

Web Infrastructures and Online Attention Ecology

HARSH TANEJA¹

University of Illinois at Urbana Champaign, USA

ANGELA XIAO WU

New York University, USA

The Web has given rise to novel infrastructures for content curation and user content generation. Existing literature connecting website features and user behavior concentrates on a small number of popular websites such as Facebook, Twitter, Google Search, and Wikipedia. This study raises the analytical scheme from the particularities of singular websites to “formats” shared across sites. Through measures we employ to gauge distinct aspects of attention, we analyze traffic to the world’s most popular sites at three time points and determine their website formats with the assistance of human coders. We found that, compared with websites invested in producing original content, the curatorial websites tend to serve as attention hubs, and websites driven by user-generated content serve as attention bridges. Methodologically, this study advances the project of capturing the interconnections within the online attention ecology. Empirically, our findings show how this ecology evolved in relation to Web infrastructures.

Keywords: format, UGC, curation, Web usage, Web infrastructures, online attention ecology

In addition to traditional media’s logic of attracting and holding attention, Web infrastructures enable new mechanisms of content generation and attention curation. The Web ecosystem entails a large number of intermediaries between content producers and media users. These intermediaries, companies such as Facebook and Google, use algorithmically generated hyperlinks and recommendations to curate attention (i.e., direct people to specific corners of the Web). In the process, they profit enormously from advertising and harvesting user data. Further, the Web-based industries also design and deploy technical facilities to solicit contributions from ordinary users to maximize the user base. While digital curation amounts to rent extraction from the user’s exchange of attention for original content, websites based on user-generated content essentially utilize user contribution as the fuel for growth.

Harsh Taneja: harsh.taneja@gmail.com

Angela Xiao Wu: angelaxwu@nyu.edu

Date submitted: 2018–06–01

¹ Both authors contributed equally to this manuscript.

Copyright © 2019 (Harsh Taneja and Angela Xiao Wu). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at <http://ijoc.org>.

The recent literature focused on Web usage has predominantly concentrated on social media and, to a smaller extent, search engines. Our study intends to move the discussion from case studies of the usage of singular websites (e.g., Facebook, Twitter, Google Search, Wikipedia, and so on) to a broader view in which larger online attention ecology is included, and usage patterns of these sites are understood in relation to Web use more holistically. Our study considers attention as the aggregated usage of websites globally, a conceptualization we draw from Webster (2011) that "public attention is the extent to which multiple individuals . . . are exposed to cultural products across space and/or time" (p. 46). Operationalized in this manner, gaining public attention is tied directly to revenue potential for website owners (Davenport & Beck, 2013; Franck, 1999; Napoli, 2003). Shifting the perspective, our study makes the following contributions to the literature on Web usage: (1) It raises the analytical scheme from the particularities of singular websites to "formats" shared across sites, (2) it methodologically advances the project to capture the interconnections of online attention ecology, and (3) it provides empirical findings to address the research question: How did online attention ecology evolve in relation to Web infrastructures?

First, we draw from existing literature on Web usage to formulate a rationale that differentiates website formats along the dimension of attention curation and that of content generation. Next, to complement the common measure of discrete websites' popularity, we use two distinct measures from network analysis to gauge the relational aspects of attention, which are crucial to addressing the research question at hand. Our attention data consist of shared usage between the world's 850 most popular websites at three historical time points sourced from a global Internet measurement panel. We classify all the websites in terms of infrastructural features and examine how these technological features are related to the attention patterns.

Web Use and the Digital Native Formats

The burgeoning literature on online information consumption and social, cultural, and political activities more generally tends to focus on major social media and search engines. Two key features that these websites embody emerge from this literature: content generation and attention curation. First, the feature of user-generated content (UGC) constitutes a major topic of social media research. Signaling a blurred line between the producers and consumers of content (Carpentier, 2011; Napoli, 2014), UGC sustains the "participatory cultures" on the Web (Jenkins, 2006). In terms of attention ecology, case studies of various social media sites have noted that the UGC feature can facilitate user growth by involving individuals who are socially related to the amateur content producers (i.e., "network effects"). Furthermore, created bottom-up without institutional selection or sanction with a priori criteria, UGC may attract the attention of consumers from diverse backgrounds. For instance, bilingual blogs effectively bridge different language spheres (Zuckerman, 2007). Likewise, many YouTube videos gain popularity in cultures quite distant from their countries of production (Platt, Bhargava, & Zuckerman, 2015). In addition to social media, the transnational appeal of Wikipedia also has been primarily driven by content generation from large communities of user editors (Bruns, 2008).

The second feature, traffic curation, has garnered increasing scholarly and media attention in more recent years. A growing literature has noted that various websites have built-in mechanisms that perform, albeit latently, the gatekeeping function in traditional media environments (Dylko, 2015; Thorson & Wells, 2015). Social media again are prominent sites for attention curation that supposedly exert power to shape the

whole attention ecology (DeNardis & Hackl, 2015). A small number of mostly American-based social media sites are seen as acting as the new centers of global media consumption (Jin, 2013). Research has established connections between social media users' personal networks and news usage, which inform the site's curatorial algorithms, and their levels of ideological polarization and civic engagements (e.g., Kim & Chen, 2015; Lee, Choi, Kim, & Kim, 2014). Facebook, for instance, has also taken over large parts of the distribution of specialized content providers, including news organizations (Bell, 2016).

