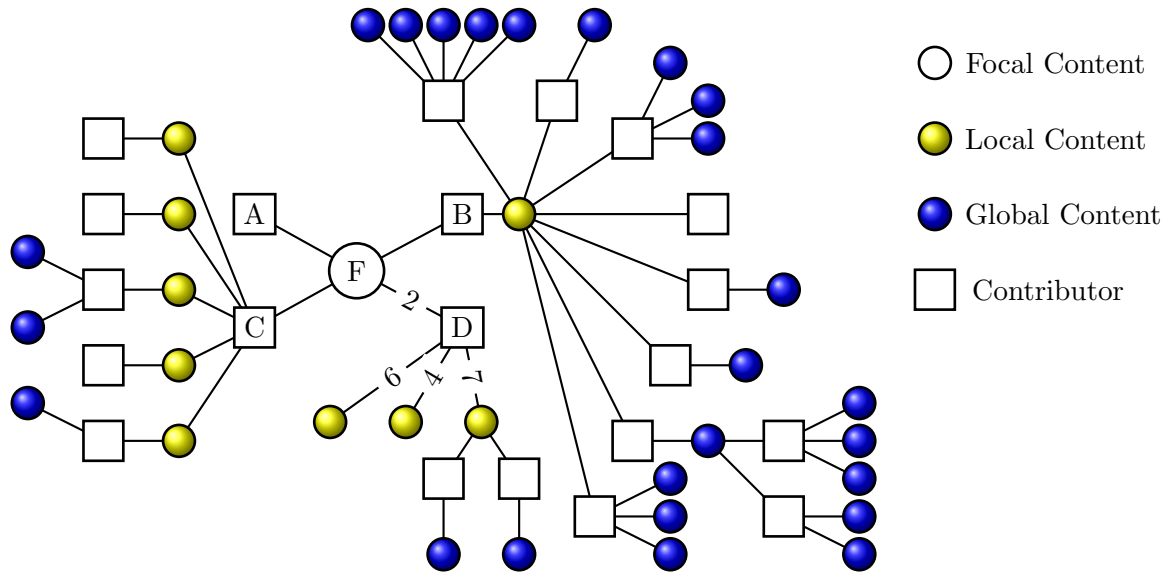


Figure 1 Example of a Collaborative User-Generated Content Network

### 3.4. Focal Independent Variables

We use a small, simplified network to illustrate our two-mode network conceptualization. Figure 1 depicts potential relationships among creators of collaborative user-generated content and the content they create. Circles represent content sources; in Figure 1, the focal content is labelled (F). Individual contributors are represented by squares; in Figure 1, contributors to the focal content are labelled (A–D). Arcs between nodes indicate contributions to content; in Figure 1, numbers indicate the number of contributions made by Contributor D to different content sources.

**3.4.1. Number of Unique Contributors** We measure the number of distinct contributors to user-generated content during the monthly observation period. In Figure 1, four contributors (A–D) contribute to focal content (F).

**3.4.2. Network Embeddedness** Network centrality assesses the embeddedness of a node in a network. Network researchers have developed a number of centrality measures, such as degree, closeness, betweenness, eigenvector, and flow betweenness, to name a few (Faust 1997, Wasserman and Faust 1994). Although these measures are often correlated, each captures slightly different aspects of network structure. Choice of a particular centrality measure should be based on the network content and the type of resources to which the network provides access. Borgatti (2005) argues that two centrality measures are most appropriate for studying knowledge and expertise as resources that can be replicated and spread in many directions simultaneously—degree centrality and closeness centrality. Since each of these measures reflects somewhat different aspects of the

network, we employ both to capture two dimensions of network embeddedness—local and global network centrality.

*Local network centrality* reflects the direct collaborative activity of contributors to the focal content and other sources of collaborative user-generated content. For example, in Figure 1, Contributor A contributes to no other content sources, Contributor B contributes to one other content source, Contributor C contributes to five other sources, and Contributor D to three. We measure local centrality by calculating the degree centrality of the content in the incidence matrix of contributors and content sources. Degree centrality assesses the number and strength of direct connections possessed by a node, capturing the level of social capital immediately available to the node (Wasserman and Faust 1994). We operationalize degree centrality as the number of connections to other articles made by shared contributors, weighted by the number of contributions made. Because this value is correlated with the number of contributors, we divide degree centrality by the number of contributors to yield a relative measure. For the example in Figure 1, Contributor D adds 34 ( $2 \times (6 + 4 + 7)$ ) to the degree centrality of article F; the total degree centrality of article F also incorporates the activity through Contributors A–C (although Contributor A adds nothing). For scaling purposes, we divide local network centrality by 1000.

*Global network centrality* assesses the position of the focal content source relative to all other sources of collaborative user-generated content in the network as a whole. Although degree centrality captures the social capital available through direct connections (i.e., the local network), it does not reflect the social capital of indirect connections (i.e., the global network; Faust 1997). For instance, in Figure 1, although Contributor B only contributes to one other content source, many other contributors work on that content source and those contributors work on many other content sources. Any one of these other contributors could add valuable information obtained through their collaborative activity to the article that Contributor B works on, making it accessible to the focal article F. That is, once a contributor uses information gained from one source to improve a second source, all subsequent contributors to the second source now have access to this information to improve the other projects on which they are working. Thus, in Figure 1, Contributor B provides a short path between focal content (F) and an extensive amount of collaborative activity occurring in the network. This should enhance the value of focal content (F).

To measure global network centrality, we use closeness centrality (Freeman 1979), which measures a node's average distance to all other nodes in the network. Nodes that are closer on average to all other nodes in the network will have better access to the resources embedded in the network as a whole. Another appropriate and frequently used measure to capture the global centrality

in knowledge networks is eigenvector centrality (Borgatti 2005). We focus on closeness centrality because of its stability (Costenbader and Valente 2003) and its prior use to reflect information flow (Borgatti 2005).

Closeness centrality is the mean geodesic distance between a node,  $v$ , and all other nodes in the graph. For our two-mode network, we focus only on the distances between nodes of the same type (articles). For an article node  $v_a$ , we calculate the closeness centrality,  $C_c(v_a)$ , within graph  $G(V_a)$  as

$$C_c(v_a) = \frac{(|V_a| - 1)}{\sum_{v'_a \in V_a \setminus v_a} d(v_a, v'_a)} \quad (1)$$

where  $d(v_a, v'_a)$  is the shortest path between article node  $v_a$  and another article node  $v'_a$ . For the 35-node example network in Figure 1, the closeness centrality of article F is  $0.53 = (35 - 1) / ((9 \times 1) + (20 \times 2) + (5 \times 3))$  since there are 9 articles with distance of 1, 20 articles with a distance of 2 and 5 articles with a distance of 3. (Closeness centrality is also frequently calculated as the reciprocal of Equation 1; in the calculation we use, higher values represent more central, closer, articles.)

**3.4.3. Content Age** Age equals the time in days since the article first appeared in Wikipedia; we use the natural log of the number of days. Article ages range from one day to 8.1 years, with an average of 2.9 years.

In our analysis, we lag the number of district contributors and network centrality measures by one month. Editing itself may cause viewing and viewership could drive collaborative activity. Because the measures of viewership and network characteristics instantaneously reflect actual viewership and network collaborative activity, using lagged values decreases the likelihood that our results are driven by reverse theoretical mechanisms that we have not hypothesized. Our results are robust, however, to the use of non-lagged network characteristics.

### 3.5. Relative Topic Popularity

Rather than the causal relationship we hypothesize, article views and network characteristics could simply reflect the underlying popularity of the article topic. Therefore, we explicitly control for the popularity of the article topic using its relative search frequency in Google during the time period under study. Using results from *Google Insights for Search*, we determined the number of times that users of Google search for keywords from the article title for each week of our analysis period (December 2007-June 2009). Unfortunately, although an excellent control for popularity, Google Insights has important limitations: 1) it provides search volume relative to other topics rather than absolute search volume; 2) it allows a maximum of five topics per request.<sup>4</sup> Therefore,

<sup>4</sup> Another important pragmatic limitation is that Google caps the number of requests per day at 100 per a range of Internet addresses. We thank Google and Hal Varian for helping us bypass the maximum number of daily requests.

we created a set of 6,367 distinct search requests that each contained one common topic and four new topics. The use of the common topic allowed us to create an index of search popularity relative to the common topic. Because Google Insights reports weekly search volume, we summarized the weekly relative popularity by month to match the Wikipedia article viewing measure. (Where weeks crossed monthly boundaries, we interpolated based on the number of days assuming a uniform distribution of search volume during the week.) Thus our measure independently captures the popularity of the keywords in the article title relative to other articles in our sample over time.

