Working Paper

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 Lurie
College of Management, Georgia Institute of Technology, Atlanta, GA 30332, nicholas.lurie@mgt.gatech.edu

User-generated content increasingly is created through the collaborative efforts of multiple individuals. Characteristics of the network associated with the creation of collaborative content should therefore influence content value. A social network analysis, applied to Wikipedia's Medicine Wikiproject, reveals a curvilinear relationship between the number of distinct contributors to user-generated content and viewership. Globally central content—characterized by connections to more prominent collaborative content in the overall network—generates greater viewership. Contrary to previous theory, locally central content—characterized by greater intensity of work by contributors to multiple content sources—is negatively associated with viewership. In addition, network effects are stronger for newer collaborative user-generated content. A recursive relationship between contribution and viewership activity suggests a virtuous cycle between the value of—and contribution to—user-generated content, but this dynamic matures and stabilizes over time. Finally, effects of network characteristics on value differ for the most and least viewed content. These findings have implications for fostering collaborative user-generated content.

Key words: user-generated content; social networks; information value; wiki; hierarchical clustering; social network analysis

1. Introduction

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 objects and spaces and perform shared tasks; 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.

In this article, we argue that the characteristics of networks that connect the creators and output of collaborative user-generated content (UGC) are important predictors of the popularity, and therefore value, of this content to consumers. In particular, through their involvement with multiple collaborative projects, contributors develop content creation and collaboration skills that they apply to other projects on which they work. These skills and knowledge are reflected in emergent networks among individuals and content sources. We propose that three dimensions of this network affect the value of collaborative user-generated content: (1) the size of the network (i.e., the number of distinct contributors), (2) the centrality of the article in the local network (i.e., the number of and intensity with which collaborators work on other sources), and (3) the centrality of the article in the global network (i.e., the relative importance of other user-generated content on which collaborators work). The effects of these network characteristics on increasing the market value of collaborative user-generated content should be greater for newer relative to older content, because collaborative content is likely to stabilize over time.

We test our hypotheses by analyzing Wikipedia's Medicine Wikiproject and examining how network characteristics affect the market value of user-generated content. Our results demonstrate a curvilinear relationship between the number of distinct contributors to an article and its market value. We also find that an article's centrality in the global network is associated with more valuable content. Contrary to our expectations, we see some evidence that article centrality in the local network is associated with less valuable content. These network effects are stronger for newer user-generated content, with the exception of position in the local network, which is unaffected by age. 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

Growing research offers greater understanding of user-generated content and its implications for marketing by showing, for example, that product reviews 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). Researchers also have also shown that the relative influence of user-generated content depends on the characteristics of the content, characteristics of the creators of content, and their interactions (Berger and Milkman 2009, Constant et al. 1996, Weiss et al. 2008). For example, 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. 2008a,b, 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).

Prior research on network effects has examined 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 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), and sociologists have considered the value of networks to individuals and groups (Borgatti et al. 2009). However, research on the products of collaborative networks and how network characteristics affect their creation is limited. Because collaborative user-generated content requires a complex network of contributors and content, network characteristics may help predict its value to consumers. Figure 1 depicts potential relationships among creators of collaborative user-generated content and the content they create to illustrate the network structures we examine.

In Figure 1, circles represent content sources, whereas individual contributors are represented by numbered squares. *Network size* captures the number of distinct individuals who contribute to the focal content (F); in this example, there are four distinct contributors (1–4). *Local network centrality* measures the collaborative activity of these contributors to other content sources: Contributor 1 contributes to no other content sources, Contributor 2 contributes to one other content source, Contributor 3 contributes to five other sources, and Contributor 4 to three. *Global network centrality* assesses the level of collaborative activity across the content sources to which the focal source is connected. Although Contributor 2 only works on one other content source, that

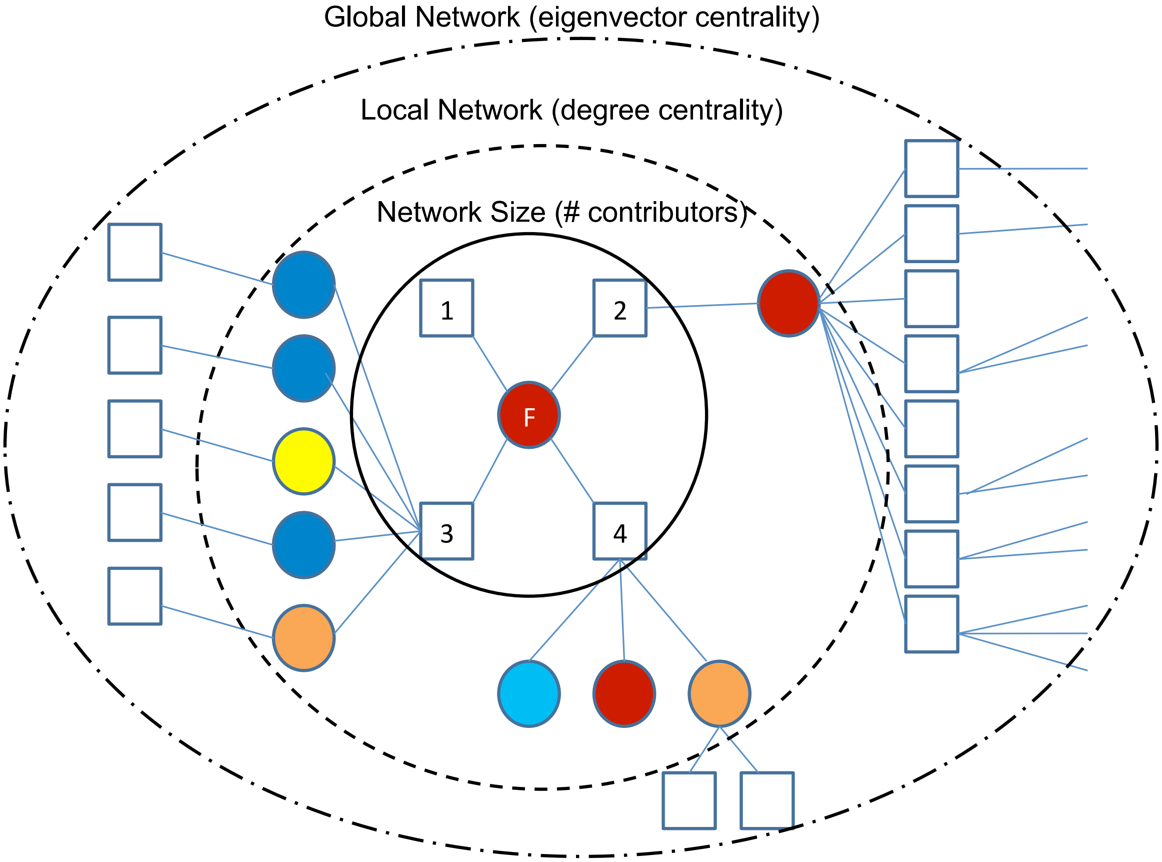

Global Network (eigenvector centrality)
Local Network (degree centrality)
Network Size (# contributors)
1
2
F
3
4

Figure 1 **Example of a Collaborative User-Generated Content Network**

particular content source exhibits a relatively high level of collaborative activity (eight additional contributors) compared with the level of activity of content sources worked on by Contributors 1, 3, and 4. Global network centrality also includes the collaborative activity of those eight additional contributors, the articles they worked on, the collaborators for those articles, and so on, accounting for the influence of all contributors and content sources in the network. We posit that each of these network characteristics affects the value of collaborative user-generated content.

2.1. Network Size

With user-generated content, the network size (i.e., number of contributors to particular content) can vary. Research on prediction markets, virtual teams, and social networks (Constant et al. 1996, Foutz and Jank 2010, Martins et al. 2004) suggests that the quality of aggregate information, number of ideas generated, and likelihood of a valuable answer increases with the number of participants. Because each contributor represents a unique source of knowledge, additional contributors can identify important missing information or factual inaccuracies. The more people who contribute, the more thorough and high-value information the content contains.

Yet additional contributors may be valuable only up to a point. Too much available information leads to information overload, making it difficult to decide what information is most valuable and salient. Less content is potentially more valuable in some settings, because the costs associated with finding the most valuable content decrease (Hansen and Haas 2001). New ideas have limited marginal value after a certain point, because they are redundant, and it is increasingly costly to filter out bad ideas (Butler 2001, Constant et al. 1996). For collaborative user-generated content, more contributors also increase coordination costs and development time and possibly decrease the quality of the final product (Brooks 1975). Although larger and more diverse teams can enhance creativity, an increasing diversity of perspectives makes it harder for teams to reach consensus (Lovelace et al. 2001). 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 that content.*

2.2. Local Network Centrality

When people work on other sources of user-generated content, they likely develop expertise in content creation. Greater expertise allows them to 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 make the transferred content more informative (Gregan-Paxton and John 1997). In particular, they learn subject matter information, how to be more effective collaborators, and the reputation of other collaborators in a community, which facilitates their creation of more valuable content. *Local network centrality* thus captures the intensity of contributions a group of collaborators makes to other sources of user-generated content.