Existing literature has also examined search engines' traffic curation based on their indexing and ranking techniques. Representing some of the most popular sites on the Web, search engines such as Google and Baidu are an absolute necessity for billions to navigate the World Wide Web. Search engines' curatorial logics, which operate through algorithms, are too sophisticated and profitable to allow public access. Moreover, people tend to employ simpler queries and view fewer result pages, albeit a possible result from constantly improving search algorithms, at the same time indicating growing reliance on the algorithm (Jansen & Spink, 2006). As many studies have revealed, though the rendering of search results may stay clear of human interference, serious biases are built into the rendering algorithm (Ballatore, 2015; Jiang, 2014; Segev, 2010). Hence, in the aggregate, these sites exert formidable influence over attention flows (Halavais, 2013; Vaidhyanathan, 2011).

To generalize from findings about these particular websites to gain a broader understanding of Web attention ecology in relation to the digital features of content generation and attention curation, we invoke the register of "format." *Format* is a widely used term in media and communication research. The television industry has historically referred to formats as "programming ideas [such as reality, scripted, or live television] that are adapted [usually from abroad] and produced [into programs] domestically" (Waisbord, 2004, p. 359; Williams, 1974). In the context of digital media, authors have examined MP3 and the weblog as formats (Siles, 2011; Sterne, 2012). "All formats presuppose particular formations of infrastructure with their own codes, protocols, limits, and affordances" (Sterne, 2012, p. 15). There is an important analytical distinction between *format* and *genre*, which may be seen as two levels of categorization. When adopting a format, the content being produced may take a variety of recurrent themes—that is, genres (Devitt, 2004). For instance, within the microblog format arise genres such as celebrity microblogging and government microblogging. Compared with genres, website formats are more enduring because they take much longer to stabilize and gain (or lose) purchases (Siles, 2011).

The Internet comprises multiple scales of technical networks and different layers that interact with users in varied manners (Sandvig, 2013). The Internet's digital infrastructures can be seen as a kind of "media structure" in which media users enact their preferences (Webster, 2014). Such infrastructures are composed of algorithmic rendering of search results, hyperlinks embedded in social media feeds, recommender systems of shopping sites, and so forth. For this study, we conceptualize content generation and attention curation as two orthogonal dimensions of website-level format. In other words, each website may be seen as taking on a particular curatorial format and also a particular generative format. Like an "ideal type," each format stresses certain elements that are common to most cases and not meant to perfectly correspond to all the

characteristics of any particular case (i.e., website). In our following discussion and operationalization, we consider a website embodying one format from each dimension based on its most prominent features.²

Curatorial Formats: From Open to Closed

As the literature reviewed suggests, the multifarious “curatorial technologies” native to the Internet have garnered much concern regarding their capacities to organize, capitalize, and manipulate public attention. Our first dimension thus addresses curatorial structures, which builds on Webster’s (2014) “open-closed” typology of digital media structures. Webster considers media structures “open” if they enable people to go anywhere on the Web (e.g., search engines), and “closed” if they limit people to a predetermined corner of the Web (e.g., news websites that link only to their affiliate sites). Viewed through this lens of curatorial formats, the Web consists of a complex combination of attention holders and attention curators. The former provide texts, images, and services that attract attention, whereas the latter direct attention along various curatorial logics, functioning as an interface between users and attention holders. Transforming attention into traceable and “actionable” resources, the power of attention curators vis-à-vis attention holders is one major aspect that distinguishes today’s digital environment from the traditional media environment.

As the prototypical curatorial infrastructure of the Web, “Search Engine” as a curatorial format³ is located at open end of the spectrum. All a search engine does is transport user attention elsewhere. Characteristic of social media sites, “Social Network” is a particular curatorial format that both curates and holds online attention.⁴ This means that it carries numerous outward hyperlinks but also hosts content and benefits from retaining attention on-site. While social media sites curate content *and* attention, what sets them apart from search engines is the central role of personal networks in constituting their logic of curation. To a nonmember, sites such as Facebook and Pinterest simply provide no front page; even on websites boasting the range of their content offerings to encourage registration, the user experience is contingent on creating one’s own personal network. Because of this curatorial logic, the user’s exposure to content on social media is directly shaped by his or her network of friends, connections, or other subscriptions that have been personally assembled over time and built into the infrastructure of the interface.

Toward the closed end of the spectrum of attention curation, we identified three more curatorial website formats. Among the most typical attention holders, also the closest extension of the model followed in traditional broadcasting, are sites adopting what we call the “Content Producer” format. These sites focus on

² The eventual typology is informed by both the existing literature and an iterative process whereby the authors worked through 50 randomly sampled Web domains from the lists of the world’s 1,000 most popular websites (the methods section has details about these lists).