### 3.6. Control Variables

To control for factors other than network characteristics and topic popularity that may affect the number of article views, we include length, reading complexity, anonymity of contributors, amount of multimedia content, information presentation, external references, and internal links as covariates. In Table 1, we present descriptive statistics and, in Table 2, correlations.

**3.6.1. Length.** Although Wikipedia has length guidelines (Wikipedia 2010), one group of active Wikipedians argues that, because it is not bound by the confines of traditional printed encyclopedias, an article should contain all possible relevant information about a particular topic (McAfee 2007). In short, an article may be more valuable simply because it has more; not better, information. To control for this possibility, we include the length of each article, expressed in thousands of characters of text (for scaling purposes), which ranges from 0 (for stub articles) to 1,094,011 characters. We use the natural log of article length in the statistical models.

**3.6.2. Reading complexity.** Articles may be more valuable if written in a more sophisticated style. Alternatively, articles may be incomprehensible if they are difficult to read. We control for the reading complexity of each article using the automated readability index (Smith and Senter 1967). (We applied models using the Coleman-Liau index and found similar results.) The ARI equals  $ARI = (4.71 \times letters/words) + (0.5 \times words/sentences) - 21.43$ , and estimates the U.S. school grade required to understand the article.<sup>5</sup> While many measures of reading complexity exist, the ARI and Coleman-Liau measures are well suited to automated processing of large data since they do not require dictionary matching or syllable breakdowns.

<sup>5</sup> The numeric constants (and the remainder of the formula) represent the estimates developed through a long history of research on reading complexity. The formula predicts the school grade reading level (U.S.) for a particular text; the constants are only used to scale the formula result. For our analysis, we focus on relative differences in reading complexity, not the absolute value.

**Table 1** Descriptive Statistics

Variable	Minimum	Maximum	Mean	Std. Dev
Monthly Article Views (/1000)	0.01	3,675.59	11.15	29.57
Age (days)	29.00	2,979.00	1,311.77	615.69
Length (characters/1000)	0.00	1,094.01	10.30	12.64
Complexity (ARI)	0.01	1,281.56	19.48	7.11
Section Depth	1.00	6.00	2.42	0.73
External References	0.00	303.00	10.37	21.96
Internal Links	0.00	3,508.00	60.32	73.89
Multimedia Content	0.00	35.00	0.05	0.50
Anonymity (percentage)	0.00	1.00	0.29	0.16
Relative Popularity	0.00	43.08	0.03	0.58
Distinct Contributors	0.00	1,592.00	28.08	37.34
Local Centrality	0.00	2,147.48	98.30	198.57
Global Centrality	0.00	12.20	2.18	2.55

**3.6.3. Anonymity of contributors.** People can contribute to an article, whether they log in and identify themselves in the Wikipedia system or not. If a contributor is not logged in, his or her identity is recorded as an anonymous IP address. Anonymous contributors represent part of the collaborative network we cannot capture, though anonymity may affect the nature of collaborative interactions—helping in some situations and hurting in others (Sia et al. 2002). Because the raw number of anonymous contributors is highly correlated with the total number of contributors, we used the percentage of anonymous contributors, calculated as the total number of anonymous contributors divided by the total number of contributors to an article. On average, 29% of the total contributors to each article were anonymous.

**3.6.4. Information presentation.** Because multimedia content may enhance the value of information (Schlosser 2003), we control for the total number of multimedia files in an article using a measure labeled *multimedia content*. Similarly, an article’s organization may influence its market value, because well-organized information should be more accessible to readers. Articles in Wikipedia can contain up to six levels of nested sections. To control for this effect, we include the maximum section level reached in the article, which we refer to as *section depth*.

**3.6.5. References and links.** Wikipedia policy states that all contributions should be supported by an authoritative external reference. In the Medicine Wikiproject, only peer-reviewed medical journals are considered authoritative. Contributors may attempt to manipulate the market value of an article by including more references, or the number of references could indicate the

**Table 2** Variable Correlations

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Article Views (ln)	1.00											
2. Age (ln, days)	0.55	1.00										
3. Length (ln, chars)	0.39	0.22	1.00									
4. Complexity (ARI)	0.07	0.01	0.30	1.00								
5. Section Depth	0.33	0.19	0.62	0.14	1.00							
6. External References	0.22	0.07	0.50	0.07	0.38	1.00						
7. Internal Links	0.41	0.29	0.55	0.17	0.46	0.53	1.00					
8. Multimedia Content	0.04	0.01	0.05	0.06	0.08	0.03	0.34	1.00				
9. Anonymity (%)	0.53	0.50	0.27	0.06	0.20	0.01	0.23	0.01	1.00			
10. Relative Popularity	0.10	0.06	0.05	0.01	0.06	0.03	0.10	0.01	0.05	1.00		
11. Contributors	0.48	0.36	0.36	0.05	0.33	0.35	0.55	0.02	0.38	0.28	1.00	
12. Local Centrality	-0.10	-0.09	0.18	0.03	0.11	0.23	0.06	0.01	-0.29	-0.02	-0.11	1.00
13. Global Centrality	0.30	0.19	0.30	0.06	0.21	0.22	0.26	0.01	0.23	0.04	0.30	0.03

Correlations for the 131,109 Wikipedia Medicine monthly panel observations from December 2007 to June 2009. All correlations greater than 0.01 are significant.

popularity of a topic in the medical literature. For example, although lung cancer is the leading cause of U.S. cancer deaths, it is relatively underfunded and under-researched compared with other forms of cancer (Khullar and Colson 2009). Articles also often contain links to other Wikipedia articles that may be sources of views or a reflection of market value. Accordingly, we control for the number of *external references* and the number of *internal links*.

## 4. Results

We first analyze the effects of network structure on article views using hierarchical models that incorporate unobserved within-article clustering, verifying that the results are robust to a range of modeling assumptions. Then, we use two alternative samples that demonstrate the robustness of our findings beyond the Medicine Wikiproject. Finally, we extend the analysis with Bayesian hierarchical models using Markov Chain Monte Carlo simulation.

### 4.1. Hierarchical linear latent models

Our initial analysis utilizes general linear latent mixed models that incorporate the correlation between the interdependent and correlated repeated observations of the same article. The latent article level effects help control for article heterogeneity not captured by our control or focal independent variables. For article  $i$  at time  $t$ , we first incorporate heterogeneity in general article viewing through a random intercept hierarchical model. In this model (Equation 2), article views

( $v_{i,t}$ ) depend on monthly control covariates (vector  $\mathbf{C}_{i,t}$ ) and monthly focal network covariates (vector  $\mathbf{N}_{i,t-1}$ ); these are nested within overall article level random intercepts ( $\zeta_{0,i}$ ).

$$\ln(v_{i,t}) = (\beta_0 + \zeta_{0,i}) + \beta_c \mathbf{C}_{i,t} + \beta_n \mathbf{N}_{i,t-1} + \epsilon_{i,t} \quad (2)$$