Local network centrality may increases content value in two ways. Contributors can directly and intentionally transfer information and knowledge from one source to another or they may possess content, process, and reputational knowledge that influence the development of the sources they work on. Regardless, local network centrality should be positively associated with the market value of user-generated content.

Hypothesis 2. *The market value of collaborative user-generated content relates positively to the local network centrality of the content.*

2.3. Global Network Centrality

In the same way that collaborators' ability to create valuable content depends on the intensity of their contributions to other content sources, the particular content sources on which they choose to work should affect the market value of the content they create. In most user-generated content settings, collaborative activity involves a relatively small number of content sources (Barabasi 2003). Most of the millions of articles on Wikipedia are article stubs that involve minimal collaboration; only a fraction earns recognition as featured (best) articles[1]. Although a user may gain experience by working intensely on multiple content sources, far greater subject, process, and reputational insight likely comes from collaborating on highly active sources. *Global network centrality* therefore captures the intensity of collaborative activity on all directly and indirectly connected content and should relate positively to the market value of collaborative user-generated content. Again, two mechanisms might drive this effect: Contributors recognize the most valuable content sources and seek to transfer information from the richest sources or contributors are particularly prolific, skilled, or influential, and content sources become popular or active due to their contributions.

Hypothesis 3. *The market value of collaborative user-generated content relates positively to the global network centrality of the content.*

2.4. Content Age

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 places functional limits on the length of an article (Wikipedia 2010). Groups generally move through distinct phases of collaboration, and later collaboration is distinct from that which occurred earlier (c.f., Gersick 1988, Tuckman 1965). Early collaboration likely is dedicated to brainstorming and generating new content, because no content exists; middle phases involve organizing disparate ideas from a generated critical mass; and later phases entail maintaining and defending the content amidst ongoing collaboration, after the group has developed the organized whole (Kane et al. 2009b). These characteristics 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 more difficult to change, the effects of network size, local network centrality, and global network centrality on market value should decline.

Hypothesis 4. *The impact of (a) network size, (b) local network centrality, and (c) global network centrality on the market value of collaborative user-generated content declines with content age.*

[1] See http://en.wikipedia.org/wiki/Wikipedia:1.0#Statistics.

3. Research Method and Setting

Social network analysis (SNA) is an insightful approach for studying collaborative environments (Borgatti and Cross 2003, Cross and Prusak 2002, Cummings 2004, Reagans and McEvily 2003). 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).

Yet SNA can be applied more broadly to study interactions that are not explicitly social. In such an analysis, the nodes can be any entity in a network, and a tie can be any connection between them. For example, SNA methods have been used to investigate airline networks (airports and routes; Amaral et al. 2000), the working of the human brain (neurons and synapses; Newman et al. 2006), shared contributions to open source software projects (projects and developers; Grewal et al. 2006, Oh and Jeon 2007), the structure of the Internet (webpages and hyperlinks; Wellman 2001), blog networks (blogs and trackbacks; Wattal et al. 2010), and purchase patterns in online recommendation networks (i.e., "customers who bought this, also bought…"; Carmi et al. 2009, Oestreicher-Singer and Sundararajan 2010). Even Google's famed PageRank algorithm uses measures based in SNA to prioritize results (Brin and Page 1998).

One classic approach to SNA employs a two-mode network (Wasserman and Faust 1994) with two distinct types of nodes, such that one type of node represents a tie that connects the other type. A paradigmatic example is Davis et al. (1941) study of southern women, in which social parties were nodes, connected by the women who attended them. Two-mode networks characterize various types of shared interactions, including the opinions of Supreme court justices, bills proposed by lawmakers, structures of corporate boards, and the contributions of scientific communities (contributors and papers; Carrington et al. 2005). 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). We employ a two-mode network analysis to examine relationships among different creators and sources of collaborative user-generated content, with creators as the ties that connect different content sources. This approach is consistent with the idea that contributors transfer information from one content source to another as they work on multiple sources; the underlying knowledge of a contributor affects all the articles to which he or she contributes.

We test our hypotheses by analyzing the effects of the network structure on the viewership of articles in the Medicine Wikiproject. To confirm the robustness of our findings, we use a holdout sample, in which parameter estimates from the first ten months of data are used to predict monthly article viewing for the remaining nine months. As an additional test, we use parameter estimates from the Medicine Wikiproject to estimate viewership for the fashion and auto Wikiprojects, and compare predicted to actual viewership for these additional samples. To assess the extent to which the effects of network characteristics depend on market structure, as revealed through consumer behavior (Kohli and Jedidi 2007), we divide our original sample into five clusters of articles, each of which exhibits similar viewing patterns, and examine effects for each cluster individually.

3.1. Research Setting

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.

We assess how network characteristics affect the value of collaborative user-generated content by examining the relationships among 16,068 Wikipedia articles in the Medicine Wikiproject (i.e., all articles in this project during the study period) and the creators of these articles. 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 the network and the market value of information created by the network.

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, which 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. (2009a). Previous studies have affirmed the quality of medical information on Wikipedia (Clauson et al. 2008, Devgan et al. 2007, Laurent and Vickers 2009), 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. 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 article stubs)[2]. 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.

[2] See http://stats.wikimedia.org/EN/BotActivityMatrix.htm.

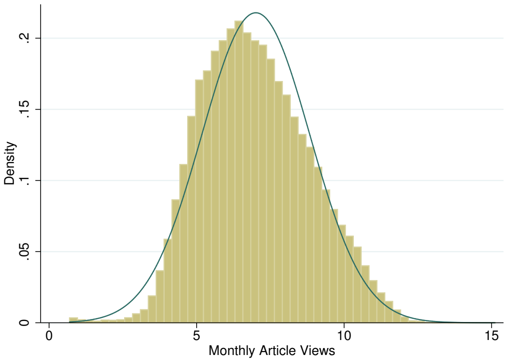


Figure 2 Histogram of Article Views (ln)

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 × 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. Figure 2 depicts a histogram of this transformation, along with a fitted normal distribution curve. A Shapiro-Francia test fails to reject the null hypothesis that the distribution is normal ($W' = 0.9914$, $p < 0.5$).

Because anyone can edit content in Wikipedia, an endogenous relationship might exist between viewership and contribution activity. People might seek information on Wikipedia, realize they possess the knowledge and interest to improve the content, and then decide to contribute. Those new contributors improve the article, which makes the content more valuable and attracts additional viewers. To address this potential endogeneity we employ a three-stage least squares (3SLS) regression and identify the system of equations through the variable *protected*, indicating articles protected from changes by Wikipedia administrators (i.e., to prevent vandalism). Protection should reduce the number of contributors to an article since fewer people can contribute. Protection does not affect the ability to view the article, so protection status affects editing activity but not viewing. Thus, for article i during period t, we use 3SLS to estimate the following equations simultaneously:

$$ln(views_{i,t}) = \beta_1\, X_{i,t} + \beta\, ln(views_{i,t-1}) + \beta\, authors_{i,t} + \epsilon_1 \tag{1}$$

$$authors_{i,t} = \beta_2\, X_{i,t} + \beta\, authors_{i,t-1} + \beta\, ln(views_{i,t}) + \beta\, protected_{i,t} + \epsilon_2 \tag{2}$$

where X is a vector of article covariates.

3.4. Independent Variables

3.4.1. Network size. We measure network size as the number of distinct contributors to user-generated content. Technically, it is the degree centrality of the untransformed bimodal matrix. To avoid confusion with our other centrality measures, we use the term “network size.”

3.4.2. Local network centrality. We measure this variable as the degree centrality in the incidence matrix of contributors and content sources, that is, as the number and strength of direct connections possessed by a node (Scott 2000, 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 scaling purposes, we divide local network centrality by 1000.

3.4.3. Global network centrality. We measure global network centrality as eigenvector centrality in the incidence matrix. Eigenvector centrality summarizes the centrality of a node in the global network of all nodes and ties that compose the network (Scott 2000, Wasserman and Faust 1994). It reflects the position of a particular content source among all collaborative activity in the network. Similar to network strength, eigenvector centrality accounts for not only the number but also the intensity of collaborative activities associated with each content source. The measure is recursive in that, in our network of contributors and content, a content source has a high eigenvector centrality score if it is connected to other content sources that have high eigenvector centrality. Bonacich and Lloyd (2004) define eigenvector centrality as follows:

> Let $\mathbf{A}$ be a symmetric adjacency matrix, where $a_{ij} = a_{ji} = 1$ if i and j are connected in a network and $a_{ij} = a_{ji} = 0$ otherwise. The eigenvector measure of centrality x is the solution to the following matrix equation: $\mathbf{A}x = \lambda x$.