³ In this article, we use capitalized phrases to refer to these ideal type formats.

⁴ To refer to the curatorial format embodied by social media sites, we use *Social Network* instead of *Social Media* because the phrase *Social Network* places more emphasis on the logic of curation rather than content (Murthy, 2012). To be clear, we consider social media sites to exhibit the Social Network (curatorial) format and the UGC (generative) format (see the next section). For example, MySpace’s owner once illustrated the site’s nature with a circle drawn between “content” and “distribution”: “it shares aspects of both” (Reiss, quoted in Gillespie, 2010, p. 351).

the production of original content, or the “primary accumulation” of attention on the Web. The scope may range from timely information or commentaries about current affairs (e.g., *Wired* magazine and *Huffington Post* sites) to leisurely or specialized content that is not time sensitive (e.g., *Harvard Business Review* and *Yelp*). One of the earliest Web formats, “Web Portal,” embodies the most straightforward mechanism to “give order amid chaos,” presenting the user with a uniformly organized portfolio comprising aggregated content (Tatnall, 2005). Through selecting and foregrounding content from diverse sources and likely with a top-down approach relying on professional editors, Web portals shape the attention ecology. Like content producers, Web portals attempt to hold attention rather than direct it elsewhere on the Web. However, unlike content producers that build a distinct brand for content production, portals concentrate more on *content* curation. Examples include generic megasites offering a wide range of (acquired) content and services such as Yahoo, AOL in the United States, and China’s NetEase.

Unlike traditional media, the Web enables a combination of both symbolic and utilitarian consumption. While content producers and portals deal with “symbolic” or “experience goods,” e-retail sites—embodying the “E-Retail” format—manage “material” or “utilitarian goods.” These sites profit from transactions that users make in exchange for products or services, although sometimes they provide both free and premium accounts (e.g., Dropbox). For example, in our scheme, Apple’s online store for selling its gadgets is an e-retail website, but we regard the sites of video distributor Netflix and game distributor Steam as portals that curate experience goods, even though the latter extracts a middleman’s fee from the transactions it facilitates. E-retail sites rarely point their users outward and thus typify the most closed of the basic Web formats. Figure 1 schematizes our typology.

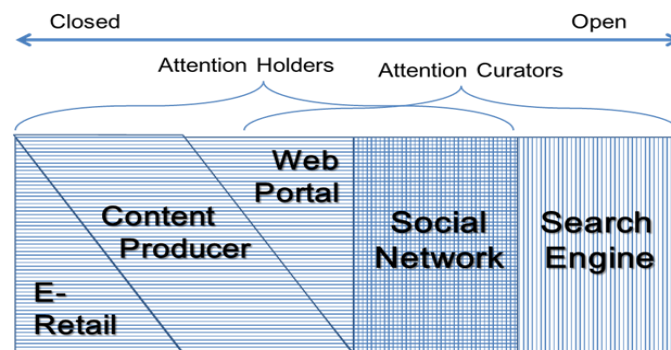


Figure 1. Classification of curatorial website formats.

Generative Formats: UGC and Non-UGC

The design and deployment of technical facilities to solicit contributions from ordinary users is another distinct characteristic that fuels growth of Web-based industries. As seen in the case of social media, when websites provide a toolkit enabling users to create and populate content, they acquire the additional potential to shape the attention ecology. To capture this affordance, we consider a separate, “generative” dimension that addresses the website format for content production. Along this dimension, when a website

embodies a “UGC” format, the site owner designs and maintains the technological platform, but the main content is left for users to supply. In the absence of these generative features, the site embodies a Non-UGC format.

The generative format and the curatorial format are orthogonal. According to our definition, Wikipedia and Yelp adopt the Content Producer format (curatorial) and the UGC format (generative), while those content producers employing or subcontracting to recognized professionals such as *The New York Times* are Non-UGC. A UGC e-retail site is one that hosts numerous sellers or service providers, facilitating and benefiting from transactions between them and the site’s users. These individual sellers, rather than the website itself, have the rights to close deals. Examples include eBay, Uber, and Airbnb. A UGC search engine would rely primarily on users’ active contributions to building its index, an example being isoHunt, a search engine for torrents. In the case of social media, its Social Network format designates a type of curatorial infrastructure and its UGC format a mode of content generation.

Having generated a typology of formats along the two dimensions, we next advocate three distinct measures to problematize the contours of the online attention ecology in relation to Web infrastructures.

Capturing the Attention Ecology: Popularity, Junctionality, and Transversality

In scholarship and popular forums alike, *popularity*—that is, the volume of usage—tends to be the main indicator of attention. For instance, at 1.5 billion monthly unique users, Google.com gets more “attention” than a website with 1.5 million users. Beyond the amount of attention each discrete online outlet receives, popularity measures provide no information about how users’ attention is shared between websites or how a site is situated in “connected and interlinked minds and eyes” (Read, 2014, para. 18). With this dimension alone, we are unable to comprehend the ecological and relational structure of online attention and, in turn, how emergent curatorial and generative infrastructures are embedded and how they connect to the more traditional formats in terms of the attention ecology on the Web.