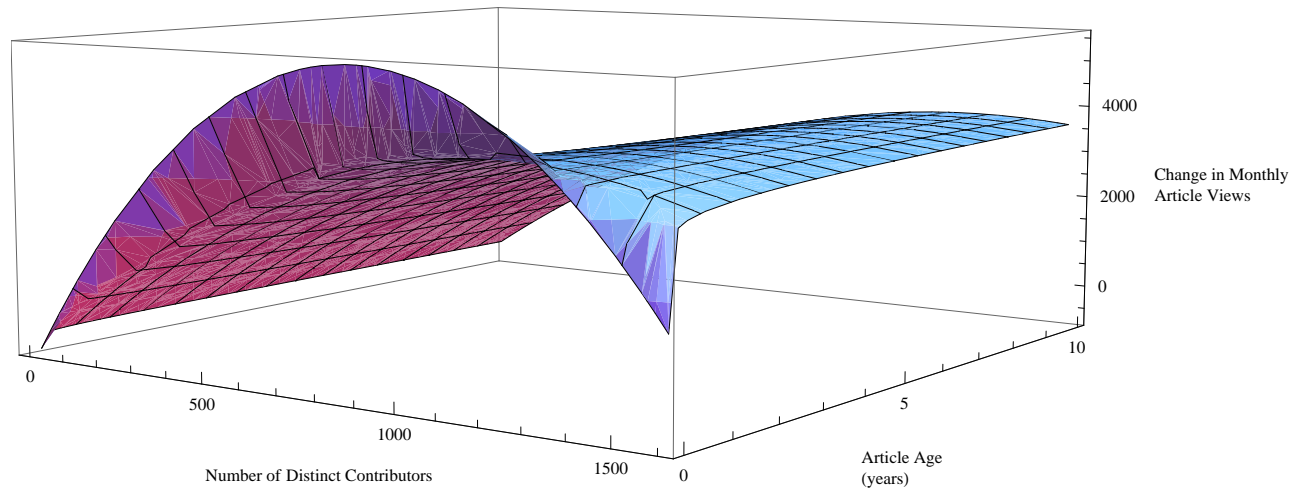
We then extend this analysis to allow for heterogeneity in coefficients. This extension ensures that the main effects of our focal network characteristics are not overly influenced by extreme viewership of some articles and also captures interesting variation in article activity. In this model (Equation 3), article views ( $v_{i,t}$ ) depend on monthly control covariates (vector  $\mathbf{C}_{i,t}$ ) and monthly focal network covariates (vector  $\mathbf{N}_{i,t-1}$ ); these are nested within overall article level random intercept ( $\zeta_{0,i}$ ) and random coefficients ( $\zeta_{n,i}$ ) for network covariates (vector  $\mathbf{N}_{i,t-1}$ ).

$$\ln(v_{i,t}) = (\beta_0 + \zeta_{0,i}) + \beta_c \mathbf{C}_{i,t} + (\beta_n + \zeta_{n,i}) \mathbf{N}_{i,t-1} + \epsilon_{i,t} \quad (3)$$

Table 3 summarizes the results of this analysis. All continuous variables are standardized. For computational tractability, we focus our analysis on the 7,535 articles that progress beyond the stub classification.

Model 1 includes only control covariates and allows for heterogeneity in article intercepts. Model 2 adds our focal network covariates. To assess statistical significance, we used Markov Chain Monte Carlo sampling based on 10,000 iterations. In support of Hypothesis 1, we find a curvilinear relationship between distinct authors and article viewing; the linear term is positive ( $\beta = 57.04$ ,  $p < 0.001$ ) while the second order term is negative ( $\beta = -4.45$ ,  $p < 0.001$ ). In support of Hypothesis 2, we find a significant and positive ( $\beta = 7.02$ ,  $p < 0.001$ ) effect of local (degree) centrality on article views as well as a significant and positive effect ( $\beta = 2.90$ ,  $p < 0.001$ ) of global (closeness) centrality. These effects are practically as well as statistically significant. A one standard deviation increase in local centrality increases viewership seven-fold; similarly, a one standard deviation increase in global centrality almost triples viewership. We considered alternative specifications of the relationship between authors and viewing; the curvilinear model has a slightly improved AIC over linear functional forms (a reduction in AIC of 40).

Model 3 extends Model 2 by allowing for heterogeneity in network covariate coefficients; the hypothesized relationships are seen again in the mean coefficients. (We did not include article-level random coefficients for the quadratic term for authors since the variance was highly correlated with the linear authors term.) For the random coefficient models, we report  $p$ -values for Models 3 and 4 based on whether zero falls outside of the 5%, 1%, and 0.1% confidence intervals. Although it is difficult to establish degrees of freedom in models with random coefficients, our sample size allows



**Figure 2** Effect of Number of Contributors and Article Age on Article Views

conservative underestimation of degrees of freedom and indicates high probabilities of rejecting the null hypothesis and finding support for our hypotheses.

Model 4 adds interactions of article age (natural log) with the network covariates. We find support for Hypothesis 3—content age reduces, but does not completely offset, the effect of network covariates on the market value of user-generated content. Content age reduces the effects of number of authors (linear  $\beta = -38.25$ ; quadratic  $\beta = 3.63$ ), local network centrality ( $\beta = 2.17$ ) and global network centrality ( $\beta = 0.76$ ). Figure 2 illustrates the relationships among number of contributors, article age, and viewing activity. Although stronger in the first years of an article’s existence, the curvilinear relationship endures throughout the observed lives of the articles. Because we model the natural log of age, the effects of increasing age diminish as the content gets older. In addition to the natural log transformation of age, we considered both linear and quadratic transformations; both slightly increase the root mean squared error (from 29.481 to 29.557 and 29.552 respectively) and the mean absolute percentage error (from 2.117% to 2.182% and 2.169% respectively). Thus, we retained the log transformation as the best fitting functional form.

#### 4.2. Alternative Samples

To assess the predictive validity of the models in alternate contexts, we used two alternative samples—the fashion and auto Wikiprojects. These Wikiprojects are comparatively smaller (2,503 and 6,890 articles, respectively) but are interesting to study because of the importance of marketing to these industries. We collected the full text of 1,026,892 revisions in the Auto Wikiproject and

**Table 3** Linear Latent Mixed Model Regression on Article Views

Variable	Model 1	Model 2	Model 3	Model 4
Article Intercepts	Heterogeneous	Heterogeneous	Heterogeneous	Heterogeneous
Article Coefficients	Homogeneous	Homogeneous	Heterogeneous	Heterogeneous
<b>Article Level standard deviations</b>				
Residual	49.885	48.583	31.796	31.788
Intercept	166.164	146.053	272.766	274.305
Contributors			841.386	844.109
Local Centrality			313.213	314.020
Global Centrality			14.659	14.557
<b>Revision Level coefficients and standard errors</b>				
Monthly Fixed Effects	indicators	indicators	indicators	indicators
Constant	691.361*** (2.316)	764.928*** (2.217)	899.250*** (5.355)	910.139*** (5.489)
Age (ln, years)	12.673*** (0.455)	25.020*** (0.668)	18.642*** (0.686)	0.480 (2.480)
Length (ln, characters)	11.969*** (0.430)	11.245*** (0.436)	6.435*** (0.359)	6.395*** (0.359)
Complexity (ARI)	0.582** (0.206)	0.498** (0.202)	0.600*** (0.140)	0.603*** (0.140)
Section Depth	3.624*** (0.478)	3.140*** (0.491)	2.314*** (0.473)	2.257*** (0.473)
External References	8.569*** (0.490)	3.847*** (0.518)	2.158*** (0.651)	1.886** (0.652)
Internal Links	3.380*** (0.591)	1.452** (0.590)	-1.407** (0.465)	-1.448** (0.465)
Multimedia Content	-1.650** (0.613)	-0.882 (0.601)	-28.015*** (0.472)	-0.012 (0.472)
Anonymity (percentage)	58.337*** (3.957)	53.941*** (4.158)	41.857*** (4.224)	26.165*** (4.233)
Relative Popularity	30.865*** (3.463)	20.485*** (3.078)	88.042*** (11.757)	83.801*** (11.732)
Contributors		57.042*** (1.694)	387.713*** (14.274)	422.019*** (14.737)
Contributors <sup>2</sup>		-4.451*** (0.163)	-3.159*** (0.489)	-6.512*** (1.070)
Local Centrality		7.018*** (0.338)	31.678*** (4.612)	40.246*** (4.725)
Global Centrality		2.896*** (0.192)	2.075*** (0.250)	2.155*** (0.250)
Age × Contributors				-38.250*** (5.099)
Age × Contributors <sup>2</sup>				3.625*** (0.693)
Age × Local Centrality				-2.172*** (0.384)
Age × Global Centrality				-0.760*** (0.225)
log-likelihood	-755708	-714054	-680962	-680907
deviance	1511446	1428135	1361952	1361848
AIC	1511477	1428175	1362007	1361906

Linear Latent Mixed Model Regression on 131,109 monthly observations of the natural log of views (divided by 10,000) of 7,535 articles; standard errors in parentheses; continuous variables standardized; coefficient significance calculated using Markov Chain Monte Carlo sampling (10,000 samples) \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ; significance estimates for the random coefficient models based on the likelihood of zero falling outside 5%, 1%, and 0.1% confidence intervals. Restricted maximum likelihood estimation.