Eigenvector centrality can also be calculated on a valued matrix, so we define $a_{ij} = a_{ji} = n$, where n is the total number of contributions that contributor a makes to articles i and j. We then determine the vector λ of eigenvalues for each month from December 2007 until June 2009.

3.4.4. 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.

3.5. Control Variables

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

3.5.1. 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 1,094,010 characters. We use the natural log of article length in the statistical models.

3.5.2. Reading complexity. Articles may be more valuable if written in a more sophisticated style. That is, articles may be perceived as containing valuable information if they sound authoritative, whether they actually are or not. 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 (ARI; 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. For our analysis, relative values are more important than absolute values; the structure of Wikipedia articles results in relatively high ARI scores. For scaling purposes, we divide reading complexity by 1000.

Table 1 Descriptive Statistics

Variable	Minimum	Maximum	Mean	Std. Dev
Monthly Article Views (/1000)	0.001	3,675.587	10.512	28.534
Age (days)	1.000	2,979.000	1,289.149	614.370
Length (characters/1000)	0.000	1,094.011	10.018	12.355
Complexity (ARI/1000)	0.010	1,281.560	19.444	7.093
Section Depth	1.000	6.000	2.407	0.724
External References	0.000	303.000	9.987	21.435
Internal Links	0.000	3,508.000	58.741	71.963
Multimedia Content	0.000	35.000	0.047	0.499
Anonymity (percentage)	0.000	1.000	0.288	0.163
Distinct Contributors	0.000	1,743.000	50.767	120.549
Local Centrality	0.000	2,147.484	98.300	198.567
Global Centrality	0.000	1,999.522	46.700	180.601

3.5.3. 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, anonymous contributors made 29% of the contributions per article.

3.5.4. Information presentation. Because multimedia content may enhance the market value of information (Schlosser 2003), we control for the total number of multimedia files using a measure we call *multimedia images*. 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*.

3.5.5. References and links. Wikipedia policy states that all contributions should be supported by an authoritative external reference. On the Medicine Wikiproject, only peer-reviewed medical journals are considered authoritative. Contributors may attempt to manipulate the market value of the article by including more references, or the number of references could indicate the popularity of a topic in the medical literature. For example, although lung cancer is the leading

Table 2 Variable Correlations

	Variable	1	2	3	4	5	6	7	8	9	10	11
1.	Article Views (ln)	1.000										
2.	Age (ln, days)	0.566	1.000									
3.	Length (ln, chars)	0.435	0.210	1.000								
4.	Complexity (ARI)	0.078	0.004	0.308	1.000							
5.	Section Depth	0.362	0.176	0.617	0.148	1.000						
6.	External References	0.244	0.065	0.493	0.066	0.379	1.000					
7.	Internal Links	0.454	0.279	0.550	0.170	0.463	0.526	1.000				
8.	Multimedia Content	0.044	0.002	0.047	0.066	0.082	0.019	0.336	1.000			
9.	Anonymity (%)	0.575	0.486	0.268	0.062	0.196	0.012	0.226	0.006	1.000		
10.	Contributors	0.524	0.333	0.352	0.053	0.329	0.351	0.545	0.019	0.377	1.000	
11.	Local Centrality	−0.107	−0.062	0.178	0.036	0.116	0.230	0.068	0.006	−0.278	−0.107	1.000
12.	Global Centrality	−0.111	−0.387	0.044	0.025	0.023	0.076	0.014	0.022	−0.085	0.028	0.191

Correlations for the 132,447 Wikipedia Medicine monthly panel observations from December 2007 to June 2009. All correlations greater than 0.01 are significant.

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*.

3.5.6. Monthly fixed effects. Article viewing may also change over time. Therefore, we include indicator variables for each month.

Finally, we include the lagged value of each dependent variable to control for unchanging article-specific features (e.g., topic). Our analysis thus focuses on article views that result from collaborative activity.

4. Results and Discussion

We first analyze the effects of network structure on article views for the entire sample. Then, we use a holdout sample to test predictive validity as well as two alternative samples that demonstrate the robustness of our findings beyond the Medicine Wikiproject. Finally, we divide our original sample into five clusters of articles with similar viewing patterns and examine each cluster individually.

4.1. Aggregate Analysis

In Table 3, we provide the full results of a simultaneous equation 3SLS regression on the natural log of article views (scaled by 1000 for this presentation) using the sample of 132,447 monthly observations of articles from December 2007 until June 2009.

Table 3 Three Stage Least Squares Model of Article Views

Model	Variable	Model 1		Model 2	
Equation 1: Article Views (ln/1000)					
	Monthly Fixed Effects	indicators		indicators	
	Constant	744.439^{***}	(5.215)	755.126^{***}	(5.387)
	Article Views (ln, lagged)	0.971^{***}	(0.001)	0.971^{***}	(0.001)
	Age (ln, years)	18.287^{***}	(1.285)	17.964^{***}	(1.384)
	Length (ln, characters)	9.135^{***}	(1.164)	8.904^{***}	(1.166)
	Complexity (ARI)	-27.702^{*}	(11.181)	-26.060^{*}	(11.181)
	Section Depth	1.975^{*}	(0.847)	1.795^{*}	(0.848)
	External References	1.038	(0.690)	0.629	(0.693)
	Internal Links	3.261^{**}	(1.129)	2.990^{**}	(1.135)
	Multimedia Content	-0.014	(0.661)	-0.031	(0.661)
	Anonymity (percentage)	91.676^{***}	(5.857)	75.019^{***}	(6.171)
	Contributors	12.280^{***}	(1.424)	48.579^{***}	(4.627)
	Contributors^2	-6.780^{***}	(1.115)	-61.869^{***}	(6.901)
	Local Centrality	-1.993^{**}	(0.672)	-2.738^{***}	(0.729)
	Global Centrality	3.266^{***}	(0.685)	1.908^{*}	(0.765)
	Age × Contributors			-27.487^{***}	(3.483)
	Age × Contributors^2			42.469^{***}	(5.158)
	Age × Local Centrality			-0.856	(0.747)
	Age × Global Centrality			-1.694^{***}	(0.391)
	R^2	98.099		98.100	
	$\chi^2(\times 10^6)$	6.44^{***}		6.44^{***}	
Equation 2: Contributors					
	Monthly Fixed Effects	indicators		indicators	
	Constant	-0.250^{***}	(0.053)	-0.251^{***}	(0.053)
	Contributors (lagged)	1.021^{***}	(0.001)	1.021^{***}	(0.001)
	Article Views (ln)	0.103^{***}	(0.006)	0.103^{***}	(0.006)
	Age (ln, years)	-0.121^{***}	(0.011)	-0.121^{***}	(0.011)
	Length (ln, characters)	0.002	(0.012)	0.002	(0.012)
	Complexity (ARI)	-1.034^{***}	(0.113)	-1.034^{***}	(0.113)
	Section Depth	0.041^{***}	(0.009)	0.041^{***}	(0.009)
	External References	-0.052^{***}	(0.007)	-0.052^{***}	(0.007)
	Internal Links	0.139^{***}	(0.011)	0.139^{***}	(0.011)
	Multimedia Content	-0.025^{**}	(0.007)	-0.025^{**}	(0.007)
	Anonymity (percentage)	0.407^{***}	(0.056)	0.407^{***}	(0.056)
	Article Protected? (1=yes)	-5.787^{***}	(0.087)	-5.789^{***}	(0.087)
	R^2	99.958		99.958	
	$\chi^2(\times 10^8)$	3.02^{***}		3.02^{***}	

124,711 observations; standard errors in parentheses; significance $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Model 1 contains the focal network variables. Because of our large sample size, we use a low threshold of statistical significance ($p < 0.001$) to test our hypotheses. We find support for our first hypothesis; the number of unique contributors has a curvilinear relationship with the content's market value. Both the linear and squared coefficients are significant ($\beta = 12.28$, $p < 0.001$; $\beta = -6.78$, $p < 0.001$, respectively), which implies an inverted U-shaped relationship with article views. Additional contributors working on an article increase its market value up to a point, then detract from the ability of the article to attract viewers. We also considered models with a linear effect of contributors or a log of the number of contributors; the quadratic model provides a slightly