To do so, first, instead of analyzing websites as isolated entities, one needs to consider them in relation to one another through shared attention, gauged from shared usage (Hindman, 2009). Hypothetically, for instance, if, of Pinterest.com’s 100 visitors, 10 also visit AllRecipes.com (which has 60 visitors), these 10 visitors’ visits represent the shared attention between Pinterest.com and AllRecipes.com. Our conceptualization of shared attention online resembles audience duplication between websites, a focal variable of several recent studies on digital media use (e.g., Taneja & Wu, 2014).

Theoretically speaking, in the network of shared attention—that is, the attention ecology—visitors of equally popular websites could differ in their usage patterns across the Web. Different from popularity or the amount of attention a site receives, the quality of the site’s attention would depend on how many and which other sites its visitors also visit in conjunction. Thus, using this lens, attention is no longer treated as fungible and discrete. Therefore, drawing from network analysis, we advocate two additional dimensions to capture how a site is located in relation to the larger ecology of attention, as well as the nature of attention it receives.

Junctionality refers to a website's ability to attract attention from people who use other sites, which makes it "well connected," occupying the concentrated spots in the attention ecology. Put differently, a site takes the position of a "junction" when a user tends to use it regardless of what online repertoires he or she has. For example, if people choose Yahoo as the vantage point of their browsing routine, or simply set it as their browser's home page, the junctionality of Yahoo is more likely to be high. The same goes for search engines such as Google, if these are what everyone constantly turns to when surfing the Web. Junctionality corresponds with weighted degrees in network analysis, which is a weighted sum of the number and value of a node's direct connections to all other nodes. In the case of an attention network, the weighted degree is thus the number of websites a site shares attention with, as well as the number of duplicated audiences with every such site. In our conceptualization, nodes with high weighted degrees can be considered sites with high junctionality, or junctions of shared attention.

Despite its conceptual elegance and computational simplicity, junctionality/weighted degrees only factor a node's direct connections. Relying on this dimension alone, one may miss out on nodes that are not knit into tightly connected clusters, but bridge different ones. For this, we use *transversality*—which in mathematics describes how differential spaces intersect—to identify a website's ability to attract attention from people with disparate attention patterns. Put differently, if two populations of users each focus on distinct sets of websites, the few websites that both populations frequent would be high in transversality (Figure 2). In fact, we know that Web usage around the world when aggregated appears as a mosaic of regional cultures; each comprises a distinct set of websites, and the "global" websites—that is, sites used by culturally diverse populations—are rapidly diminishing in number (Wu & Taneja, 2016). Therefore, in the context of global attention ecology, where websites cluster based on shared attention, the transversal sites are those that occupy the bridging areas between these clusters. For example, Chinese Wikipedia has high transversality; it attracts attention from Chinese-speaking users worldwide, as well as some in China, because of its liberal ethos. Transversal websites may not be the highest in junctionality. In the previous example, this highly transversal Chinese Wikipedia has much lower junctionality compared with Baidu Baike, which, thanks to temporary blockage by the Chinese government, has thrived and retained China-based users. Similarly, junctionality cannot fully capture transversality. Baike is low in transversality because it receives attention from people who largely visit China-based websites, rather than dissimilar groups of websites, as in the case of Chinese Wikipedia.

Transversality corresponds with *current flow betweenness*, an adaption of the more commonly used *betweenness centrality*, a measure of a node's bridging function in social networks. A node's betweenness centrality is the proportion of all shortest paths between any two nodes that pass through it. By definition, the measure is more appropriate to use in unweighted networks. Adapted for weighted networks, betweenness centrality can mislead because all information would not flow only through the shortest path, but in inverse proportion to a tie's value. To capture the bridging function in such networks more efficiently, network scientists have derived a measure called current flow betweenness, which assumes that instead of flowing through the shortest path, information spreads through all available paths like an electric current through all available paths, in inverse proportion to their resistance. We used this measure because in our network, a node's current flow betweenness indicates the capacity of that website to bridge attention between any two websites (which are otherwise not directly connected).

Thus, sites with high current flow betweenness can be considered high in transversality, or sites that bridge the various clusters of global online attention.