644,336 revisions in the Fashion Wikiproject, then built the variables described in Section 3 and analyzed monthly viewing over the same period (December 2007–June 2009). In both samples, we find that the overall modeling strategy has good predictive validity and that the focal network variables improve the fit further. For the Autos sample, we see that the focal network covariates reduce the Mean Absolute Percentage Error (MAPE) from 3.95% in the control model to 2.51% in the network models; for the Fashion sample, the MAPE is reduced from 3.22% to 2.37%. The alternative samples indicate the models can be generalized outside the context of our original sample.

**Table 4 Predictive Validity**

Model	Medicine	Autos	Fashion
<b>Root Mean Squared Error</b>			
Control Model	47.237	62.939	42.282
Network Predictors (Heterogenous)	29.489	40.070	30.005
Network Predictors and Age Interaction (Heterogenous)	29.481	40.082	30.014
<b>Mean Absolute Percentage Error</b>			
Control Model	3.20%	3.95%	3.22%
Network Predictors (Heterogenous)	2.17%	2.51%	2.37%
Network Predictors and Age Interaction (Heterogenous)	2.17%	2.51%	2.37%
<b>Median Absolute Percentage Error</b>			
Control Model	1.48%	1.63%	1.66%
Network Predictors (Heterogenous)	1.03%	0.95%	1.26%
Network Predictors and Age Interaction (Heterogenous)	1.03%	0.95%	1.26%
Observations	131,109	46,936	23,839
Articles	7,535	4,462	1,856

Predictive validity assessed using hierarchical linear models as in Table 3. Because division by zero is undefined, Absolute Percentage Error calculations do not include observations where there was no monthly article viewing.

### 4.3. Hierarchical Bayesian Models

The analysis summarized in Table 3 also indicates that there is heterogeneity in article-level random coefficients for the network covariates. For example, Model 3 indicates dispersion in the authors ( $\sigma = 841.39$ ), local centrality ( $\sigma = 313.21$ ), and global centrality ( $\sigma = 14.66$ ) coefficients. To understand this variation better, we used Bayesian models with Markov Chain Monte Carlo (MCMC) methods. Our empirical context, a hierarchical model with repeated observations, is “ideally suited for MCMC methods” (Rossi and Allenby 2003, p. 312).

Hierarchical Bayesian analysis can be computationally challenging. First, it is not atypical for models to require in excess of ten thousand iterations of burnin with multiple Markov chains before reaching convergence. Second, once convergence is achieved, thousands of iterations are often used to sample coefficients. Our sample size (131,109 revisions nested within 7,535 articles) exacerbates these challenges. To address these challenges, we used a 70-node Linux computing cluster, combining the R statistical software (R Development Core Team 2011) with the JAGS software for Bayesian analysis with Gibbs Sampling using Markov Chain Monte Carlo simulation (Su and Yajima 2011), coding our model in the BUGS declarative language. To reduce overall burnin time, we ran 50 instances of a random subset of 5% of the articles sampling on 40,000 iterations and discarding the first 20,000 iterations for burnin. The computing cluster allowed us to run these instances in parallel. Within the subsets, there was no evidence of failure to converge (mean  $\hat{r} = 1.0077$ ; minimum  $\hat{r} = 1.0005$ ; maximum  $\hat{r} = 1.1859$ ). These analyses were done with multiple starting values and indicate that starting values had little effect on parameter estimates. We averaged the burnin results from the subsets and used these averages as initial parameters for the full sample analysis to reduce overall burnin time. Second, for the full sample analysis, the computing cluster allowed us to run our control, random slope, random intercept, interaction, and all robustness checking models in parallel. We used at least 50,000 iterations for burnin and 3,000 iterations for coefficient sampling. For the full sample models, there was no evidence of failure to converge after the burnin period.

Table 5 summarizes the results of our Bayesian analysis. Model 1 is a control model that includes random intercepts at the article level. Model 2 adds the focal network variables for each monthly observation. As in Table 3, the number of unique contributors has a curvilinear relationship with the content's value. The estimated linear coefficient is  $\beta = 76.12$  and the squared coefficient is  $\beta = -5.87$ , implying an inverted U-shaped relationship with article views. Again, we see that additional contributors working on an article increase its viewership up to a point, then detract from the ability of the article to attract viewers. We also find the estimated coefficient for degree centrality per contributor is  $\beta = 6.33$ ; the coefficient for closeness centrality is  $\beta = 2.91$ . Model 3 allows for additional heterogeneity in the response coefficients by allowing for random slopes and coefficients at the article level. The deviance information criterion (DIC) generalizes the Akaike and Bayesian information criterion for hierarchical modeling; models with smaller values are preferred to those with larger values. In the random coefficients model, the DIC is improved (a reduction of 54,258 in the DIC) indicating a better fit; however, the results are consistent with the random slope model. Again, as in Table 3, content age reduces, but does not completely offset, the effects

**Table 5 Hierarchical Bayesian Model of Article Views**

Variable	Model 1		Model 2		Model 3		Model 4	
Monthly Fixed Effects	indicators		indicators		indicators		indicators	
Article Intercepts	Heterogeneous		Heterogeneous		Heterogeneous		Heterogeneous	
Article Coefficients	Homogeneous		Homogeneous		Heterogeneous		Heterogeneous	
Constant	797.398	(34.655)	845.966	(10.195)	201.715	(22.252)	920.649	(12.215)
Age (ln, years)	28.569	(0.612)	28.707	(0.589)	23.376	(0.525)	−4.001	(1.459)
Length (ln, characters)	10.869	(0.404)	11.453	(0.411)	7.958	(0.650)	7.918	(0.343)
Complexity (ARI)	0.375	(0.201)	0.409	(0.199)	58.764	(0.155)	−0.547	(0.155)
Section Depth	5.182	(0.447)	3.780	(0.480)	2.333	(0.447)	2.083	(0.421)
External References	9.989	(0.488)	3.687	(0.473)	2.227	(0.658)	1.897	(0.543)
Internal Links	7.538	(0.559)	2.684	(0.576)	−21.393	(0.523)	−0.618	(0.450)
Multimedia Content	−2.235	(0.614)	−0.608	(0.612)	−0.035	(0.494)	0.032	(0.482)
Anonymity (percentage)	123.627	(4.124)	88.565	(3.462)	71.604	(3.811)	59.851	(3.865)
Relative Popularity	26.983	(1.660)	14.972	(2.005)	57.165	(4.654)	45.312	(3.093)
Contributors			76.119	(1.346)	373.415	(7.285)	418.422	(9.916)
Contributors <sup>2</sup>			−5.867	(0.127)	−6.658	(0.407)	−17.441	(0.447)
Local Centrality			6.329	(0.354)	14.299	(17.184)	2.897	(14.561)
Global Centrality			2.912	(0.198)	64.872	(2.773)	48.683	(11.401)
Age × Contributors							−56.474	(2.766)
Age × Contributors <sup>2</sup>							8.949	(0.314)
Age × Local Centrality							−0.319	(0.326)
Age × Global Centrality							−1.424	(0.239)
<i>pD</i>	11715.1		10504.9		37875.92		37865.09	
<i>DIC</i>	1404383		1402301		1348043		1347957	
deviance	1392668		1391796		1310167		1310092	

Hierarchical Bayesian analysis of 131,109 monthly observations of the natural log of views (divided by 10,000) of 7,535 articles; at least 60,000 iterations of burnin before sampling based on at least 3,000 iterations; standard deviations in parentheses.

of the network covariates. (Because Bayesian inference is not based on significance, there are no *p*-values to report.) Overall, we find that characteristics of the collaborative network influence the value of user-generated content.

