JOURNAL OF COMMUNICATION

Journal of Communication ISSN 0021-9916

ORIGINAL ARTICLE

The Flow of Digital News in a Network of Sources, Authorities, and Hubs

Matthew S. Weber[1] & Peter Monge[2]

1 School of Communication and Information, Rutgers University, New Brunswick, NJ 08901, USA
2 Annenberg School for Communication and Journalism, University of Southern California, Los Angeles, CA 90089, USA

This article presents an analysis of the flow of information in a network of online news sites. Social network theory and research on hyperlinked networks of Web pages are used to develop a model of information flow among Web sites. Kleinberg's authority-hub model is extended by introducing sources of information in the network. Significant support was found for a Source–Authority–Hub model, which shows the source, directionality, routing, and destination of news information flow through a network of authorities and hubs. This model demonstrates the ability of key Web sites to control the flow of news and information. Applications of the model to over-time data have the potential to predict future changes in the online news industry.

doi:10.1111/j.1460-2466.2011.01596.x

Online news first appeared in the mid-1990s when news articles were reprinted from traditional papers onto the Internet, primarily through Usenet groups (Cohen, 2002). The Internet has continued to grow as a news medium, and as of 2008 the Web replaced print as a primary medium for day-to-day news among young adults (Kohut, 2008). It is apparent, however, that as the news industry transforms from traditional to online forms, traditional newspaper business models do not apply. The acceleration in the volume of information transmitted online (Cao & Li, 2006), the decrease in economic costs (Varian, Farrell, & Shapiro, 2004), and the proliferation of user control (Lin & Salwen, 2006) have led to a decrease in the economic value of traditional printed news content and an increase in the production of electronic information (Boczkowski & Ferris, 2005). The vast amount of online content paired with the rapid growth in users has, on average, reduced organizations' abilities to charge a premium for online content. Similarly, the vast amount of Web sites and the dispersion of users has reduced the amount that organizations can charge advertisers for online ads (Boczkowski & Ferris, 2005; Cao & Li, 2006). Thus, the emergence of an Internet news market is leading news organizations to seek new business models (Patterson, 2007).

Corresponding author: Matthew S. Weber; e-mail: matthew.weber@rutgers.edu

To understand the significance of ongoing changes in industry structure and concentration, a new model for analysis of information flow in online networks is needed. A new model will help researchers and practitioners alike to understand changing patterns of information flow. The continued global growth of Internet users, increased access speeds, and wireless connectivity suggests that overall online news sources will continue to gain readers at the expense of print alternatives (Boczkowski, 2004; Savage & Waldman, 2005). Instead of a traditional push-model, users are free to navigate between sites to seek the information they desire and select their own versions of the daily news. Ultimately, the shift toward the Internet as a primary source of global news will empower individual users and create a fluid information landscape. Despite all these changes, few studies examine the movement of digital news from producers through to consumer-targeted Web sites. In order to provide a predictive model of the online news environment, an existing model of online information flow is extended based on a review of relevant literature. The new model is tested with a data set collected from online news sites and the results are presented as a foundation for understanding the continued transformation of the industry, as well as the effects of trends such as continued media convergence.

Information flows and networks

The study of communication networks and information flow is well established in the field of organizational communication. Communication networks have been utilized to study social relations, information exchange, and organizational networks since the 1970s (Monge & Contractor, 2003; Wigand, 1988). For example, Rice's (1982) longitudinal study of information flows within and between workgroups found that information flows clustered into distinct patterns, and demonstrates that complex communication networks can be deconstructed into clearly delineated clusters representing the patterns of information movement. From a similar perspective, Rogers (1985) and Rogers & Kincaid (1981) showed that ideas and information diffuse at varying rates in distinct patterns through networks of individuals and organizations, dependent on the information communicated, the maturity of the network, and the frequency of communication.

More recently, this line of research has been expanded to the study of hyperlink networks. For instance, Schumate and Lipp's (2008) research showed that non-governmental organizations (NGOs) utilize hyperlink networks to drive collective action toward a common social cause. Other work has demonstrated that organizations cluster together via their hyperlink affiliations in diverse contexts such as HIV/AIDS organizations (Schumate & Dewitt, 2008), networks of environmental activists (Ackland, O'Neil, Bimber, Gibson, & Ward, 2006), and communication patterns utilized by terrorists (Arquilla & Ronfeldt, 2001). Adamic and Adar (2003) showed that blogs cluster according to political affiliation, highlighting the importance of network location; a user's starting point in seeking information in a network will guide the perspectives that a given user views.

Towards a new model of online information flow

Relatively little previous work examines the overall flow of content through digital information networks such as online news, that is, from producers through to dissemination. There are, however, exceptions. Barnett and Park (2005) laid a foundation for examining structure and movement in their study of hyperlinks as a means of connecting organizations in an online information network. Their research, framed in world systems theory's notion of a core, periphery and semiperiphery, showed that the global media system functions as an interconnected network with major nations such as the United States occupying core positions, and nations with a less-developed physical infrastructure, such as Lithuania and Morocco, occupying periphery positions. Kleinberg, Raghavan, and Gibson (1998) developed a computer algorithm based on the flow of information through online networks that facilitates searching for information on the Internet. They decomposed Internet Web pages into two categories: hubs and authorities. This algorithm, called Hyperlink-Induced Topic Search (HITS), reduces the Web to a directed graph where Web sites are nodes and hyperlinks between pages are directed links showing one-way flow.

Using a combination of real-world data and computer simulations, the HITS algorithm demonstrated that searches for information are guided by authorities, which are trusted sources of information and content. Authorities are identified as such because multiple link directories, or hubs, point users to authorities and identify those sites as trustworthy (Kleinberg & Lawrence, 2001). There is generally a balance between hubs and authorities; the relationship between the two is mutually reinforcing. Furthermore, Kleinberg (1999) found that collections of pages coalesce into information "communities" or "subgraphs." This model has proven to be robust and accurate in the identification of topical clusters of Web sites (Asano, Tezuka, & Nishizeki, 2007). Although this model has generally been applied to online search, this research is not altogether disparate from work in communication which has utilized the mapping of hyperlink networks to illustrate online representations of organizational networks (Halavais, 2008; Park, 2003).