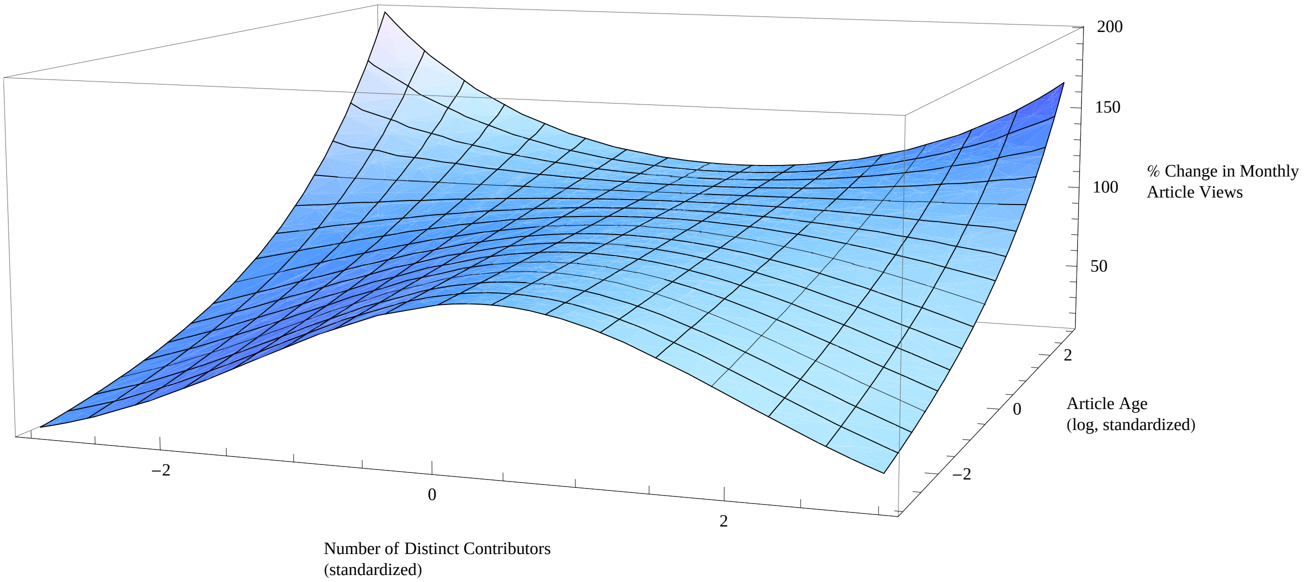


Figure 3 **Effect of Number of Contributors and Article Age on Article Views**

better fit (cf., AIC decreases by 18 and 7, respectively). However, we do not find support for our second hypothesis, because the coefficient for degree centrality per contributor does not meet our strict test for significance and is opposite from the hypothesized direction ($\beta = -1.99$, $p < 0.010$). Nevertheless, we do believe that this surprising result is worth further analysis, which we explore in a later section. In support of our third hypothesis, the market value of user-generated content relates positively to global network centrality; the coefficient for eigenvector centrality is positive and significant ($\beta = 3.27$, $p < 0.001$).

With Model 2, we find support for hypotheses 4a and 4c: Content age reduces the impact of network size and global network centrality on the market value of user-generated content. Specifically, age reduces the impact of the number of distinct contributors, with a negative coefficient for the linear interaction ($\beta = -27.49$, $p < 0.001$) and a positive coefficient for the squared interaction ($\beta = 42.47$, $p < 0.001$). For both the linear and squared terms, the net effect of age is to reduce, but not completely offset, the direct effect of the number of distinct contributors, as shown in Figure 3. The effect of age on local network centrality is not significant ($\beta = -0.86$, $p < 0.244$), whereas global network centrality has a decreasing effect ($\beta = -1.69$, $p < 0.001$) as articles age. Figure 3 illustrates the relationships among number of contributors, article age, and viewing activity.

Both models imply a recursive effect of article viewership on the number of contributors. From Equation 2, the coefficient for article views is significant and positive ($\beta = 0.10$, $p < 0.001$). More contributors lead to greater viewing, but more viewing yields more contributors. The protection variable, which affects contributions but not viewing, is significant ($\beta = -5.787$, $p < 0.001$) and behaves as expected. It is also interesting to note that age has an opposite effect in the contributor model—it relates positively to viewing but negatively to the overall number of contributors. This finding suggests that collaborative user-generated content matures and stabilizes over time, such that more people view older content but are less likely to contribute to it. It may be that more mature content attracts a more general audience that is less likely to have the knowledge or inclination to contribute, or perhaps viewers of the content find it relatively complete and have nothing to add.

4.2. Predictive Validity

To assess the predictive validity of the models, we used both an internal holdout sample (to assess temporal predictive validity) and external samples (to assess predictive validity in alternate contexts). First, we generated coefficient estimates using only the first ten months of data. We then used these estimators to predict monthly article viewing for the remaining nine months. The correlation of the predicted value with the observed values in the remaining nine months was 0.9891 indicating the models yield accurate estimates of future viewership within the same context. In the original Medicine sample, the correlation between the predicted values and actual values was 0.9905. Second, we used the estimators from the Medicine Wikiproject sample to generate predicted values of monthly article viewing for 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 644,336 revisions in the Fashion Wikiproject and 1,026,892 revisions in the Auto Wikiproject, then built the variables described in Section 3 and analyzed monthly viewing over the same period (December 2007–June 2009). For the Fashion sample, the correlation was 0.9848; for the Auto sample, it was 0.9885. That is, correlations for the alternative samples are slightly lower but remain quite high and confirm that the models can be generalized outside the context of our original sample.

4.3. Article Heterogeneity

Results from the aggregate analysis shown in Table 3 offers support for many of our hypothesized relationships. However, viewership of user-generated content is not homogenous; often a small percentage of content receives a high proportion of viewership (Barabasi 2003). In addition, viewership of some articles is relatively stable while others experience a high degree of variance. For example, a multi-part series on autism by CNN in 2005 increased viewership of autism-related Wikipedia articles. This could lead to very different network effects. To address this, we re-estimated our models for articles characterized by different viewership patterns over time.

4.3.1. Hierarchical clustering. Following previous research that infers market structure on the basis of differences in consumer behavior (Kohli and Jedidi 2007), we used hierarchical clustering to identify related groups of articles. We identified three key measures of differences in monthly viewing activity for each article over the study period—mean, variance and skewness—that correspond to the first three sample distribution moments (generally, $E\left[(X-\mu)^k\right]$, $\forall k \in \{1,2,3\}$). First, some articles attract a consistently higher volume of viewing activity. The mean, $E\left[(X)^1\right]$, reflects these differences in overall viewing activity. Second, to capture differences in the variability of viewing activity over time, we use the variance, $E\left[(X-\mu)^2\right]$. For clustering, we use the square root of the variance (standard deviation). Third, some articles had noticeable spikes in activity, so we use skewness, $E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$, to indicate the symmetry of viewing activity, relative to the mean. (For robustness, we also examined clusters based on the fourth moment, kurtosis, but found few qualitatively interesting distinctions in the resultant clusters.)

With these three measures, we grouped articles using agglomerative hierarchical clustering. To evaluate if two groups should be merged, we assessed similarity through the complete linkage, based on the farthest pair of articles in a group. (For robustness, we also considered alternative linkage methods, including single, average, and centroid linkages.) Our qualitative analysis, detailed subsequently, suggests that the clusters generated through the complete linkage best reflect differences in underlying viewing activity. To calculate the distance between the farthest articles, we used the unweighted Euclidean distance based on mean, standard deviation and skewness. (Again, we evaluated alternative measures, including squared Euclidean distance, absolute value distance, maximum value distance, and correlation coefficient similarity; we retain the Euclidean distance based on our qualitative analysis of the resultant clusters.)