## 5. Discussion

We study the entire compendium of 16,068 Wikipedia articles in the Medicine Wikiproject to determine the effect of collaborative network structure on the value of user-generated content, as measured by viewership. Consistent with our hypotheses, we find that the number of contributors to a content source relates curvilinearly to viewing and that network embeddedness (as measured

through local and global centrality) relates positively to it. We also find that both effects are stronger for newer sources of content than for established ones. Analyses using external samples demonstrate the models accurately predict viewership of articles on different topics (i.e., fashion and autos). As a whole, these results support the core idea that characteristics of the network of contributors and content affect the value of collaborative user-generated content. These results suggest interesting new opportunities and avenues for researchers and managers.

### 5.1. Theoretical Contributions

This article has several implications for theory. First, we demonstrate the need to consider network characteristics of peer-production environments and how these relationships affect the value of user-generated content. Even if a particular collaborative environment is not explicitly social, information and knowledge can still flow from one content source to another as contributors work on multiple sources. Further research should consider how heterogeneous types of relational ties (Borgatti et al. 2009) among content sources affect content creation.

Second, our random coefficient models reveal that network effects on different sources of user-generated content are not equal. We see considerable between-article dispersion indicating that, although there is a strong main effect of network characteristics on viewership, there is also interesting between-article variation to explore. Ongoing research should examine factors that lead to different network effects across different sources of collaborative user-generated content.

Third, our models show that the effects of collaborative inputs on content value depend on the maturity of the content. We find that the impact of both the number of contributors and network embeddedness are stronger for earlier rather than later collaborative efforts. This finding suggests that researchers should consider the state of production when determining the value of collaborative inputs.

### 5.2. Methodological Contributions

This article also makes a number of contributions of broad interest to marketing science. First, we draw attention to the insights that can be obtained by using social network analysis to analyze the complex network that connects objects and people, and what this means for behavior, rather than simply examining social relations. Although the focus of this article is on user-generated content, the approach we use is applicable to many domains of marketing. For example, a two-mode approach could be used to examine how the embeddedness of particular brands within a consumer-brand network impact the success of brand extensions; similarly, such an approach could be used to understand the extent to which products are connected to each other through the consumers that

purchase them and what this implies for cross-promotional strategies. Others could apply these ideas to a wide range of topics in organizational buying settings.

Second, we demonstrate an approach to analyzing databases much larger and comprehensive than those traditionally examined in the marketing literature. The information revolution makes such databases increasingly common but increases in computing power allow researchers to gain insights into this data within a reasonable time frame. Computer clusters like the one we use are increasingly available to university and industry researchers. The ability to tackle these more comprehensive databases allows researchers to better assess the extent to which the effects they observe generalize or vary.

### 5.3. Managerial Contributions

The findings of this article should be of particular interest to managers seeking to cultivate collaborative content. First, the curvilinear relationship between number of contributors and value of collaborative user-generated content suggests that managers should not necessarily pursue a more-is-better strategy toward the number of contributors. Although it is important to generate sufficient participation, once content attains a critical mass of contributors, it may be necessary to redirect new contributors to other content—particularly if there is a virtuous cycle in which increased viewing leads to more contributors. Our data should not be used to predict the optimal number of contributors to a particular content source though, because the optimal number differs by article. Yet we argue that the search for contributors becomes unnecessary or even counterproductive after a point.

Second, our model indicates that all contributors are not equally valuable. Certain contributors with greater experience and knowledge in peer production settings may be more valuable; managers should intentionally seek to recruit top contributors from other collaborative user-generated content sources to work on their important projects. Alternatively they could explicitly establish mechanisms to enable contributors to share best practices for collaboration, such as a forum in which top contributors share their experiences, or encourage contributors to move from one collaborative effort to another to learn and spread these lessons.

Third, the relatively strong influence of the number and network of contributors early in the development of collaborative user-generated content suggests that managers might focus efforts to support the development of content at these early stages of development. Such support might involve seeding early collaborative efforts with experienced contributors, explicitly hired or tasked by the manager to contribute, who can assist with the initial content development. As collaborative user-generated content begins to mature and attract sufficient numbers of contributors to sustain

collaboration, initial contributors may stop contributing in order to focus their efforts on developing new sources of content.

## 6. Limitations and Conclusions

Several limitations of this study suggest the need for further research. First, in this study we only examine the potential for content to flow between nodes (i.e., connections through shared contributors) rather than measuring the actual flow of information and knowledge between articles. Although our approach is consistent with previous applications of SNA (Borgatti et al. 2009), additional research could examine how specific content and process knowledge is transferred through collaboration networks. Also, network measures likely reflect broader characteristics of the network involved in creating user-generated content, such as creators' experience in creating content as well as their content knowledge. Our research nevertheless demonstrates that the two-mode network conceptualization has explanatory power for studying collaborative user-generated content. Future research should investigate applications of this methodology in other contexts (e.g. blog postings, online reviews, locations and the people that connect them).

Second, although we control for many aspects of the user's search for content, there may be lingering issues with endogeneity. Our analysis assesses market value through viewership; there may be other aspects of market value not captured by this measure. Furthermore, underlying topic popularity may not be perfectly captured through Google search results. For instance, we cannot capture whether the Wikipedia article was the first result returned in a Google search at a particular point in time and the Google Insights for Search algorithm may change over time.

Third, the number of contributors and the number of viewers are correlated processes that follow the same lifecycle and diffusion process. Although we incorporated article age directly and through interactions with out focal variables, lifecycle effects may be more complex than the log-linear relationship models we use. Future research might investigate how these lifecycle effects influence collaborative user-generated content, both qualitatively and empirically.

In conclusion, this article represents an initial attempt to examine how characteristics of the networks involved in creating collaborative user-generated content affect the content's market value. We find that more contributors improve viewership of user-generated content, but only up to a point. Too many contributors complicate the development efforts and reduce viewership. By conceptualizing Wikipedia as a two-mode network of content and contributors, we find clear evidence of the content-contributor network's effect on viewership, suggesting that it is a mistake to view a given source of user-generated content as independent from other sources of content. Rather,

content is influenced not only by those who create it, but also through connections to other contributors and other, sometimes quite distal, content. Finally, we find that the effect of collaboration changes over time, with newer user-generated content being relatively more sensitive to network characteristics than more mature content. Understanding these effects is particularly important given the increasingly collaborative nature of user-generated content and the growing interest by firms in generating revenue from such content.

### Acknowledgments

We are grateful to Hal Varian and Google, Inc., for providing extensive Google Insights for Search information. Gerald Kane acknowledges funding for this research from the National Science Foundation (CAREER 0953285). The authors also thank Steven C. Lacey and Zoey Chen for their assistance with data collection and analysis.

### References

- Adler, P. S., S.-W. Kwon. 2002. Social capital: Prospects for a New Concept. *Academy of Management Rev.* **27**(1) 17–40.
- Alba, J. W., J. W. Hutchinson. 1987. Dimensions of Consumer Expertise. *J. of Consumer Res.* **13**(March) 411–454.
- Aral, S., M. Van Alstyne. forthcoming. Networks, information and brokerage: The diversity-bandwidth tradeoff. *Amer. J. Sociology* .
- Asvanund, A., K. Clay, R. Krishnan, M. D. Smith. 2004. An Empirical Analysis of Network Externalities in Peer-to-Peer Music-Sharing Networks. *Inform. Systems Res.* **15**(2) 155–174.
- Berger, J., C. Heath. 2007. Where Consumers Diverge from Others: Identity Signaling and Product Domains. *J. Consumer Res.* **34**(2) 121–134.
- Berger, J., K. Milkman. 2011. Virality: What Gets Shared and Why? M. C. Campbell, J. J. Inman, R. Pieters, eds., *Advances in Consumer Res.*. Association for Consumer Res., 118–119. Vol. 37.
- Borgatti, S. P. 2005. Centrality and Network Flow. *Social Networks* **27**(1) 55–71.
- Borgatti, S. P., M. G. Everett. 1997. Network Analysis of 2-mode Data. *Social Networks* **19**(3) 243–269.
- Borgatti, S. P., A. Mehra, D. J. Brass, G. Labianca. 2009. Network Analysis in the Social Sciences. *Science* **323**(5916) 892–895.
- Brandes, U., P. Kenis, J. Lerner, D. van Raaij. 2009. Network Analysis of Collaboration Structure in Wikipedia. *Proceedings of the 18th International Conference on the World Wide Web*. ACM, 731–740.
- Brooks, F. P. 1975. *The Mythical Man-Month: Essays on Software Engineering*. Addison-Wesley, Reading, MA.