The validity of the HITS algorithm is well established for modeling relationships between Web sites and is suitable for mapping information flows when it is possible to divide the entire corpus of Web sites into either authorities or hubs. By building on analogous work in communication, it provides a useful starting point for describing an information network typology. In Kleinberg's work, the HITS model is applied to a cluster of Web pages to identify likely information sources based on linking patterns. In other words, ideal matches are identified based on the pattern, or "flow," of content through links to key hubs. Similarly, in the online news environment there is a contingent of wire services feeding information through to a core of online newspapers that aggregate news.

There are two critical modifications made here to apply the HITS model to the study of online news; the model is recast to accommodate the complex movement of online news and to capture the feeding of information into the network. First, the model proposed here is based on the transmission of information rather than

the direction of hyperlinks, as is the case in the HITS model. In the case of online news, sources feed content into the network, rather than simply form hyperlinks. This is an important departure; although in certain cases hyperlinks can represent the sharing of information, they are often indiscriminate linkages between Web sites and require minimal investment to create (Halavais, 2008). The movement to sharing of content, however, represents a conscious intention to connect two Web sites, and a clear exchange of information between two Web sites; this is, therefore, a stronger measure of relationships between Web sites.

A second modification results from the growth of Web 2.0 technology and the resulting complexity as news Web sites have become more prodigious (Constantinides & Fountain, 2008; Xia, Huang, Duan, & Whinston, 2007). In information-rich online networks, specific nodes have the function of feeding information into the network. News wire services such as the Associated Press (AP), Reuters, and Agence France-Presse (AFP) supply a steady stream of content into the network, functioning as sources of information. Thus, sources are proposed as an addition to authorities and hubs in order to model the communication of news through online networks. These source sites feed content into a vast network of intermediaries—major metropolitan news sites, regional newspapers, and topic specialists—that validate and edit this content. Authorities specialize either on topics or in regions, and form a dense central set of intermediaries. One example is the *San Francisco Chronicle*, a paper that primarily covers news about San Francisco and northern California. Intermediary authorities publish news fed in from source Web sites, but also add content and information depending on the specialty of a given authority. For example, *Business Week* publishes articles based on information from wire services, but also publishes feature articles about current trends in business based on internal reporting. In this way, the Business Week Web site performs two functions: first, the Web site filters information from sources, and second, the Web site adds original content specific to its role as an authority. As in the original HITS model, a core set of hubs also exist, parsing together information into a directory format. These hub Web sites collect stories primarily from authorities, but stories are also posted occasionally in a raw format directly from sources. For example, Google News collects news stories from thousands of online news authorities, and provides them in a directory format to users, but also posts content directly from wire services.

Kleinberg's (1999) model of authorities is therefore extended here to include sources of information, and in turn, it is used to model the full news media network of sources, authorities, and hubs. The general Source–Authority–Hub (SAH) model is illustrated in Figure 1. The SAH model accounts for the increasing complexity of online information networks. Sources feed content into the network through authorities, filtering information from sources but also adding knowledge and insight to the network. The original function of authorities as topical aggregators is retained, augmented by the addition of unique knowledge. Finally, in line with the original model, hubs collect links and direct users to the most relevant or appropriate information for a given topic. Information may flow in a reverse direction, with

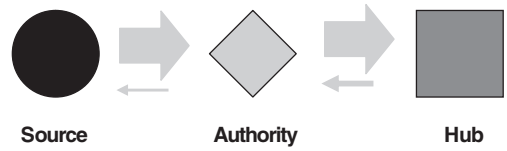

Figure 1 The flow of online news as a Source–Authority–Hub model.

small amounts of information transmitted from hubs to authorities, and even less transmitted from authorities back to sources. While ultimately information is provided out to end-users, this model focuses on information within the network. Users connect into the network when they access information from a given Web sites; additionally, users often navigate the network via hyperlinks. This network is generally expected to have a small number of centralized sources connected through a high number of intermediate authorities, through to a small number of centralized hubs. Sources incur significant costs associated with the production of information and the provision of information to the network. Authorities are generally regional or niche, and function on a smaller scale in terms of the amount of content flowing through a given authority. Authorities represent the bulk of the network, and correspond to the vast number of regional centers across the country, as well as the long tail of niche content markets. Finally, hubs incur search costs seeking out information and reliable sites from which they can aggregate information. The resulting SAH model is tested as a basis for examining online information flow.[1]

To test the SAH model, the presence of sources is represented by both the source parameter and the k-out-star network parameters. The source parameter models a node that only has outlinks. A positive source parameter indicates the presence of Web sites that only connect outwards to other Web sites. The k-out-star parameter indicates that a Web site primarily links outwards to other Web sites, but may have reciprocal ties. These two parameters capture the possible range of source configurations. It is hypothesized that both the source and k-out-star parameters will be significant and positive, indicating that they are a critical part of the network.

H1: Digital news networks will have a greater number of sources and k-out-star configurations than random information networks.

Authorities, on the other hand, are sites that filter information provided by sources and feed content onwards to hubs. The alternating transitive 2-path k-triangle (TK-2-P) parameter is used to model this communicative pattern. When this parameter is positive, it indicates a given Web site will transmit information to a partner Web site through a set of intermediary nodes that serve as connectors. This parameter predicts the presence of a high number of intermediary Web sites between sources and hubs and in the SAH model it is likely to be significant and positive, predicting the existence of authorities.

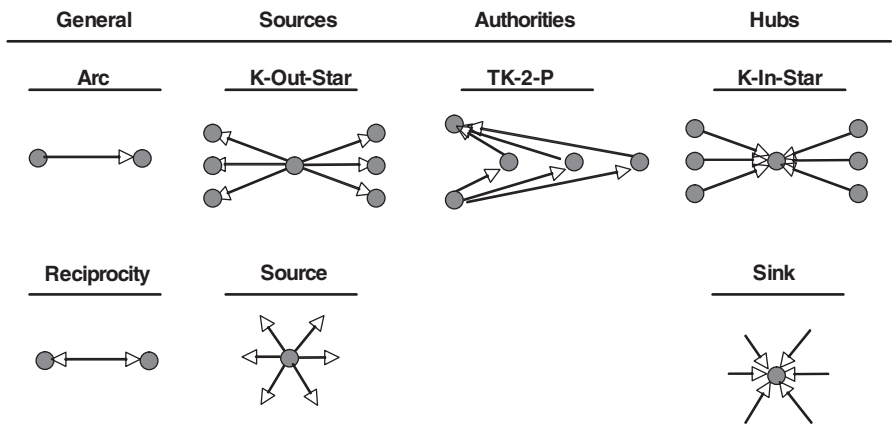

Figure 2 PNET parameters used in estimation.