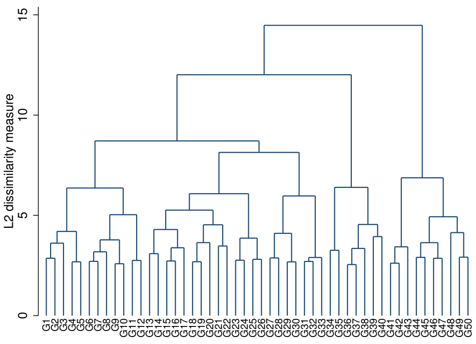


Figure 4 Dendogram for Cluster Analysis

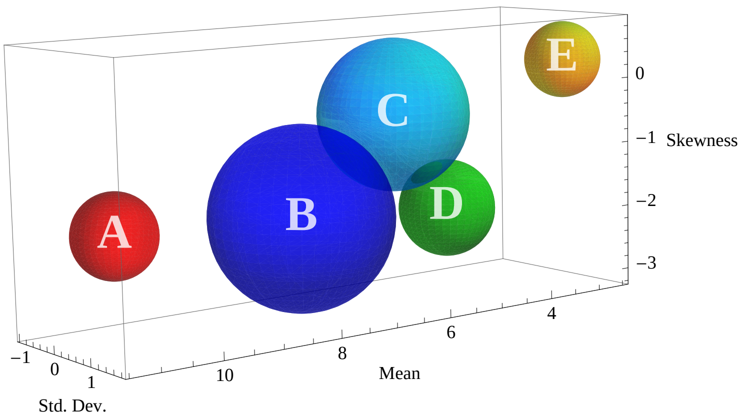


Figure 5 Relative Size and Viewing Attributes of Clusters

Table 4 Cluster Statistics

Variable	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Count	763	8,515	4,861	1,294	635
Traffic (ln) Mean	10.746	7.571	5.861	4.878	3.147
Traffic (ln) Std. Dev.	0.379	0.391	0.445	0.493	1.252
Traffic (ln) Skewness	−1.654	−1.783	−0.390	−1.917	0.297
Age (days)	2,067.246	1,310.568	935.815	770.576	968.706
Length (characters/1000)	23.280	9.065	7.288	5.938	8.000
Complexity (ARI/1000)	20.450	19.524	18.877	18.882	17.758
Section Depth	3.022	2.400	2.197	2.012	2.187
External References	25.491	8.453	7.973	5.682	9.533
Internal Links	152.473	52.502	37.894	31.026	40.377
Multimedia Content	0.088	0.051	0.016	0.031	0.031
Anonymity (percentage)	0.444	0.295	0.216	0.138	0.208
Distinct Contributors	272.387	30.928	13.823	5.930	20.773
Local Centrality	426.474	69.830	41.600	24.717	57.693
Global Centrality	44.902	31.634	85.158	98.750	96.434

Observations are monthly.

Figure 4 depicts the dendogram generated by hierarchical clustering. Using the dendogram and qualitatively analyzing the results, we identified five clusters. (Compared with two, three, four, six, ten, fifteen, and twenty cluster alternatives, five clusters balanced parsimony with more precise modeling.) In Table 4, we list the mean values of the clustering variables for the resultant five clusters. Because we used three variables for clustering, a three-dimensional plot can describe the cluster relationships. Figure 5 illustrates the relative position of each cluster according to the mean values of the clustering variables. The volume of each sphere is proportional to the number of articles in the cluster.

4.3.2. Article viewing by clusters. To examine differences among clusters in article viewing patterns over time, we plotted monthly viewing activity (both raw and log-transformed) of 20 articles from each cluster for which viewing mean, variance, and skewness were close to the cluster average. In Figure 6, we plot the results for a typical article in each cluster. Clusters A and B are clearly distinct in their average viewing. Cluster A is small but experiences the most viewership by far. Articles in this cluster also on average are considerably older and longer and have many more contributors than other articles in the sample. They represent the major topics in medicine and deal with the most common diseases, procedures, and medications (e.g., heart failure, muscular dystrophy, triglycerides). Cluster B is large (more than half the articles in our sample) and exhibits notably less viewership than Cluster A, though still considerably more than the others. These articles focus on less common but still mainstream topics, such as monocular vision, hand surgery, and egg allergies.

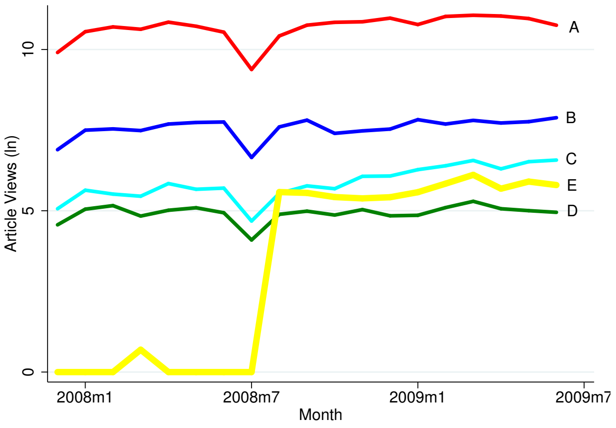


Figure 6 **Viewing of Representative Articles**

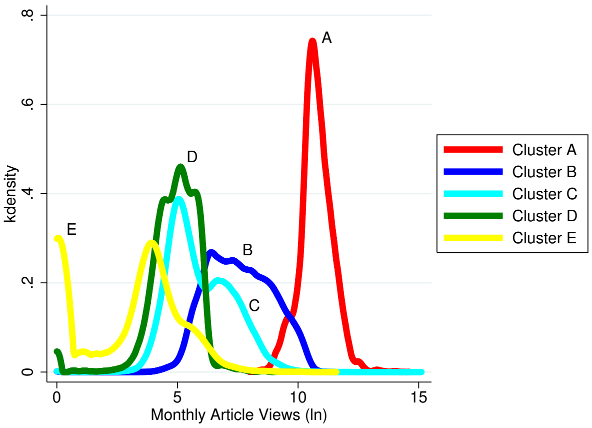


Figure 7 **Density of Article Views by Cluster**

The remaining clusters are more similar to one another in average viewership; all involve moderate viewing levels. The articles in these clusters pertain to niche and focused topics, such as periorbital dermatitis and staphylococcal dermatitis in Cluster C. The most specific topics, in Clusters D and E, include an article in D focused on the “Meningeal branch of the mandibular nerve” (it received around 100 views per month). The clusters also differ from each other in their growth patterns over time. For example, articles in Clusters C and E experienced increased viewership during the sample period. Cluster C represents approximately one-third of the articles in our sample and articles in this cluster move from relatively modest viewership to greater viewing rates, which may reflect the increasing popularity of both viewing and contributing to information on Wikipedia during this time. Articles in Cluster E instead move from virtually no viewership to moderate viewing rates and include articles that started as mere placeholders with no content and then attracted some content and viewing. Articles in Cluster D show virtually no growth in viewership. Figure 7 provides an alternative perspective using viewing distribution densities to show differences in the variance and skewness of viewership across the clusters.

4.4. Heterogeneity and Endogeneity

Because we identified a recursive relationship between viewership and contributions, and because contribution behavior may differ for articles with different viewership patterns, we reassess our analysis in light of these potential differences. In Table 5, we provide the results of the same 3SLS

Table 5 **Three Stage Least Squares Model of Article Views by Cluster**

Variable	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Equation 1: Article Views (ln/1000)					
Constant	1393.047***	678.483***	834.934***	1057.112***	794.992***
Article Views (ln, lagged)	0.919***	0.985***	0.950***	0.878***	0.800***
Monthly Fixed Effects	indicators	indicators	indicators	indicators	indicators
Age (ln)	18.895**	4.334**	23.735***	14.557*	26.801
Length (ln)	6.278*	2.279	12.248***	14.322*	82.180**
Complexity (ARI)	−4.836	−32.592*	−20.339	50.034	168.810
Section Depth	1.194	0.882	−0.647	21.121**	−66.398*
External References	1.096	1.605*	1.504	26.644***	−1.908
Internal Links	−0.453	1.183	−1.006	−33.229***	82.232
Multimedia Content	−2.113	−0.125	2.559	37.516**	83.102
Anonymity (%)	−77.986*	49.333***	70.382***	−6.444	−39.375
Contributors	7.886**	10.611***	61.185***	541.240***	209.255**
Contributors^2	−1.721	−7.659**	−55.704**	−1351.000***	−97.117*
Local Centrality	−52.251***	−2.017**	−0.132	0.573	−0.558
Global Centrality	0.526	2.071*	5.618**	0.373	4.102
R^2	91.53	98.20	92.81	90.70	80.71
$\chi^2(\times 10^5)$	1.40	45.00	3.20	0.41	0.06
Equation 2: Contributors					
Constant	−15.484***	−0.232***	−0.033	0.822***	0.255
lagged Contributors	1.019***	1.020***	1.020***	1.027***	1.017***
Article Views (ln)	1.507***	0.072**	0.037***	−0.148***	0.103***
Monthly Fixed Effects	indicators	indicators	indicators	indicators	indicators
Age (ln)	−0.839***	−0.101***	−0.027*	−0.012	−0.088**
Length (ln)	−0.036	0.021*	0.014	0.017	0.026
Complexity (ARI)	−2.109***	−0.282**	−0.132	0.091	−0.057
Section Depth	0.190***	−0.007	0.029*	0.020	0.020
External References	−0.018	−0.022***	0.047***	0.050***	0.054
Internal Links	0.185***	0.071***	−0.038	−0.008	−0.019
Multimedia Content	0.036	−0.013**	0.007	−0.029	0.025
Anonymity (%)	9.712***	0.234***	0.147*	−0.165**	0.373
Article Protected?	−7.030***	−1.489***	0.388	–	−1.443
R^2	99.94	99.94	99.85	99.94	99.98
$\chi^2(\times 10^7)$	2.2	15.0	1.6	0.7	0.6
N	12,522	81,800	24,728	4,184	1,477

significance $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

regression as in Table 3, broken out by the five clusters. The independent analysis of each cluster suggests that the effects of the collaborative network are not universal.