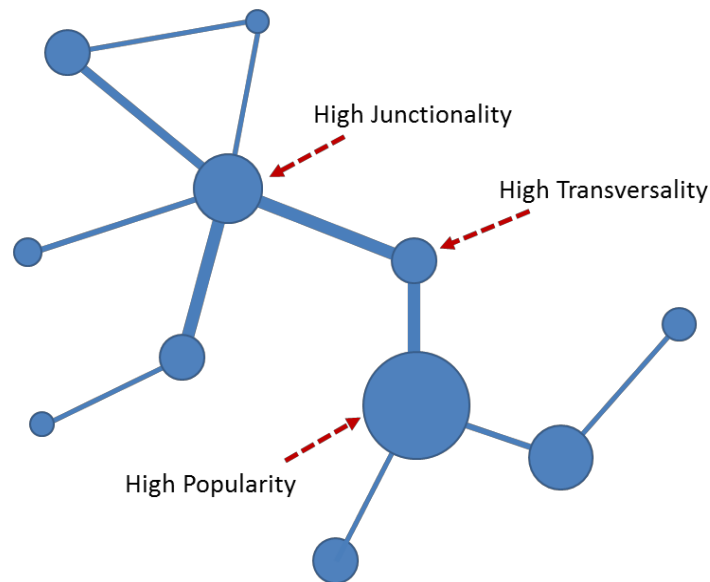


Figure 2. Three measures of attention in an attention ecology.
Each node represents a website, and the node size represents the amount of attention it receives. Each tie represents the attention that two websites share, and the tie thickness represents the amount of shared attention.

We hypothesize some possibilities based on the characterizations of Web usage in the literature (as discussed in a preceding section). First, for junctionality, we expect Social Network and Search Engine (curatorial formats at the open end of the spectrum) to have, on average, higher junctionality than Web Portal, Content Producer, and E-Retail (formats at the holding end). Second, between a site's junctionality and its generative format, we do not anticipate any such aggregate effects. In other words, UGC websites would not, on average, have higher junctionality than Non-UGC sites. Third, for transversality, we expect, on average, sites with a UGC format to have higher transversality than sites without.

Method

Web Usage Data

We sourced Web usage (traffic) data from comScore, a panel that provides continuously metered Internet audience measurement data once a month from 2 million users worldwide in 170 countries. At three time points (September 2009, 2011, and 2013), we obtained shared user traffic between the world's

top 1,000 Web domains (ranked by monthly unique users). The aggregated "WorldWide" data from comScore are based on Web usage on personal computers (PCs) and do not include usage from mobile devices. Given the shift in devices used for Web access, we believe that analyzing PC data alone for 2014 or later would be a partial representation of global online attention and incomparable to previous years for our research question.

The websites in our sample account for 99% of the global Web user visits, thus ensuring an adequate representation of sites in different languages and for different geographies. We excluded advertising networks or advertising servers because the majority of traffic to such sites is a by-product of users' attention to other websites. After excluding these, our final sample consisted of 849 sites in 2009 and 850 sites each in 2011 and 2013. The 2013 sample had 407 sites that were not present in 2011, and the 2011 sample had 348 sites that were not present in 2009.

Coding Formats

After closely examining 50 randomly sampled websites in light of curatorial and generative dimensions, respectively, we finalized the typology and described the rationale in a codebook. Next, we employed a group of coders with specific language skills and contextual knowledge to collectively code our entire sample (top 1,000 websites annually appearing in 23 languages) according to the classification of Web formats, as described earlier. In addition to having native-level fluency in the languages in which they coded, we required that all coders be proficient in English.

Given the difficulty of finding at least two coders in each language, it is impossible to test intercoder reliability in the traditional sense. However, to mitigate reliability concerns, we developed an iterative process to train coders. The largest number of websites in our sample are in English and Chinese. We first walked each coder through the codebook and assigned a select set of English-language websites. We then examined the results and interviewed the coder to ensure that he or she followed the logic of our codebook. Likewise, our Chinese language coder coded a subset of Chinese websites together with one of the coauthors with Chinese proficiency. Through this exercise, only when we were convinced of the coders' understanding of our codebook did we recruit them for coding sites in their own languages (such as Russian, Chinese, and German). For many languages with smaller online populations, the sites included in our sample (which is restricted to the world's 1,000 most popular sites) tend to be fairly well known in their respective linguistic spheres, reducing the chances of coding errors. In cases that involved fewer than a dozen websites (e.g., Turkish and Dutch websites), one of the coauthors coded the sites in the presence of a native speaker.

After compiling all the codings, we also performed "sanity checks" to investigate anomalies by rather unlikely combinations of "curatorial" and "generative" codes. These checks include, for example, looking for the co-occurrence of Content Producer and UGC and that of Social Network and Non-UGC. After the initial coding was completed, we employed another coder to comb through the compiled codings to ensure that ad networks were correctly identified (to be further eliminated from our final sample). Because of prior involvement in our research projects, this last coder was experienced with our coding scheme; we had extensively perused his former coding results, which proved highly reliable.

For the original codebook we used to train and guide our coders, see <https://www.dropbox.com/s/sq8pzt6cgufahyr/Codebook-form3.pdf?dl=0>. Based on these manual coding endeavors, we are able to determine which of the five curatorial formats a website belonged to, and whether it is UGC or not.

We randomly sampled 200 websites, of which 89 had content in English and 62 in Chinese. We hired two undergraduate seniors as research assistants with native Chinese proficiency from a large midwestern American university. We had both of them code these 151 sites after reviewing our codebook. We achieved high intercoder reliability with Krippendorff's alpha = 0.81 for the five curatorial formats and Krippendorff's alpha = 0.87 for classifying a site as UGC/Non-UGC.