H2: Intermediary authorities are more likely to exist in a digital news network than in random information networks.

Hubs are the third component of the SAH model. With regards to information flow, hubs are indicated by the combined presence of the sink parameter and the k-in-star parameter. The sink parameter indicates the presence of a Web site that only receives links from other Web sites. The k-in-star parameter indicates that a Web site primarily receives links from other Web sites, but may have reciprocal ties. These two parameters are used to capture the possible range of configurations of hubs. Thus, it is hypothesized that both the k-in-star parameter and the sink parameter will be positive and significant in the SAH model:

H3: Digital news networks will have a greater number of sinks and k-in-star configurations than random information networks.

In addition to the above parameters, arc and reciprocity parameters were included as baseline measures of the network. For a network of Internet sites, the arc parameter is expected to be negative, corresponding with low network density. Figure 2 summarizes the hypothesized parameters as they correspond to the predicted SAH model. (For further technical specifications of parameters see Robins, Pattison, Kalish, & Lusher, 2007).

Method

Data collection

Examination of hyperlinks is considered the default tool for extracting structural relationships among sites in online networks, and a substantial body of literature examines the general geography and interconnectedness of Web sites (Barnett & Sung, 2006; Lin, Halavais, & Zhang, 2007; Park, 2003; Park, Barnett, & Nam, 2002; Park

& Jankowski, 2008; Park & Thelwall, 2008). Hyperlink analysis can be problematic, however, as linking has a very low cost barrier; all that is required is a string of code to connect two Web sites. As a result, hyperlink analysis may contains extensive erroneous links, irrelevant information, and inconsequential relationships (Thelwall, 2004). An alternative approach to hyperlink analysis is to study online organizational relationships through the context or content of relationships between organizations on the Internet (Cohen & Mitra, 1999; Cohn & Hofmann, 2001; Li & Zaiane, 2004). This research examined the flow of information between online organizations by examining the sharing of content. For news organizations, the direct sharing of content between two organizations is acknowledged in two ways. News organizations often acknowledge contributions from other sources in a byline of the article. Alternatively, an organization may create a hyperlink to contributing partners. Both linkages establish the movement of information between online news sites. Therefore, this study examined the flow of content by examining bylines acknowledging the contribution of information from another source, and hyperlinks acknowledging contributions from a partner. In this manner, the network distinguishes flow from the representational network that would be created by using only hyperlinks, in line with distinctions made in previous research examining network growth (Leskovec, Kleinberg, & Faloutsos, 2007).

A subset of online news Web sites were collected using a snowball sample (Erickson, 1979), and the largest online news aggregator in the United States, Google News, was selected as the starting point. Google News draws content from a sample of more than 4,500 sites; at any given moment, however, a subset of only 40 to 50 distinct links is displayed on the Google News home page. The selection of Google News as a starting point insured that a diverse set of links was collected. Web crawls were conducted on two dates, November 5 and November 8, 2007. Links were recorded by hand, moving stepwise outward from Google News to collect a complete network. The crawl was conducted by hand to verify the nature of the linking relationships and insure that the links extracted represented content sharing relationships. A link was recorded when a given organization acknowledged another organization directly in a byline, or a hyperlink acknowledging content sharing between two organizations.

News originates from diverse sources, and as a result a variety of different types of Web sites were collected. The sample included news aggregators, blogs, Web sites of television stations, newspapers, magazines, and news services. Blogs were collected only if they were associated or directly connected to an online news service or newspaper. This delineation eliminated Web sites that focused on commentary, as well as general discussion sites. Online newspaper sites account for the bulk of information flow in digital news networks (Gao & Vaughn, 2006). Web sites in this initial collection were classified as sources, authorities, or hubs. These classifications were based on their stated mission, generally obtained from the "About" Web page. Sources are those sites that serve as wire services or content providers, feeding information into the network. Authorities are reputable intermediaries: generally online newspapers, blogs, news television, Web sites, and news magazines that filter a

significant amount of content from wire services. Finally, Hubs sites were identified as those sites that collected information and served as information directories. In addition, a reciprocal link was recorded when a Web site collected news from another source, and provided a link back to the story on the original Web site. A nonreciprocal link occurred when information was collected from a partner or other news site, but no link was provided back to the original source.

Data analysis

Exponential random graph (ERG) models were utilized to test the SAH model. This type of modeling, previously referred to as p* models, can be used to analyze a wide array of social networks (Anderson, Wasserman, & Crouch, 1999; Snijder, Pattison, Robins, & Handcock, 2006). ERG models use maximum likelihood estimates based on Markov Chain Monte Carlo (MCMC) procedures for optimal parameter estimates (Robins et al., 2007). ERG methods estimate the probability that certain linking patterns exist in the communication network compared to what would occur on the basis of random chance alone (Robins et al., 2007). These parameters are used here to test the fit and appropriateness of the SAH model. Alternatively, the proposed network model could be analyzed using betweenness, outdegree, and indegree centrality measures. Betweenness measures the degree to which a node is between other nodes; outdegree measures the degree to which a node is connected outward to others, and indegree measures the degree to which other sites connect inward to a node (Valente, Coronges, Lakon, & Costenbader, 2008). Although these measures would account for each Web site's general position in the network, and the site's role feeding information out or collecting information in, they would not provide a general model for the flow of online news.