The results generally hold for the two largest clusters, B and C. Together, these two clusters represent approximately 13,000 articles, or 83% of the sample. However, for articles in Clusters D and E (i.e., articles with the lowest viewership), although the number of contributors continues to have a curvilinear effect on viewership, the wider network interactions of these contributors—as measured through local and global centrality—have no significant relationship with content value. These relatively peripheral articles (by topic or contributor) do not benefit from the network in the same way as typical articles.

Conversely, the most viewed articles, represented by Cluster A, exhibit no significant negative effect of network size, the strongest and a significant negative effect of local network centrality, and no effect of global network centrality. One interpretation of these results is that high-profile articles generate a considerable volume of content, which gets spun off into independent sub-articles. This trend appears in qualitative studies of content generation on Wikipedia as well (Kane 2010). For these articles more contributors are not detrimental to content value, because they eventually leave to work on subprojects. Furthermore, only for articles in Cluster A does anonymity have a significantly negative effect on viewing. These high profile articles may be more subject to vandalism by anonymous contributors.

The cluster analysis also provides insights into the surprising result, that an increase in local network centrality reduces viewership. Local network centrality is not significant for niche articles (Clusters C, D, E), weakly negative for typical articles (Cluster B), and significantly and strongly negative for the most heavily viewed articles (Cluster A). This suggests several possibilities. First, if heavily viewed articles create content that gets spun off into sub-articles, higher local network centrality means that more contributors leave the focal article to work on related subtopics, decreasing the value of the focal article. Thus, local network centrality may indicate the degree to which articles fragment into multiple sub-articles.

Second, collaborating on multiple articles (i.e., greater local network centrality) may offer valuable information to a particular article but also take it away. Contributors may look to the most heavily viewed articles as exemplars. Valuable content, processes, or reputation information generated in one content source can be transferred immediately to other sources on which the collaborators work. Although this effect would not decrease the objective quality of the information in the focal content, it may decrease its market value compared with other articles with which it competes for viewer attention—particularly if the information moves into articles on related topics.

Third, very highly viewed articles may attract contributors who seek to forward a particular agenda. Global network centrality creates value through greater connections to collaborative content, but local network centrality might reveal the opposite. Working on many different articles might demonstrate that contributors are not committed to high value content but rather seek simply to ensure that a particular agenda is widely disseminated. Kane and Fichman (2009) found that some contributions to Wikipedia articles (e.g., about the 2007 Virginia Tech shootings) sought to advance a gun control agenda, which induced wild swings in article content as opposing groups wrestled for control. Contributions by issue-oriented contributors may complicate the process of developing valuable collaborative user-generated content.

5. Discussion

For this article, we studied the entire compendium of 16,068 Wikipedia articles in the Medicine Wikiproject to determine the effect of collaborative network structure on the perceived market value of user-generated content, as measured by viewership. Consistent with our hypotheses, we found that network size (number of contributors to a content source) relates curvilinearly to viewing and that global network centrality (intensity of direct and indirect links to more collaborative content) relates positively to it. We also find that both effects are stronger for newer sources of content than for established ones. Local network centrality (the number of other sources of user-generated content on which a contributor works) does not meet our strict test for significance but is negatively related to article viewing, and this effect is not significantly moderated by content age. We also find that contribution to and viewership of user-generated content is recursive; greater viewership leads to more contribution, and more contributors lead to greater viewership. Analyses using internal and external holdout samples demonstrate high predictive validity for future viewership of articles in the Medicine Wikiproject as well as accurately predicting viewership of articles on very different topics (i.e., fashion and autos). Finally we clustered our results and found that, though our hypotheses hold for the vast majority of typical articles, network effects differ for the most and least prominent articles. 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 exciting 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 still flow from one content source to another as contributors work on multiple sources. Further research should consider how relationships among content sources affect content creation; and researchers cannot simply assume that stronger connections always have positive effects on content value. Although research tends to emphasize the benefits of networks, there are potentially detrimental effects (cf., Labianca and Brass 2006).

Second, we demonstrate the dynamic nature of the relationship between content generation and viewership. Content created by more contributors attracts more viewers, and more viewers increase the number of contributors. This dynamic may be particularly salient in a setting such as Wikipedia, where anyone can contribute, but it also applies more broadly to other social media platforms in which people contribute comments or feedback. People may be drawn to content precisely because they have an interest in or possess some base knowledge about it. As they read the content, they may contribute if they recognize that they possess knowledge that can improve it. Our results also suggest, however, that this dynamic stabilizes over time. Further research should explicate this recursive relationship between viewership and contribution.

Third, our cluster analysis reveals that network effects on different sources of user-generated content are not equal. For example, collaborative networks associated with older sources of user-generated content may affect viewership at the beginning of the content lifecycle and content generated about high-profile topics may affect viewership differently than content about relatively specific topics. Ongoing research should examine factors that lead to different network effects across different sources of collaborative user-generated content.

5.2. 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 cluster. 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, we raise the important question of how to maintain the value of collaborative user-generated content. The most valuable content in our sample is most harmed by contributors' activity on other collaborative projects. It is not clear whether this effect is due to the partitioning of content, a reduction in relative content value as information gets copied to other sources, or the advancement of agendas that reduce content value. It may be possible to protect content in its mature stage, but it is difficult to keep information from being copied. This finding reveals a potential downside of relying on users to generate content: They create valuable information, but this information is difficult to control.

6. Limitations and Conclusion

Several limitations to this study suggest the need for further research. Although SNA provides important insights into how the relationships among content creators and content sources affect viewership, it refers to the potential for content to flow among nodes, rather than measuring the actual flow. Additional research might examine the extent to which specific content and process knowledge gets transferred through collaboration networks. Although our data show that more and less prominent articles exhibit different network effects than do typical ones, limitations in our data set prohibit us from discerning whether these differences reflect the topic of the collaboration, time, or both. We presume that more important articles were created first; yet Wikipedia viewing data are available only beginning in December 2007, so we cannot disentangle these competing explanations. Furthermore, network measures likely reflect multiple characteristics of the network involved in creating user-generated content, such as creators' experience in creating content as well as their content knowledge. Identifying the relative importance of these different aspects is an interesting topic for future research.

We could not capture aspects of the user's search for content, such as whether the Wikipedia article was the first result returned in a Google search. Although search ranking may affect our results, it reflects content value as represented through viewership. Google's PageRank algorithm prioritizes pages with more incoming links, because presumably people link to content they find most valuable. Thus, viewership—whether directly through a related site or indirectly through a search engine—is part of the market value that drives viewing.

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. Understanding these effects is particularly important considering the increasingly collaborative nature of user-generated content and the growing interest by firms in generating revenue from this content.

References

Alba, J. W., J. W. Hutchinson. 1987. Dimensions of consumer expertise. *J. of Consumer Res.* **13**(March) 411–454.

Amaral, L. A. N., A. Scala, M. Barthélémy, H. E. Stanley. 2000. Classes of small-world networks. *PNAS* **97**(21) 11149–11152.

Barabasi, A.-L. 2003. *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. Penguin, New York.

Berger, J., K. Milkman. 2009. Virality: What gets shared and why? M. C. Campbell, J. J. Inman, R. Pieters, eds., *Advances in Consumer Res.*. Association for Consumer Research. Vol. 37, forthcoming.

Bonacich, P., P. Lloyd. 2004. Calculating status with negative relations. *Social Networks* **26**(4) 331–338.

Borgatti, S. P., R. Cross. 2003. A relational view of information seeking and learning in social networks. *Management Sci.* **49**(4) 432–445.

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.

Brin, S., L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* **30**(1-7) 107–117.

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. *Information 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 Review E* **74**(3) 036116.

Carmi, E., G. Oestreicher-Singer, A. Sundararajan. 2009. Spreading the oprah effect: The diffusion of demand shocks in a recommendation network. Available at http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1127&context=ici2009. Accessed on August 15, 2010.

Carrington, P. J., J. Scott, S. Wasserman. 2005. *Models and Methods in Social Network Analysis*. Structural analysis in the social sciences, Cambridge University Press, New York.

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.

Cross, R., L. Prusak. 2002. The people who make organizations go- or stop. *Harvard Bus. Rev.* **80**(6) 104–112.

Cummings, J. N. 2004. Work groups, structural diversity, and knowledge sharing in a global organization. *Management Sci.* **50**(3) 352–364.

Davis, A., B. B. Gardner, M. R. Gardner. 1941. *Deep South: A social anthropological study of caste and class*. University of Chicago Press, Chicago.

Denning, P., J. Horning, D. Parnas, L. Weinstein. 2005. Wikipedia risks. *Communications of the ACM* **48**(12) 152.