Analysis

Popularity: Visitor Numbers

Overall, in each of the three years, the popularity of websites in our sample follow a "long-tail" distribution, with a small share of websites commanding the lion's share of users.

In each year, about 7% of the sites were search engines. In 2009 and 2011, 11% of the sites were social networks, and this number dropped to 5% in 2013. The combined incidence of websites belonging to these two attention-curating formats was between 12% and 18%. Among attention holders, Web portals made up more than one third of all sites in each year, and their proportion ranged from 34% in 2009 to 37% in 2013. The distribution of formats in our sample across the three years remained largely stable, except these minor changes just noted. This is remarkable given the churn in our sample; between 2009 and 2011 and between 2011 and 2013, our sample had about 40% replacement of websites. In Figure 3, we show the incidence of various website formats in our sample. Other than the base format, the chart includes the composition of each format in terms of UGC sites or otherwise (Non-UGC). UGC sites constitute between 16% and 19% of our sample.

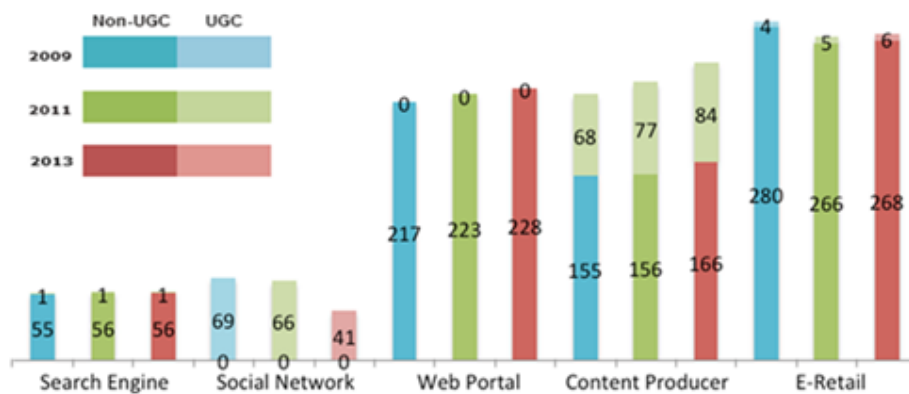


Figure 3. Incidence of popular websites by format.

Given the long-tail distribution of traffic patterns in general, we provide a closer examination of the “megapopular” sites to see how attention was distributed across formats within this group. We followed Webster and Ksiazek (2012) and considered megapopular sites as those with a reach greater than 2% of total Web users. Figure 4 shows the ratio of the incidence of each format in this subsample to its incidence in the overall sample as a base, indexed at 100. An index of 100 indicates a 1:1 ratio, and an index of 150 denotes a ratio of 1.5:1, indicating that the category occurred 50% more often in the subsample. Overall, the incidence of formats among the megapopular sites is not very different from their incidence in the overall sample. Comparatively, the attention-curating formats Search Engine and Social Network are more likely to be megapopular, and Content Producer less so, among attention-holding formats. Taken together, among the megapopular sites, attention curators have a greater incidence (17%–21%) compared with their proportion in the full sample (12%–15%).

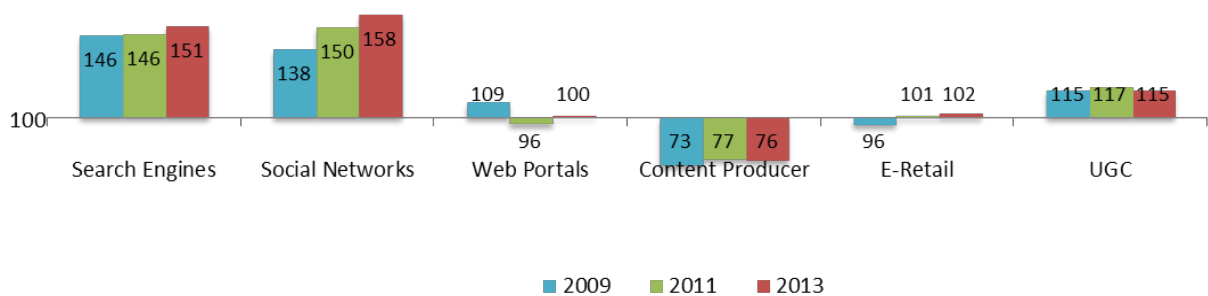


Figure 4. Ratio of incidence of megapopular websites to total websites by format.

To assess the average popularity of each format, we first calculated the percentage of the total Internet user base that each site received each year (as a means of normalization) and then calculated the mean of the percentages of all sites belonging to a format. As Figure 5 shows, in terms of content generation, compared with non-UGC sites, UGC-driven websites on average are far more popular. Along curatorial dimension, Content Producer has the least average popularity, followed by E-Retail and Web Portal. All these three attention-holding formats fall short compared to the attention-curating formats (i.e., Search Engine and Social Network). Moreover, the latter, especially Social Network, have become more popular on average.