ERG models illustrate communication patterns as parameters corresponding to different linking patterns. MCMC ERG modeling allows for the computation of the hypothesized triangle parameters, as well as transitivity parameters. This facilitates the modeling of more complicated linking patterns (Robins et al., 2007), such as those seen in hyperlink data. The ability of ERG modeling to capture network complexity makes it optimal for examining online news networks. In addition, the NetDraw 2.0 package (Borgatti, 2002) was used for network visualizations. Data analysis was done with the PNET 1.0 software package (Wang, Robins, & Pattison, 2005). The software tests the fit of hypothesized parameters and estimates a general model based on simulations of the data. The model is said to fit the data when all parameters have $t < 0.10$ (Snijder et al., 2006). This indicates that the standard error of each estimated parameter is within a tolerable range of the actual value of the parameter, based on the original data and as compared to randomly generated networks of the same size. Specific parameters are significant when the values are within 1.96 standard errors of the parameters estimated by the model ($p < .05$; Robins, 2007, p. 32).

Goodness-of-fit model was assessed by generating simulated networks based on the estimated parameters, selecting a random sample, and measuring the overall fit compared to the original data. In the goodness-of-fit test, all parameters are modeled

including those not measured originally, in order to assess the global fit of the model. The goodness-of-fit test reports three calculations: the mean of each given configuration observed in the sample networks, the standard error of the sample mean, and a t value comparing the sample mean to the expected outcome based on the input parameters. For parameters that were estimated in the original model, the model is acceptable if all values fit at $t < 0.10$. For parameters that were not estimated originally, the model is acceptable if values fit at $t < 2.00$. In complex simulations, models can be accepted even when a handful of parameters are outside acceptable bounds.

In addition to the goodness-of-fit test, the overall fit of indegree and outdegree centrality were assessed (Wang, Robins, & Pattison, 2006). Given the SAH model's focus on degree centrality, this goodness-of-fit assessment follows in line with previous work by Goodreau (2007) and Goodreau, Hunter, & Morris (2005). The values for indegree and outdegree centralities for the simulated models were graphed as a boxplot against the actual data. Standard deviation and skew are also reported. In addition, a blockmodel analysis was conducted to assess the likelihood of information movement between sources, authorities, and hubs. A basic ERG model was generated with MultiNet v5.17; this package was used as it allows for the generation of blockmodels using ERG estimation (Seary & Richards, 2000).

Results

The snowball sample yielded a network of 239 Web sites with 23 news services, 7 news aggregators, 178 online news sites, 9 online magazines, 15 online sites for television stations, and 7 online blogs. This analysis included sites from six continents, representing an international information network. Overall, the network was sparsely populated with a density of 0.0103, consistent with sparsely populated networks found in other studies of online information (Barabási, 2003).

The SAH model

Preliminary baseline analysis

A baseline analysis was conducted before the primary analysis to eliminate parameters that do not occur in the network. This process eliminated several triangle parameters that do not appear either theoretically or empirically in the data set. The seven theoretical parameters that represent the SAH model were then fitted to the data.

Primary SAH model analysis

First, the arc parameter was significantly negative (-7.88, $t = 0.73$, $SE = -0.03$) corresponding with the expected sparseness of the network. Second, the reciprocity parameter was positive (3.38) indicated a notable and significant presence of reciprocal relationships between Web sites. The first hypothesis stated that a significant number of sites would function as sources in a digital news network as indicated by the positive, significant values for the k-out-star and source parameters.

Table 1 PNet Results for Optimized Model

Parameter	Value	*SE*	*t*
Arc	−7.88*	0.73	−0.03
Reciprocity	3.38*	0.31	0.06
Source	2.08*	0.40	0.02
k-out-star	3.46*	0.35	−0.04
Sink	−0.37*	0.16	−0.08
k-in-star	3.16*	0.78	−0.03
TK-2-P	0.17*	0.02	−0.01

*$p < .10$.

The k-out-star parameter had a positive, significant value of 3.46 ($t = -0.04$, $SE = 0.35$) and the source parameter had a value of 2.08 ($t = 0.02$, $SE = 0.40$), thus H1 was supported. The second hypothesis stated that there would be a significant number of intermediary authorities in digital news networks as indicated by the positive, significant value for the TK-2-P parameter. The TK-2-P parameter had a significant, positive value of 0.17 ($t = -0.01$, $SE = 0.02$), supporting H2. The last hypothesis stated that there would be a significant number of hubs in digital news networks as indicated by significant, positive values for the k-in-star and sink parameters. The k-in-star parameter had a positive, significant value of 3.16 ($t = -0.03$, $SE = 0.78$), and the sink parameter had a slightly negative, significant value of −0.37 ($t = -0.08$, $SE = 0.08$), partially supporting H3. The full results of the fitted model are summarized in Table 1, which provides each parameter value, standard error, and *t* value. Table 1 shows that all seven network parameters are statistically significant at $p < .10$ and are essential structural components of the network.

Primary goodness-of-fit analysis

To test the overall fit of the estimated parameters, a goodness-of-fit simulation was conducted. The estimated parameters, as well as a subset of general network characteristics, were simulated over an extended period (millions of estimations) and the appropriateness of the estimations were reported with global values, standard errors, and *t* values. The results are shown in Table 2. All values are under the specified limit of 2.0, indicating that there is good fit between the SAH model and the data.

Secondary goodness-of-fit analysis

The two box plots, shown in Figures 3 and 4, model the overall fit of the combined Markov and higher-order parameters based on the distribution of the outdegree and indegree distributions with the actual data (the solid black line). The diagrams show fairly close alignment between the distributions, although the actual data tend to skew lower in value. Although the results are not perfect, these analyses provide additional support for the overall fit of the model to the data.

Table 2 PNet Goodness-of-Fit Results

Parameter[a]	Sample Mean	*SE*	*t*
Arc	**588.00**	**n/a**	**n/a**
Reciprocity	**120.98**	**17.38**	**0.06**
Isolates	60.90	87.55	−0.68
In-2-star	19696.99	11337.14	−1.46
Out-2-star	19972.03	1155.94	−1.26
In-3-star	1426158.71	982126.14	−1.42
Out-3-star	1454828.73	1001641.86	−1.70
Mixed-2-star	39646.04	22803.11	−1.70
T1	48.00	42.07	−1.41
T2	422.55	319.62	−1.29
T3	587.38	427.40	−1.30
T4	293.61	211.61	−1.26
T5	294.40	211.66	−1.32
T6	18660.27	11710.53	−1.59
T7	38231.70	23125.98	−1.65
T8	38492.28	23286.86	−1.65
T9	795.67	580.71	−0.85
T10	264.80	197.33	−1.24
Sink	**33.13**	**4.96**	**−0.03**
Source	**80.73**	**6.51**	**0.04**
k-in-star	**743.41**	**259.15**	**−0.08**
k-out-star	**749.58**	**257.28**	**−0.09**
AKT-T	472.07	296.07	−0.93
TK-2-P	**831.09**	**884.36**	**−0.10**

[a]Bold indicates parameters that were included in the estimation.