- Brown, J. J., P. H. Reingen. 1987. Social Ties and Word-of-Mouth Referral Behavior. *J. of Consumer Res.* **14**(3) 350–362.
- Butler, B. S. 2001. Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures. *Inform. Systems Res.* **12**(4) 346–362.
- Capocci, A., V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, G. Caldarelli. 2006. Preferential Attachment in the Growth of Social Networks: The Internet Encyclopedia Wikipedia. *Physical Rev. E* **74**(3) 036116.
- Carlile, P. R., E. S. Rebentisch. 2003. Into the Black Box: The Knowledge Transformation Cycle. *Management Sci.* **49**(9) 1180–1195.
- Chevalier, J. A., D. Mayzlin. 2006. The Effect of Word of Mouth on Sales: Online Book Reviews. *J. of Marketing Res.* **43**(August) 345–354.
- Clauson, K. A., H. H. Polen, M. N. K. Boulos, J. H. Dzenowagis. 2008. Scope, Completeness, and Accuracy of Drug Information in Wikipedia. *The Annals of Pharmacotherapy* **42**(12) 1814–1821.
- Constant, D., L. Sproull, S. Kiesler. 1996. The Kindness of Strangers: The Usefulness of Electronic Weak Ties for Technical Advice. *Organ. Sci.* **7**(2) 119–135.
- Costenbader, E., T.W. Valente. 2003. The Stability of Centrality Measures when Networks are Sampled. *Social Networks* **25** 283–307.
- Denning, P., J. Horning, D. Parnas, L. Weinstein. 2005. Wikipedia Risks. *Communications of the ACM* **48**(12) 152.
- Duan, W., B. Gu, A. B. Whinston. 2008. The Dynamics of Online Word-of-Mouth and Product Sales—An Empirical Investigation of the Movie Industry. *J. of Retailing* **84**(2) 233–242.
- Elsner, M. K., O. P. Heil, A. R. Sinha. 2009. Spreading the Word: Assessing the Factors that Determine the Popularity of User-Generated Content. *Paper presented at the Emergence and Impact of User-Generated Content*. Philadelphia, PA.
- Espinosa, J. A., S. A. Slaughter, R. E. Kraut, J. D. Herbsleb. 2007. Team Knowledge and Coordination in Geographically Distributed Software Development. *J. of Management Inform. Sys.* **24**(1) 135–169.
- Faust, K. 1997. Centrality in Affiliation Networks. *Social Networks* **19**(2) 157–191.
- Ferguson, T., G. Frydman. 2004. The First Generation of e-Patients. *British Medical J.* **328**(7449) 1148–1149.
- Forman, C., A. Ghose, B. Wiesenfeld. 2008. Examining the Relationship between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. *Inform. Systems Res.* **19**(3) 291–313.
- Foutz, N. Z., W. Jank. 2010. Prerelease Demand Forecasting for Motion Pictures using Functional Shape Analysis of Virtual Stock Markets. *Marketing Sci.* **29**(3) 568–579.
- Fox, S., S. Jones. 2009. The Social Life of Health Information. Tech. rep., Pew Research Center, Washington, D.C.

- Freeman, L.C. 1979. Centrality in Social Networks Conceptual Clarification. *Social Networks* **1**(3) 215–239.
- Frels, J. K., T. Shervani, R. K. Srivastava. 2003. The Integrated Networks Model: Explaining Resource Allocations in Network Markets. *J. of Marketing* **67**(1) 29–45.
- Frenzen, J., K. Nakamoto. 1993. Structure, Cooperation, and the Flow of Market Information. *J. of Consumer Res.* **20**(December) 360–375.
- Frenzen, J. K., H. L. Davis. 1990. Purchasing Behavior in Embedded Markets. *J. of Consumer Res.* **17**(June) 1–12.
- Godes, D., D. Mayzlin. 2004. Using Online Conversations to Measure Word of Mouth Communication. *Marketing Sci.* **23**(4) 545–560.
- Granovetter, M. 1985. Economic Action and Social Structure: The Problem of Embeddedness. *Am. J. of Sociology* **91**(3) 481–510.
- Gregan-Paxton, J., D. R. John. 1997. Consumer Learning by Analogy: A Model of Internal Knowledge Transfer. *J. of Consumer Res.* **24**(3) 266–284.
- Grewal, R., G. L. Lilien, G. Mallapragada. 2006. Location, Location, Location: How Network Embeddedness Affects Project Success in Open Source Systems. *Management Sci.* **52**(7) 1043–1056.
- Gu, F. F., K. Hung, D. K. Tse. 2008. When Does Guanxi Matter? Issues of Capitalization and its Dark Sides. *J. of Marketing* **72**(4) 12–28.
- Gulati, R., M. Gargiulo. 1999. Where do Interorganizational Networks Come from? *Am. J. of Sociology* **104**(5) 1439–1493.
- Hansen, M. T., M. R. Haas. 2001. Competing for Attention in Knowledge Markets: Electronic Document Dissemination in a Management Consulting Company. *Admin. Sci. Quart.* **46**(1) 1–28.
- Hansen, M. T., M. L. Mors, B. Lovas. 2005. Knowledge Sharing in Organizations: Multiple Networks, Multiple Phases. *Academy of Management J.* **48**(5) 776–793.
- Haveman, H. A. 1993. Organizational Size and Change—Diversification in the Savings and Loan Industry after Deregulation. *Administrative Sci. Quarterly* **38**(1) 20–50.
- Hiltz, S.R., M Turoff. 1985. Structuring Computer-mediated Communication Systems to Avoid Information Overload. *Communications of the ACM* **28**(7) 680–699.
- Iacobucci, D., N. Hopkins. 1992. Modeling Dyadic Interactions and Networks in Marketing. *J. of Marketing Res.* **29**(1) 5–17.
- Iyengar, S. S., M. R. Lepper. 2000. When Choice is Demotivating: Can One Desire Too Much of a Good Thing? *J. of Personality and Social Psychology* **79**(6) 995–1006.
- Jones, G., G. Ravid, S. Rafaeli. 2004. Information Overload and the Message Dynamics of Online Interaction Spaces: A Theoretical Model and Empirical Exploration. *Inform. Systems Res.* **15** 194–210.