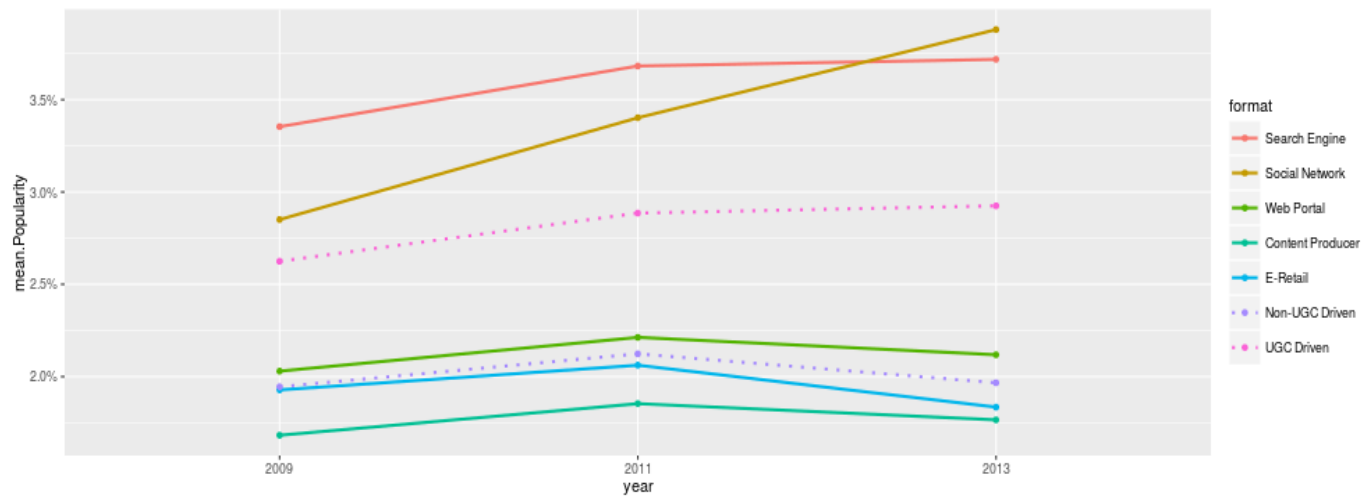


Figure 5. Normalized average popularity by format.

In Figure 6, we present the traffic distribution by formats in each year. The vertical axis indicates the site's popularity, measured as the logged unique Web users reached by the site in the month the data were collected. The horizontal axis represents the occurrence of websites corresponding to each user traffic level for each format. In 2009, the most popular site was a search engine. From 2009 to 2013, a few social networks and content producers joined search engines as the most popular websites. In contrast, the popularity of the most popular e-retailer (in 2009) diminished in 2013. Further, non-UGC sites enjoyed peak popularity in 2009, whereas UGC sites rose in popularity in 2011 and 2013, driven by a few extremely popular websites (possibly social networks and UGC content producers). Because Figure 6 shows both occurrence and individual popularity level, one can infer from it the overall prevalence of each format. Based on popularity, e-retail sites, on the whole, appear the most salient.