Blockmodel analysis

The results of the MultiNet blockmodel analysis are shown in Table 3. First, link creation (choice), reciprocity, and popularity were used to generate a basic overall model. Subsequently, the results show a significant likelihood of link creation from sources to authorities (2.90, $SE = 0.19$, $p < .01$), sources to hubs (1.43, $SE = 0.15$, $p < .01$), and authorities to hubs (1.08, $SE = 0.16$, $p < .01$). Information flow was modeled in both directions, and the results show the relatively low likelihood of link creation from hubs to authorities (-1.53, $SE = 0.37$, $p < .01$), hubs to sources (-2.25, $SE = 1.18$, $p < .10$), and authorities to sources (-0.09, $SE = 0.27$, $p > .25$).

Centrality analysis

An alternative approach to this type of network analysis is an examination of centrality measures. Table 4 presents the means and standard deviations of betweenness centrality, indegree centrality, and outdegree centrality. These data provide a comparison to traditional network measures and were calculated using an a priori classification

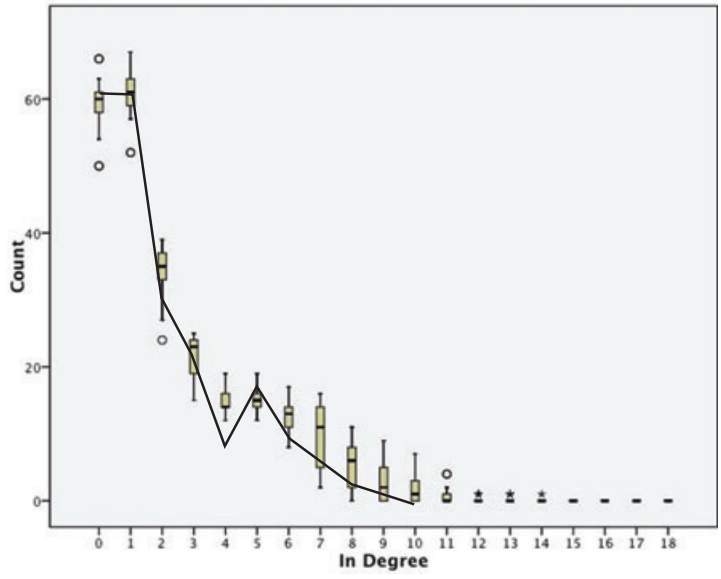

Figure 3 Boxplot of outdegree. values from model simulation.

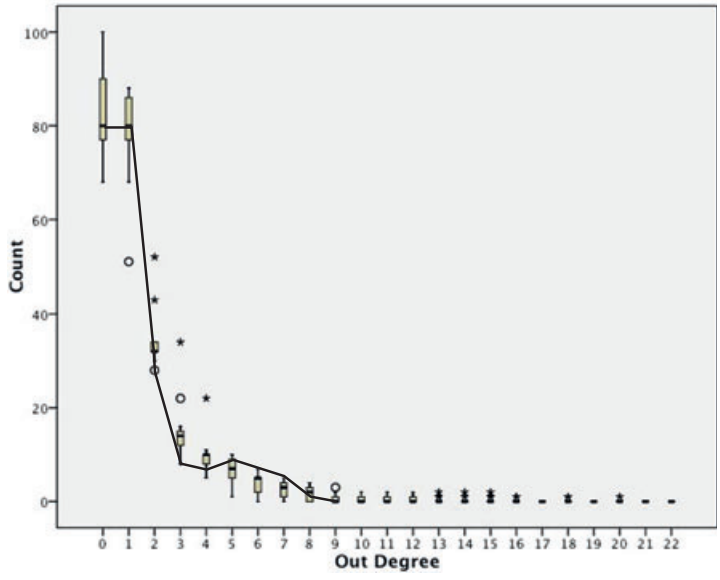

Figure 4 Boxplot of indegree values from model simulation.

of sources, authorities, and hubs. Although sources clearly have a higher outdegree centrality than authorities and hubs, and hubs have a higher indegree than sources or authorities, the distinction is less clear for betweenness centrality. The measures

Table 3 Blockmodel of Log-Likelihood of Link Creation Between Sources, Authorities, and Hubs

		B	*SE*	Wald^{PL}	*p*
Choice	Global	−5.65	0.09	4159.11	<.01
Reciprocity	Global	2.62	0.12	1053.02	<.01
Popularity	Global	0.06	0.01	185.55	<.01
Choice Block Model	Source → Authority	2.90	0.19	230.58	<.01
	Source → Hub	1.43	0.15	89.11	<.01
	Authority → Hub	1.08	0.16	43.07	<.01
	Authority → Source	−0.09	0.37	17.46	<.01
	Hub → Authority	−1.53	0.27	0.12	>.25
	Hub → Source	−2.25	1.18	3.67	<.10

Note: Model −2 log pseudo-likelihood = 7045.54.

Table 4 Betweenness, Indegree, and Outdegree Centrality for Sources, Authorities, and Hubs

	Indegree	Outdegree	Betweenness
Sources ($n = 27$)	3.62 (3.74)	6.84 (4.61)	2.06 (4.30)
Authorities ($n = 182$)	1.92 (1.93)	1.87 (2.06)	0.56 (1.40)
Hubs ($n = 30$)	4.95 (3.66)	1.95 (3.09)	0.90 (2.33)

for betweenness centrality are not clearly differentiated by role; thus, it is hard to categorize sites based on this approach. Overall, these findings support the results of the ERG model.

Discussion

The results of this analysis provide insight into the structure and flow of online news, providing a model that can be used to study patterns of information movement in the online news industry. Overall, the model parameters correspond closely with the predicted SAH structure. The significant and negative arc parameter is consistent with the traditional small world model of a sparse hyperlink network. The positive reciprocity parameter suggests that information flows in two directions rather than one, which captures the fact that the majority of Web sites occupy authoritative positions that are likely to send and receive information from partners.