Devgan, L., N. Powe, B. Blakey, M. Makary. 2007. Wiki-surgery? Internal validity of wikipedia as a medical and surgical reference. *J. of the American College of Surgeons* **205**(3) S76–S77.

Duan, W., B. Gu, A. B. Whinston. 2008a. Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems* **45**(4) 1007–1016.

Duan, W., B. Gu, A. B. Whinston. 2008b. 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.

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.

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.

Gersick, C. J. G. 1988. Time and transition in work teams: Toward a new model of group development. *Acad. of Management J.* **31**(1) 9–41.

Godes, D., D. Mayzlin. 2004. Using online conversations to measure word of mouth communication. *Marketing Sci.* **23**(4) 545560.

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.

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.

Iacobucci, D., N. Hopkins. 1992. Modeling dyadic interactions and networks in marketing. *J. of Marketing Res.* **29**(1) 5–17.

Kane, G. C. 2010. A multimethod study of information quality in wiki collaboration. *ACM Transactions on Management Information Systems* **1**(1) forthcoming.

Kane, G. C., R. G. Fichman. 2009. The shoemaker's children: Using wikis for information systems teaching. *MIS Quarterly* **33**(1) 1–17.

Kane, G. C., R. G. Fichman, J. Gallaugher, J. Glaser. 2009a. Community relations 2.0. *Harvard Bus. Rev.* **87**(11) 45–50.

Kane, G. C., A. Majchrzak, J. Johnson, G.L. Chen. 2009b. A longitudinal study of perspective development in a fluid online collective. H. Chen, S. A. Slaughter, eds., *Proceedings of the 30th International Conference on Information Systems*. Phoenix, AZ.

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.

Kohli, R., K. Jedidi. 2007. Representation and inference of lexicographic preference models and their variants. *Marketing Sci.* **26**(3) 380–399.

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.

Labianca, G., D. J. Brass. 2006. Exploring the social ledger: Negative relationships and negative asymmetry in social networks in organizations. *Acad. of Management Rev.* **31**(3) 596–614.

Laurent, M. R., T. J. Vickers. 2009. Seeking health information online: Does wikipedia matter? *J. of the American Medical Informatics Association* **16**(4) 471–479.

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.

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.

Manchanda, P., Y. Xie, N. Youn. 2008. The role of targeted communication and contagion in product adoption. *Marketing Sci.* **27**(6) 961–976.

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.* forthcoming.

Newman, M., A.-L. Barabasi, D. J. Watts. 2006. *The Structure and Dynamics of Networks*. Princeton University Press, Princeton, NJ.

Oestreicher-Singer, G., J. Goldenberg, S. Reichman. 2009. The quest for content: The integration of product and social networks in ugc environments. *Paper Presented at the Emergence and Impact of User-Generated Content*. Philadelphia, PA.

Oestreicher-Singer, G., A. Sundararajan. 2010. The visible hand of social networks in electronic markets. Available at http://ssrn.com/abstract=1268516. Accessed on August 15, 2010.

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. Accessed on January 8, 2010.

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.

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.

Scott, J. 2000. *Social Network Analysis: A Handbook*. 2nd ed. Sage, Newbury Park.

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.

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.

Tuckman, B. W. 1965. Developmental sequence in small groups. *Psychological Bulletin* **63**(6) 384–399.

Wasserman, S., K. Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, New York.

Wattal, S., P. Racherla, M. Mandviwalla. 2010. Network externalities and technology use: A quantitative analysis of intraorganizational blogs. *J. of Management Inform. Systems* **27**(1) 145–174.

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.

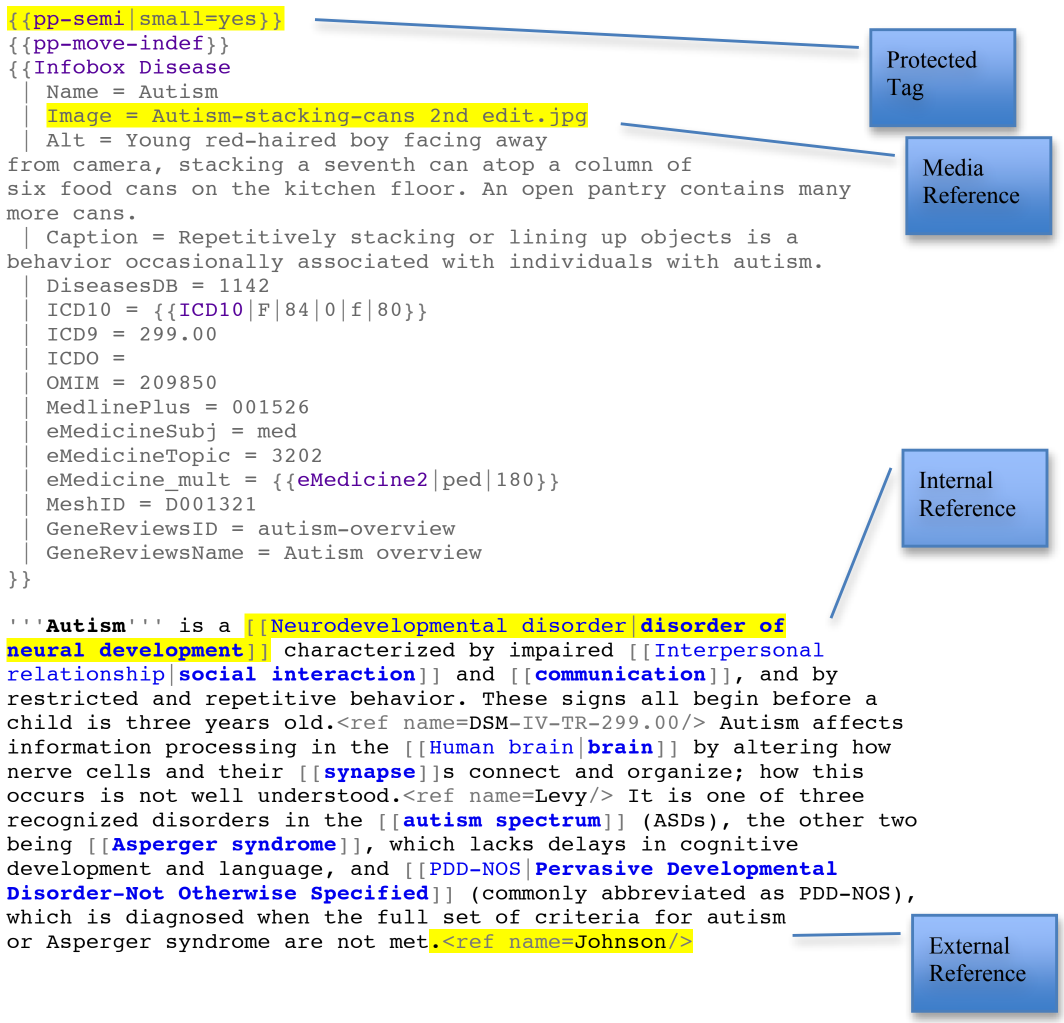
Wellman, B. 2001. Computer networks as social networks. *Science* **293**(5537) 2031–2034.

Wikipedia. 2010. Wikipedia: Article size. Available at http://en.wikipedia.org/wiki/Article_length. Accessed January 6, 2010.

Zlatić, V., M. Božičević, H. Štefančić, M. Domazet. 2006. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E* **74**(1) 016115.