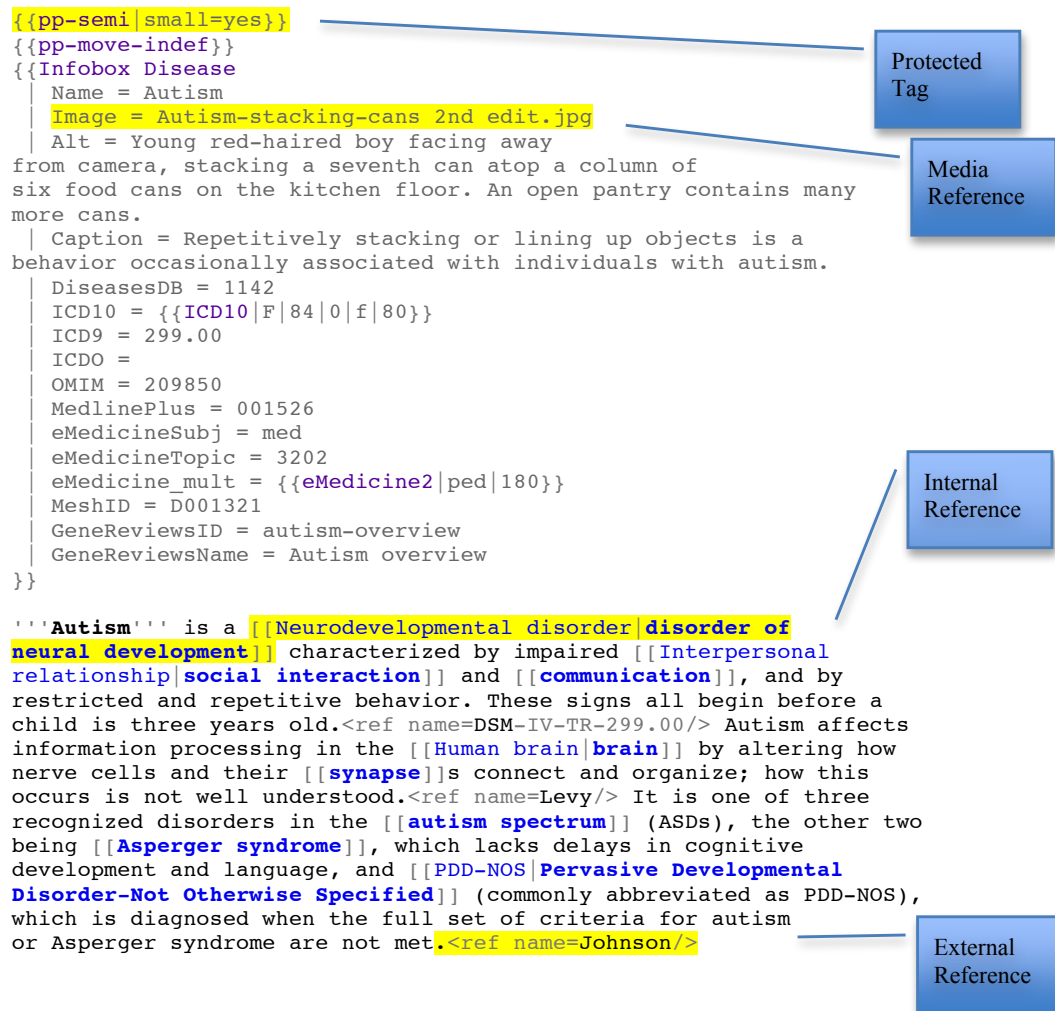
- Kane, G. C. 2011. A Multi-Method Study of Information Quality in Wiki Collaboration. *ACM Transactions on Management Inform. Systems* **1**(1) forthcoming.
- Kane, G. C., R. G. Fichman, J. Gallagher, J. Glaser. 2009. Community Relations 2.0. *Harvard Bus. Rev.* **87**(11) 45–50.
- Kane, G.C., M Alavi. 2007. Information Technology and Organizational Learning: An Investigation of Exploration and Exploitation Processes. *Organization Sci.* **18**(5) 796–812.
- Khullar, O., Y. L. Colson. 2009. The Underfunding of Lung Cancer Research. *J. Thoracic Cardiovascular Surgery* **138**(2) 275.
- Kittur, A., R. E. Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. ACM, San Diego, 37–46.
- Kozinets, R. V., A. Hemetsberger, H. J. Schau. 2008. The Wisdom of Consumer Crowds: Collective Innovation in the Age of Networked Marketing. *J. of Macromarketing* **28**(4) 339–354.
- Kriplean, T., I. Beschastnikh, D. W. McDonald. 2008. Articulations of Wikiwork: Uncovering Valued Work in Wikipedia through Barnstars. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. ACM, San Diego, 47–56.
- Kuk, G. 2006. Strategic Interaction and Knowledge Sharing in the KDE Developer Mailing List. *Management Sci.* **52**(7) 1031–1042.
- Lave, J., E. Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, New York.
- Levinthal, D. A., J. G. March. 1993. The Myopia of Learning. *Strategic Management J.* **14** 95–95.
- Li, C., J. Bernoff. 2008. *Groundswell: Winning in a world transformed by social technologies*. Harvard Business School Press, Boston.
- Li, X., L. M. Hitt. 2008. Self-selection and Information Role of Online Product Reviews. *Inform. Systems Res.* **19**(4) 456–474.
- Lin, N. 1982. *Social Resources and Instrumental Action*. Sage, Beverly Hills, CA.
- Lovelace, K, D. L. Shapiro, L. R. Weingart. 2001. Maximizing Cross-Functional New Product Teams' Innovativeness and Constraint Adherence: A Conflict Communications Perspective. *Acad. of Management J.* **44**(4) 779–793.
- Lurie, N. H. 2004. Decision Making in Information-Rich Environments: The Role of Information Structure. *J. of Consumer Res.* **30** 473–486.
- Madey, V., G., R. T. Freeh. 2004. *Modeling the F/OSS community: A Quantitative Investigation*. Idea Publishing, Hersey, PA.

- Mallapragada, G., R. Grewal, G. Lilien. 2008. Born to Win? How Foundational Network Structure Affects the Success of Open Source Development Projects.
- Manchanda, P., Y. Xie, N. Youn. 2008. The Role of Targeted Communication and Contagion in Product Adoption. *Marketing Sci.* **27**(6) 961–976.
- March, J. G. 1991. Exploration and Exploitation in Organizational Learning. *Organization Sci.* **2**(1) 71–87.
- Martins, L. L., L. L. Gilson, M. T. Maynard. 2004. Virtual Teams: What Do We Know and Where Do We Go From Here? *J. of Management* **30**(6) 805–835.
- McAfee, A. P. 2007. Wikipedia (B) (Case 608-066). Harvard Business School.
- Miller, C. C. 2009. Ad Revenue on the Web? No Sure Bet. *New York Times* (May 25) B1.
- Moe, W. W., M. Trusov. 2011. The Value of Social Dynamics in Online Product Ratings Forums. *J. of Marketing Res.* **48**(3) 444–456.
- Monge, P. R., N. S. Contractor. 2003. *Theories of Communication Networks*. Oxford University Press, Oxford; New York.
- Nahapiet, J., S. Ghoshal. 1998. Social Capital, Intellectual Capital, and the Organizational Advantage. *Academy of Management Rev.* **23**(2) 242–266.
- Naik, P., M. Wedel, L. Bacon, A. Bodapati, E. Bradlow, W. Kamakura, J. Kreulen, P. Lenk, D. M. Madigan, A. Montgomery. 2008. Challenged and Opportunities in High-Dimensional Choice Data Analyses. *Marketing Letters* **19**(3) 201–213.
- Oestreicher-Singer, G., J. Goldenberg, S. Reichman. 2009. The Quest for Content: The Integration of Product and Social Networks in UGC Environments. *Emergence and Impact of User-Generated Content*. Philadelphia, PA.
- Oh, H., M.-H. Chung, G. Labianca. 2004. Group Social Capital and Group Effectiveness: The Role of Informal Socializing Ties. *The Academy of Management Journal* **47**(6) 860–875.
- Oh, W., S. Jeon. 2007. Membership Herding and Network Stability in the Open Source Community: The Ising Perspective. *Management Sci.* **53**(7) 1086–1101.
- Phillips, L. E. 2007. Pharmaceutical Marketing Online: Stuck in Web 1.5. Available at [http://www.emarketer.com/Reports/All/Emarketer\\_2000434.aspx](http://www.emarketer.com/Reports/All/Emarketer_2000434.aspx). Accessed on January 8, 2010.
- R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. Available at <http://www.R-project.org/>. Accessed May 1, 2011.
- Ransbotham, S., G. C. Kane. 2011. Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia. *MIS Quarterly* **35**(3) 613–627.
- Reagans, R., B. McEvily. 2003. Network Structure and Knowledge Transfer: The Effects of Cohesion and Range. *Admin. Sci. Quart.* **48**(2) 240–267.