The magnitude of the source (2.08) and k-out-star (3.47) parameters indicates a strong source component. The source parameter indicates that sources exist in the network at a little over twice the rate expected by chance. The results also indicate that although "pure" sources are likely to exist, sources that also have reciprocal links are more likely. This is indicated by the greater magnitude of the k-out-star parameter as compared to the source parameter value. Regarding hubs, the sink parameter was lower than expected (−0.37) indicating that they are less likely to occur compared

to chance. k-in-stars, however, have a positive and significant parameter value (3.16) indicating the strong presence of Web sites that collect information but also connect to other Web sites. These results indicate that "pure" hubs may exist, but Web sites exist that collect information and link to other Web sites are more common. Thus, hubs are a significant part of the network, but they are also likely to have reciprocal links. Comparing sources to hubs, it is clear that the sources occur in greater numbers. Given that this study included a number of strong content sources such as the Associated Press and Reuters, this is not a surprising result. Within this global network, this finding suggests a large number of news producers whereas there are far fewer central aggregators of links and content. This corresponds with Kleinberg's model, which posits that while information flow can be modeled using authorities, they are less prominent than hubs, and in this case, than sources as well. The significant authority parameter (TK-2-P) indicates the presence of Web sites occupying positions between information sources and hubs, a pattern that corresponds with the existence of authorities, which serve as intermediaries in the network. The parameter had a slightly negative value of −0.17, indicating that although authorities are significant they are not as prevalent as other Web sites.

The results of this study show that the complexity of online information flows can be represented by concise models of information transmission that account for the structure of real-world networks. The analysis decomposes the network into its core communication patterns. As is often the case, however, the model is not as clear as predicted. Hubs were significant but they rarely existed as pure hubs (with only inlinks). This is clearly reinforced by the positive value (3.38) of the reciprocity parameter. Bidirectionality may be an indicator that the two parties involved acknowledge and actively participate in the reciprocal relationship (Rogers & Marres, 2000; Schumate & Dewitt, 2008). Although content generally flows through the network in a single direction from sources to authorities to hubs, this shows that organizations reciprocate information sharing in multiple directions. An example would be national news after first having been covered by a niche source. In this case, the story would originate from an authority before being distributed through the entire network. Overall, information generally moves from sources to authorities to hubs (S → A → H) but only occasionally moves in the reverse direction (H → A → S). This general flow of information is shown in Figure 5, which illustrates a subcomponent of the complete SAH network. Two key sources, the Associated Press and Reuters, are shown feeding information through key authorities including the *Los Angeles Times* and the *New York Times*. Information is ultimately aggregated into hubs such as Google News, Huffington Post, and Yahoo! News.

The results of the blockmodel, presented in Table 4, further support the general movement of information from sources to authorities to hubs. The results demonstrate the positive, significant probability of information being shared across organization types. The only significant parameter for reverse information flow in this model is from hubs to authorities, indicating that certain authorities may receive information back from hub organizations. This blockmodel, however, does not

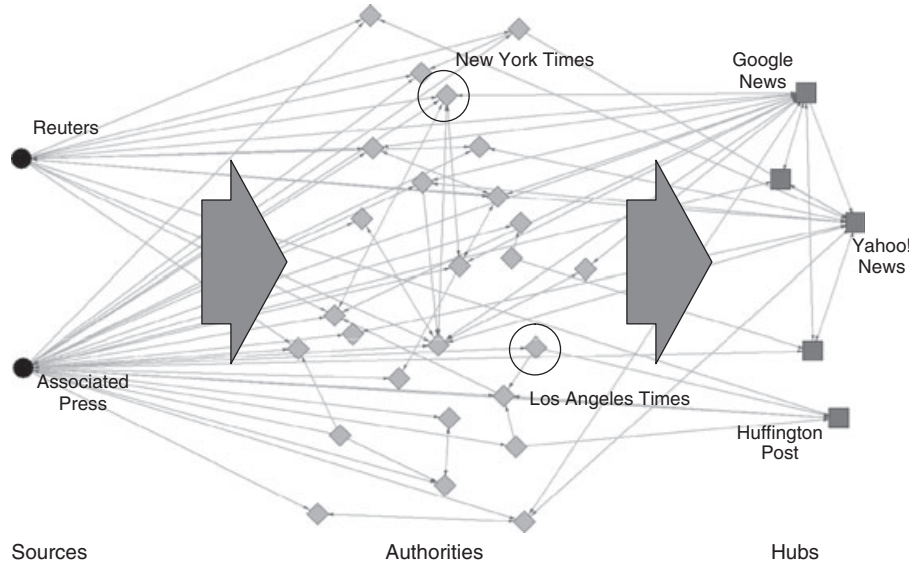

Figure 5 Visualization of overall network.

provide as accurate a fit as the PNet estimation process because it does not allow for the estimation of higher order parameters.

This overall pattern of flow from sources through to hubs highlights the critical role of sources in effectively controlling the general conversation. The work of Adamic and Glance (2005) showed that a user's starting point in the blogosphere influences the political slant of content, but this work shows that sources actually push the majority of content into the network. Thus, it is important for consumers to understand the potential biases associated with news sources such as the Associated Press or Reuters, as the news from these sources is ultimately fed through the whole network. Strength of both sources and hubs additionally points to online consolidation echoing the ongoing pattern of consolidation across traditional news media, which has dominated the industry as of late (Arsenault & Castells, 2008; Cohen, 2002). The origin of information and its distribution through increasingly complex communication networks is critical to our understanding of today's networked society, as information shapes the representation of society. (Castells 2004, p. 425) argues that the consolidation of power in society is shifting from traditional organizations to ''images of representation around which societies organize their institutions, and people build their lives.'' Evidence of this influence is seen in the dominance of hubs that aggregate news, multimedia, and information.