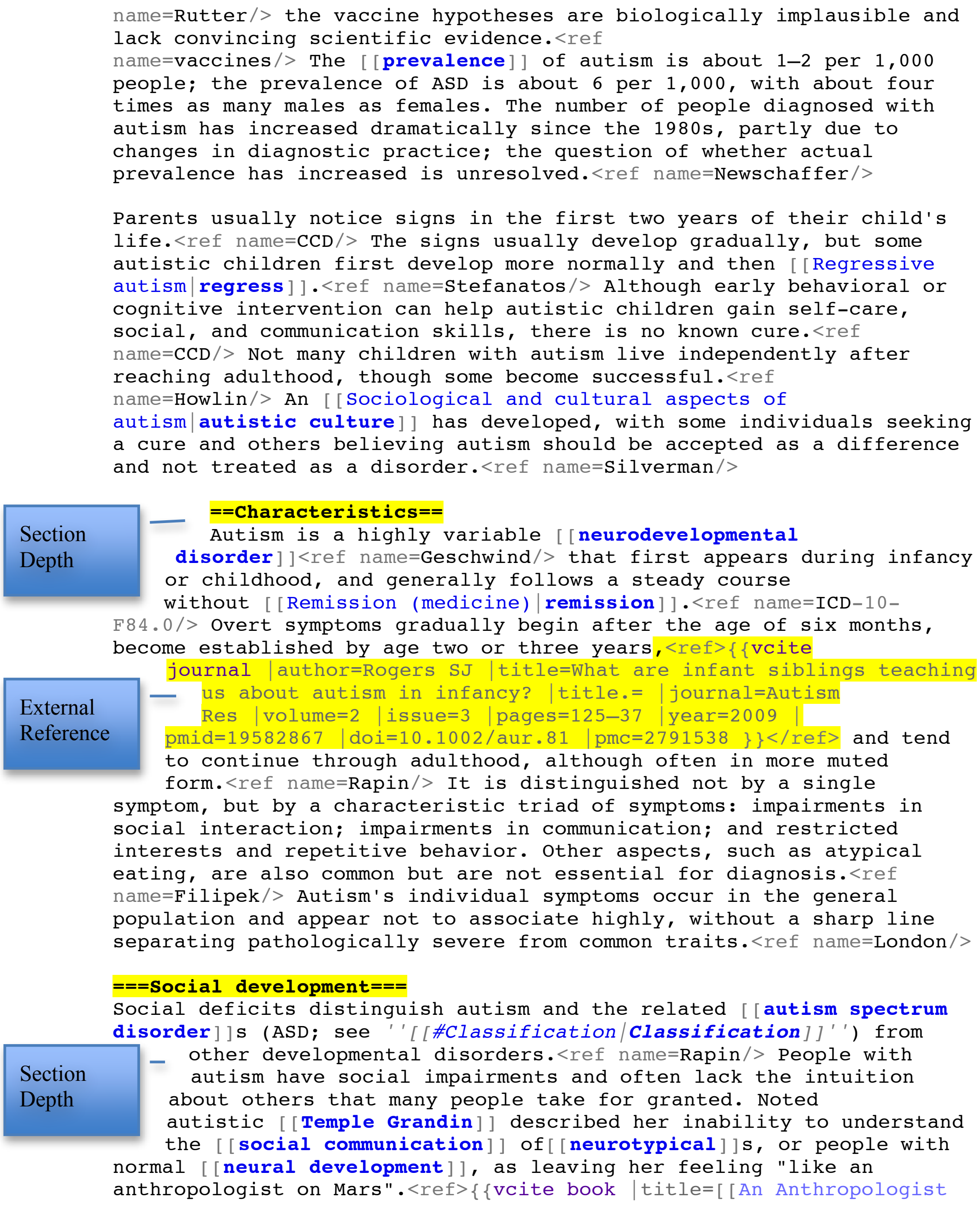
Appendix. Sample Article

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



```
{{pp-semi|small=yes}}
{{pp-move-indef}}
{{Infobox Disease
 | Name = Autism
 | Image = Autism-stacking-cans 2nd edit.jpg
 | Alt = Young red-haired boy facing away
from camera, stacking a seventh can atop a column of
six food cans on the kitchen floor. An open pantry contains many
more cans.
 | Caption = Repetitively stacking or lining up objects is a
behavior occasionally associated with individuals with autism.
 | DiseasesDB = 1142
 | ICD10 = {{ICD10|F|84|0|f|80}}
 | ICD9 = 299.00
 | ICDO =
 | OMIM = 209850
 | MedlinePlus = 001526
 | eMedicineSubj = med
 | eMedicineTopic = 3202
 | eMedicine_mult = {{eMedicine2|ped|180}}
 | MeshID = D001321
 | GeneReviewsID = autism-overview
 | GeneReviewsName = Autism overview
}}

'''Autism''' is a [[Neurodevelopmental disorder|disorder of
neural development]] characterized by impaired [[Interpersonal
relationship|social interaction]] and [[communication]], and by
restricted and repetitive behavior. These signs all begin before a
child is three years old.<ref name=DSM-IV-TR-299.00/> Autism affects
information processing in the [[Human brain|brain]] by altering how
nerve cells and their [[synapse]]s connect and organize; how this
occurs is not well understood.<ref name=Levy/> It is one of three
recognized disorders in the [[autism spectrum]] (ASDs), the other two
being [[Asperger syndrome]], which lacks delays in cognitive
development and language, and [[PDD-NOS|Pervasive Developmental
Disorder-Not Otherwise Specified]] (commonly abbreviated as PDD-NOS),
which is diagnosed when the full set of criteria for autism
or Asperger syndrome are not met.<ref name=Johnson/>
```

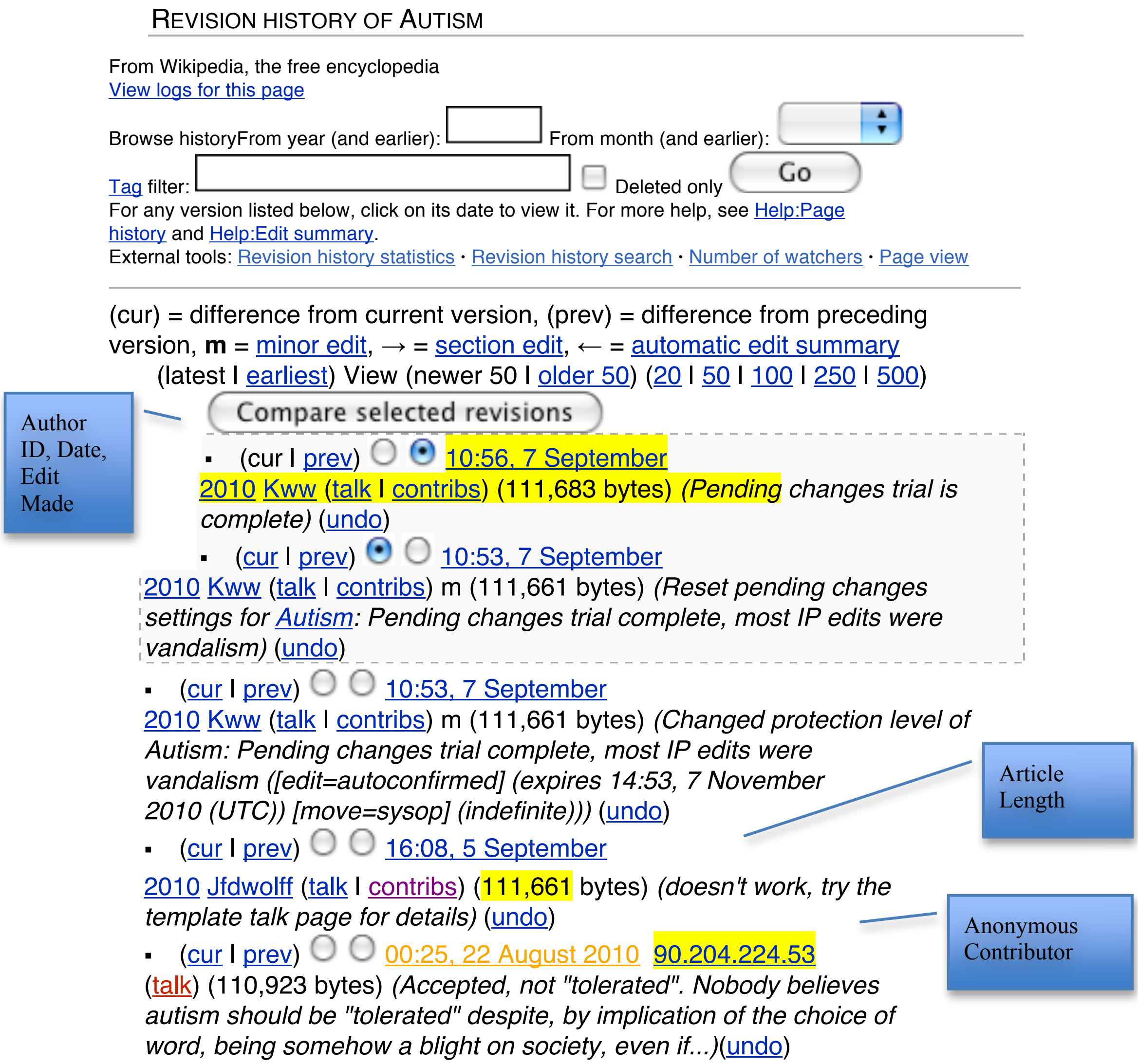


```
name=Rutter/> the vaccine hypotheses are biologically implausible and
lack convincing scientific evidence.<ref
name=vaccines/> The [[prevalence]] of autism is about 1–2 per 1,000
people; the prevalence of ASD is about 6 per 1,000, with about four
times as many males as females. The number of people 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/>

          ==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/>

===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
```



REVISION HISTORY OF AUTISM

From Wikipedia, the free encyclopedia
View logs for this page

Browse historyFrom year (and earlier): [ ] From month (and earlier): [ ]

Tag filter: [ ] □ Deleted only [Go]

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)

[Compare selected revisions]

- (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)