- Rindfleisch, A., C. Moorman. 2001. The Acquisition and Utilization of Information in New Product Alliances: A Strength-of-Ties Perspective. *J. of Marketing* **65**(2) 1–18.
- Rossi, P.E., G.M. Allenby. 2003. Bayesian Statistics and Marketing. *Marketing Sci.* **22**(3) 304–328.
- Schlosser, A. E. 2003. Experiencing Products in the Virtual World: The Role of Goal and Imagery in Influencing Attitudes versus Purchase Intentions. *J. of Consumer Res.* **30**(September) 184–198.
- Schlosser, A. E. 2005. Posting versus Lurking: Communicating in a Multiple Audience Context. *J. of Consumer Res.* **32**(2) 260–265.
- Schlosser, A. E. 2007. The Persuasiveness of Positive Online Reviews: Consumers Intuitive Theories about Evaluative-Cognitive Consistency. D. L. Hoffman, E. J. Johnson, eds., *Paper Presented at the Association for Consumer Research Pre-Conference Consumers Online: Ten Years Later*. Memphis, TN.
- Sia, C. L., B. C. Y. Tan, K. K. Wei. 2002. Group Polarization and Computer-Mediated Communication: Effects of Communication Cues, Social Presence, and Anonymity. *Inform. Systems Res.* **13**(1) 70–90.
- Singh, P. V., Y. Tan, V. Mookerjee. 2011. Network Effects: The Influence of Structural Social Capital on Open Source Project Success. *MIS Quarterly* forthcoming.
- Smith, E. A., R. J. Senter. 1967. *Automated Readability Index*. Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH.
- Spence, M. T., M. Brucks. 1997. The Moderating Effects of Problem Characteristics on Experts' and Novices' Judgments. *J. of Marketing Res.* **34**(May) 233–247.
- Su, Y.-S., M. Yajima. 2011. R2jags: A Package for Running JAGS from R. Available at <http://CRAN.R-project.org/package=R2jags>. Accessed May 1, 2011.
- Tuckman, B. W. 1965. Developmental Sequence in Small Groups. *Psychological Bulletin* **63**(6) 384–399.
- Tuli, K. R., A. K. Kohli, S. G. Bharadwaj. 2007. Rethinking Customer Solutions: From Product Bundles to Relational Processes. *J. of Marketing* **71**(3) 1–17.
- Uzzi, B. 1997. Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness. *Administrative Sci. Quarterly* **42**(1) 35–67.
- Wasserman, S., K. Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, New York.
- Weiss, A. M., N. H. Lurie, D. J. MacInnis. 2008. Listening to Strangers: Whose Responses are Valuable, How Valuable are They, and Why? *J. of Marketing Res.* **45**(August) 425–436.
- Wikipedia. 2010. Wikipedia: Article Size. Available at [http://en.wikipedia.org/wiki/Article\\_length](http://en.wikipedia.org/wiki/Article_length). Accessed January 6, 2010.
- Zlatic, V., M. Božičević, H. Štefančić, M. Domazet. 2006. Wikipedias: Collaborative Web-Based Encyclopedias as Complex Networks. *Physical Rev. E* **74**(1) 016115.
- Zukin, S., P. DiMaggio. 1990. *Structures of Capital: The Social Organization of the Economy*. Cambridge University Press, New York.

## Appendix. Sample Article

This appendix uses the Wikipedia Article on Autism to show selected information gleaned from article source code and revision history.



name=Rutter/> the vaccine hypotheses are biologically implausible and lack convincing scientific evidence.<ref name=vaccines/> The [prevalence](#) of autism is about 1–2 per 1,000 people; the prevalence of ASD is about 6 per 1,000, with about four times as many males as females. The number of people diagnosed with autism has increased dramatically since the 1980s, partly due to changes in diagnostic practice; the question of whether actual prevalence has increased is unresolved.<ref name=Newschaffer/>

Parents usually notice signs in the first two years of their child's life.<ref name=CCD/> The signs usually develop gradually, but some autistic children first develop more normally and then [Regressive autism|regress](#).<ref name=Stefanatos/> Although early behavioral or cognitive intervention can help autistic children gain self-care, social, and communication skills, there is no known cure.<ref name=CCD/> Not many children with autism live independently after reaching adulthood, though some become successful.<ref name=Howlin/> An [Sociological and cultural aspects of autism|autistic culture](#) has developed, with some individuals seeking a cure and others believing autism should be accepted as a difference and not treated as a disorder.<ref name=Silverman/>

Section  
Depth

#### ==Characteristics==

Autism is a highly variable [neurodevelopmental disorder](#)<ref name=Geschwind/> that first appears during infancy or childhood, and generally follows a steady course without [Remission \(medicine\)|remission](#).<ref name=ICD-10-F84.0/> Overt symptoms gradually begin after the age of six months, become established by age two or three years,<ref>{{vcite journal |author=Rogers SJ |title=What are infant siblings teaching us about autism in infancy? |title.= |journal=Autism Res |volume=2 |issue=3 |pages=125–37 |year=2009 |pmid=19582867 |doi=10.1002/aur.81 |pmc=2791538 }}</ref> and tend to continue through adulthood, although often in more muted form.<ref name=Rapin/> It is distinguished not by a single symptom, but by a characteristic triad of symptoms: impairments in social interaction; impairments in communication; and restricted interests and repetitive behavior. Other aspects, such as atypical eating, are also common but are not essential for diagnosis.<ref name=Filipek/> Autism's individual symptoms occur in the general population and appear not to associate highly, without a sharp line separating pathologically severe from common traits.<ref name=London/>

External  
Reference

#### ===Social development===

Social deficits distinguish autism and the related [autism spectrum disorder](#)s (ASD; see '[\[#Classification/Classification\]](#)') from other developmental disorders.<ref name=Rapin/> People with autism have social impairments and often lack the intuition about others that many people take for granted. Noted autistic [Temple Grandin](#) described her inability to understand the [social communication](#) of [neurotypical](#)s, or people with normal [neural development](#), as leaving her feeling "like an anthropologist on Mars".<ref>{{vcite book |title=[An Anthropologist

Section  
Depth

## REVISION HISTORY OF AUTISM

From Wikipedia, the free encyclopedia  
[View logs for this page](#)

Browse history From year (and earlier):  From month (and earlier):

Tag filter:   Deleted only

For any version listed below, click on its date to view it. For more help, see [Help:Page history](#) and [Help:Edit summary](#).

External tools: [Revision history statistics](#) · [Revision history search](#) · [Number of watchers](#) · [Page view](#)

(cur) = difference from current version, (prev) = difference from preceding version, m = [minor edit](#), → = [section edit](#), ← = [automatic edit summary](#)  
(latest | [earliest](#)) View (newer 50 | [older 50](#)) ([20](#) | [50](#) | [100](#) | [250](#) | [500](#))

Author  
ID, Date,  
Edit  
Made

• (cur | prev)   [10:56, 7 September 2010 Kww \(talk | contribs\)](#) (111,683 bytes) (*Pending changes trial is complete*) ([undo](#))

• (cur | prev)   [10:53, 7 September 2010 Kww \(talk | contribs\)](#) m (111,661 bytes) (*Reset pending changes settings for Autism: Pending changes trial complete, most IP edits were vandalism*) ([undo](#))

• (cur | prev)   [10:53, 7 September 2010 Kww \(talk | contribs\)](#) m (111,661 bytes) (*Changed protection level of Autism: Pending changes trial complete, most IP edits were vandalism [edit=autoconfirmed] (expires 14:53, 7 November 2010 (UTC)) [move=sysop] (indefinite))*) ([undo](#))

• (cur | prev)   [16:08, 5 September 2010 Jfdwolff \(talk | contribs\)](#) (111,661 bytes) (*doesn't work, try the template talk page for details*) ([undo](#))

• (cur | prev)   [00:25, 22 August 2010 90.204.224.53 \(talk\)](#) (110,923 bytes) (*Accepted, not "tolerated". Nobody believes autism should be "tolerated" despite, by implication of the choice of word, being somehow a blight on society, even if...)*) ([undo](#))

Article  
Length

Anonymous  
Contributor