Limitations

The overall model is robust and significant; however, there are three limitations worth noting. First, in order to apply Kleinberg's HITS model to information flow, certain modifications have been made that may affect generalizability. In Kleinberg's

original algorithm, sites that had little consequence in the filtering of information were removed from the analysis and not included in the model. Here, however, all Web sites that were collected have been included. Because data were crawled by hand and restricted to a selected category of Web sites, the data in the sample set were all of consequence. Thus, all sites were additionally categorized as sources, authorities, or hubs. Second, network data may have been subject to a centrality bias. The snowball method used for data collection dictated that a single, central site be chosen as a starting point thus inherently bequeathing that site as central to the network (Erickson, 1979). However, by selecting a naturally central hub, this bias was minimized. Also, a complete snowball sample was collected aggregating all links extended from the initial Web site, thus avoiding the pitfalls of a partial dataset. Furthermore, by choosing ERG models as the main analysis method, this study focused on the global level of online news networks rather than on the characteristics of particular sites or links. Third, limits in existing computer programs for advanced ERG models required that a relatively small data set be used for this research. Existing ERG modeling techniques are limited by significant processing time and degeneration with data sets that exceed 300 nodes.

Future research

Future studies would benefit from an analytical approach to historical data in order to examine the evolutionary mechanisms that led to the emergence of networks represented by the SAH model. While this model is accurate for studying the current shape of digital information networks, an evolutionary perspective would isolate the mechanisms that lead to its creation (Monge, Heiss, & Margolin, 2008). Changes in concentration and flow can be detected by analyzing the network structure at various periods of development over time. Additional variables such as employment levels, revenue, and information about readers of both online and offline news sources could reveal substantial insight into the underlying mechanisms of industry shift. Furthermore, building on studies such as Adamic's (2005), this model can be applied to examine the impact that source reporting has on agenda-setting across online news sites. For example, by constructing variables that account for the types of stories produced by sources, the overall effect throughout the system could more accurately be analyzed.

Scholars examining shifts in the news industry should continue to focus on longitudinal data in order to analyze market trends. Recent changes in the industry continue to point toward an ongoing evolutionary process. For instance, Google now feeds Associated Press stories directly to consumers via the Google News Web site (Claburn, 2007). This decision may push consumers to obtain news directly from wire services. Differences in global policy will also affect analysis as issues of state control and state influence over media systems are likely to influence the general movement of information, as has been seen in Russia (Rantanen, 2001; Richter, 2008) and China (Harwit & Clark, 2001; Zhang, 2007). More work is needed to understand the overall impact of both strategic changes and government policy.

Competition and concentration also creates new opportunities for expansion. Within the past 5 years evidence of future opportunities has emerged as Web sites focused on user-generated content and community news programs have gained in popularity, particularly among younger generations (Kohut, 2008). In this way, Web technology has enabled the creation of "commons-based peer production," which Benkler (2006, p. 60) describes as a new "modality of production" in response to previous concentration of information production. Thus, while concentration appears to be an ongoing trend, questions remain as to the long-term shape of the industry. The model presented in this research provides a framework for examining these issues, as well as the role that key Web sites play in agenda-setting. By extending the work in this study to questions such as these, the SAH model can be utilized to understand the nature of the changing digital information landscape.

Acknowledgments

An earlier version of this article was presented at the 58th Annual Conference of the International Communication Association, Montreal, Canada. The authors acknowledge the very helpful comments of the editor and two anonymous reviewers from *Journal of Communication*. They also thank Garry Robins, Drew Margolin, and Cindy Shen for their valuable feedback on the manuscript. This research was supported in part by grants from the National Science Foundation (IIS-0838548) and the Annenberg School for Communication and Journalism.

Note

1 Kleinberg's algorithm focused on patterns of search and information indexing, whereas the SAH model proposed in this research tests the explicit presence of sources, authorities, and hubs. The movement of information longitudinally is not explicitly tested here, but left for future study.

References

Ackland, R., O'Neil, M., Bimber, B., Gibson, D., & Ward, S. (2006). *New methods for studying online environmental-activist networks.* Paper presented at the 26th Annual International Sunbelt Social Network Conference, 24–30 April, Vancouver, Canada.

Adamic, L., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, **25**, 211–230.

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery* (pp. 36–43). New York, NY: ACM Press.

Anderson, C., Wasserman, S., & Crouch, B. (1999). A p* primer: Logit models for social networks. *Social Networks*, **21**, 37–66.

Arquilla, J., & Ronfeldt, D. (2001). *Network and netwars.* Santa Monica, CA: RAND.

Arsenault, A., & Castells, M. (2008). The structure and dynamics of global multi-media business networks. *International Journal of Communication*, **2**, 707–748.

Asano, Y., Tezuka, Y., & Nishizeki, T. (2007). *Improvements of HITS algorithms for spam links.* Paper presented at the APWeb/WAIM 2007, Heidelberg, Germany.

Barabási, A.-L. (2003). *Linked: The new science of networks.* New York, NY: Plume.

Barnett, G., & Park, H. W. (2005). The structure of international internet hyperlinks and bilateral bandwidth. *Annales des telecommunications*, **60**, 1110–1127.

Barnett, G., & Sung, E. (2006). Culture and the structure of the international hyperlink network. *Journal of Computer-Mediated Communication*, **11**, 217–238.

Benkler, Y. (2006). *The wealth of networks.* New Haven, CT: Yale University Press.

Boczkowski, P. J. (2004). *Digitizing the news: Innovation in online newspapers.* Cambridge, MA: MIT Press.

Boczkowski, P. J., & Ferris, J. A. (2005). Multiple media, convergent processes and divergent products: Organizational innovation in digital media production at a European firm. *The Annals of the American Academy of Political and Social Science*, **597**(1), 16.

Borgatti, S. P. (2002). *NetDraw: Graph visualization software.* Cambridge, MA: Analytic Technologies.

Cao, Z., & Li, Z. (2006). Effect of growing Internet newspapers on circulation of U.S. print newspapers. In X. Li (Ed.), *Internet newspapers.* Mahwah, NJ: Erlbaum.

Castells, M. (2004). *The information age: The power of identity.* Malden, MA: Blackwell.

Cohen, E. (2002). Online journalism as market-driven journalism. *Journal of Broadcasting & Electronic Media*, **46**, 532–548.

Cohen, E., & Mitra, A. (1999). Analyzing the web: Directions and challenges. In S. Jones (Ed.), *Doing internet research* (Vol. 4). Thousand Oaks, CA: Sage.

Cohn, D., & Hofmann, T. (2001). *The missing link: A probabilistic model of document content and hypertext connectivity.* Paper presented at the Advances in Neural Information Processing Systems, Denver, CO.

Constantinides, E., & Fountain, S. J. (2008). Web 2.0: Conceptual foundations and marketing issues. *Journal of Direct, Data and Digital Marketing Practice*, **9**, 231–244.

Erickson, B. H. (1979). Some problems of inference from chain data. *Sociological Methodology*, **10**, 276–302.

Gao, Y., & Vaughn, L. (2006). Web hyperlink profiles of news sites: A comparison of newspapers of USA, Canada and China. *ASLIB Proceedings*, **57**, 398–411.

Goodreau, S. M. (2007). Advances in exponential random graph (p*) models applied to a large social network. *Social Networks*, **29**, 231–248.

Goodreau, S. M., Hunter, D. R., & Morris, M. (2005). *Statistical modeling of social networks: Practical advances and results.* Seattle, WA: University of Washington.

Halavais, A. (2008). The hyperlink as an organizing principle. In J. Turow & L. Tsui (Eds.), *The hyperlinked society: Questioning connections in the digital age.* Ann Arbor, MI: University of Michigan Press.

Harwit, E., & Clark, D. (2001). Shaping the internet in China: Evolution of political control over network infrastructure and content. *Asian Survey*, **41**, 377–403.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**, 604–632.

Kleinberg, J., & Lawrence, S. (2001). The structure of the web. *Science*, **294**, 1849–1850.

Kleinberg, J., Raghavan, P., & Gibson, D. (1998). *Inferring Web communities from Link topology.* Paper presented at the HyperText 98, Pittsburgh, PA.

Kohut, A. (2008). *Audience segments in a changing news environment.* Washington, DC: The Pew Research Center for the People and the Press.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, **1**(1), 1–41.

Li, J., & Zaiane, O. (2004). *Combining usage, content, and structure data to improve web site recommendation* (Vol. 3182). Heidelberg, Germany: Springer Berlin.

Lin, C. A., & Salwen, M. B. (2006). Utilities of online and offline news use. In X. Li (Ed.), *Internet newspapers.* Mahwah, NJ: Erlbaum.

Lin, J., Halavais, A., & Zhang, B. (2007). The blog network in America: Blogs as indicators of relationships among US cities. *Connections,* **27**(2), 15–23.

Monge, P. R., & Contractor, N., S. (2003). *Theory of communication networks.* New York, NY: Oxford University Press.

Monge, P. R., Heiss, B., & Margolin, D. (2008). Network evolution in organizational communities. *Communication Theory*, **18**(4), 449–477.

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections*, **25**(1), 49–61.

Park, H. W., Barnett, G., & Nam, I.-Y. (2002). Hyperlink-affiliation network structure of top Web sites: Examining affiliates with hyperlink in Korea. *Journal of the American Society for Information Science and Technology*, **53**, 592–601.

Park, H. W., & Jankowski, N. W. (2008). A hyperlink network analysis of citizen blogs in South Korea politics. *Javanost: The Public*, **15**(2), 57–74.

Park, H. W., & Thelwall, M. (2008). Link analysis: Hyperlink patterns and social structure on politician's Web sites in South Korea. *Quality and Quantity*, **43**, 687–697.

Patterson, T. (2007). *Creative destruction: An exploratory look at news on the Internet.* Boston, MA: Joan Shorenstein Center on the Press, Politics and Public Policy.

Rantanen, T. (2001). The old and the new. *New Media & Society*, **3**(1), 85–105.

Rice, R. (1982). Communication networking in computer-conference systems: A longitudinal study of group roles and system structure. *Communication Yearbook*, **6**, 925–944. Newbury Park, CA: Sage.

Richter, A. (2008). Post-Soviet perspective on censorship and freedom of the media. *International Communication Gazette*, **70**, 307–324.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, **29**, 19.

Rogers, E. (1985). *Diffusion of innovations.* New York, NY: Simon and Schuster.

Rogers, E., & Kincaid, D. (1981). *Communication networks: Toward a new paradigm for research.* New York, NY: Free Press.

Rogers, R., & Marres, N. (2000). Landscaping climate change: A mapping technique for understanding science and technology debates on the World Wide Web. *Public Understanding of Science*, **9**, 141–163.

Savage, S., & Waldman, D. (2005). Broadband Internet access, awareness and use: Analysis of United States household data. *Telecommunications Policy*, **29**, 615–633.

Schumate, M., & Dewitt, L. (2008). The north/south divide in NGO hyperlink networks. *Journal of Computer-Mediated Communication*, **13**, 405–428.

Schumate, M., & Lipp, J. (2008). Connective collective action online: An examination of the hyperlink network structure of an NGO issue network. *Journal of Computer Mediated Communication*, **14**, 178–201.

Seary, A., & Richards, W. (2000). Fitting to p* models in Multinet. *Connections*, **23**(1), 84–101.

Snijder, T., Pattison, P., Robins, G., & Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology*, **36**(1), 99.

Thelwall, M. (2004). *Link analysis: An information science approach.* San Diego, CA: Academic Press.

Valente, T., Coronges, K., Lakon, C., & Costenbader, E. (2008). How correlated are network centrality measures? *Connections*, **28**(1), 16–26.

Varian, H. R., Farrell, J., & Shapiro, C. (2004). *The economics of information technology.* Boston, MA: Cambridge University Press.

Wang, P., Robins, G., & Pattison, P. (2005). *PNet 1.0.* Melbourne, Australia: University of Melbourne.

Wang, P., Robins, G., & Pattison, P. (2006). *PNet user manual.* Melbourne, Australia: University of Melbourne.

Wigand, R. T. (1988). Communication network analysis: History and overview. In G. H. Goldhaber & G. Barnett (Eds.), *Handbook of organizational communication.* Norwood, NJ: Ablex.

Xia, M., Huang, Y., Duan, W., & Whinston, A. B. (2007). Implicit many-to-one communication in online communities. In C. Steinfeld, A. Pentland, M. Ackerman, & N. Contractor, S. (Eds.), *Proceedings of the third communities and technologies conference* (pp. 265–274). London, England: Springer.

Zhang, X. (2007). Breaking news, media coverage and 'citizen's right to know' in China. *Contemporary China*, **16**, 535–545.