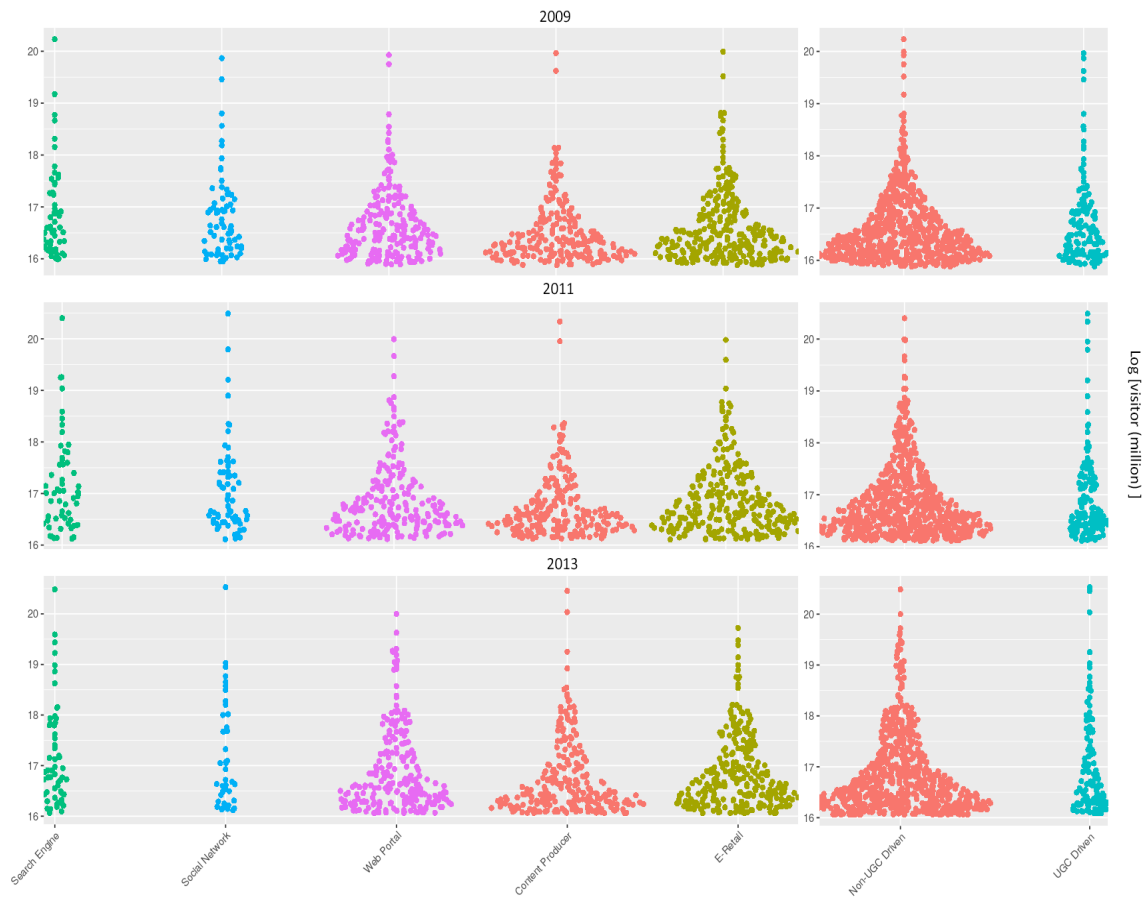


Figure 6. Distribution of website popularity (logged) by format.

Attention Ecology: Network Analysis

In addition to examining absolute user attention to each website (i.e., popularity), we also mapped global online attention through a relational lens. The focal unit of our analysis is “shared attention,” which we measured through audience duplication. The latter is the extent to which two media outlets (e.g., websites) are consumed by the same set of people in a given time period. In a hypothetical universe of 100 people, if on a given day, 30 people accessed both BBC.com and Yahoo.co.jp, the shared attention (audience duplication) between these two websites would be 30, or 30%. For each website in our sample, we obtained (using comScore) its audience duplication with all other websites in the same annual sample. Our final data set has 359,976 $[(849 \times 848) / 2]$ pairs of audience duplication denoting shared attention between websites in 2009, and 360,825 pairs $[(850 \times 849) / 2]$ each in 2011 and 2013.

We conceptualized the ecology of online attention as an attention network, with websites as the nodes and the shared attention (audience duplication) as the ties between nodes. While adapting the

general approach from recent studies of audience duplication between digital outlets (e.g., Taneja & Wu, 2014), we made one significant departure. All these studies analyzed networks with dichotomous ties, where they considered two nodes as tied (or not) if they had duplicated audiences above (or below) a statistically determined expected level of audience duplication. Instead of considering dichotomous ties, in the present study, we analyzed networks with valued ties that took into account the number of duplicated audiences between two websites. For instance, if any two websites A and B had 3.5 million audiences that visited both these sites during an observation period, we considered the network tie between them to have a value of 3.5 million. Next, we present our analysis of these resulting networks for each of the three years focusing on node (website) level centrality.⁵

Junctionality and Transversality: Weighted Degrees and Current Flow Betweenness

To characterize positions of individual websites in terms of junctionality and transversality, we calculated their weighted degrees and their current flow betweenness. To establish associations between these measures and website formats along both dimensions, we ran a series of analyses of covariance (ANCOVAs) and linear regressions. In each of these, we controlled for visitor numbers. Further, we logged all amount and count variables (such as degree centrality measures and visitor numbers) to make their positively skewed long-tail distributions symmetric.

First, we modeled the junctionality as a dependent variable (DV) with the curatorial formats as the predictors. For the initial analysis, we reclassified the combination of social networks and search engines as “attention curators” and the combination of e-retailers, professional producers, and Web portals as “attention holders.” For each year, we fitted ANCOVA models with junctionality as the DV and controlled for visitor numbers and the curatorial format (reclassified). These suggested that means of junctionality significantly differed according to the curatorial format even after controlling for visitor numbers. Post hoc tests (using Tukey HSD [honestly significant difference]) confirmed that, in each year, the mean junctionality of attention-curating websites was significantly higher than that of websites that hold attention. We repeated this analysis for junctionality retaining all five curatorial formats and found that in each year, search engines and social networks had greater mean junctionality than content producers. Likewise, social networks had higher junctionality than e-retailers in all three years, and so did search engines, although only in 2011 and 2013. Within attention-holding websites, Web portals had higher mean junctionality than content producers in 2009 and 2013, but we found no significant differences between portals and e-retail on this measure. Table 1(a) summarizes the key findings.

We repeated the described analysis with transversality as the DV. Here again, we found that attention curators had significantly higher average transversality scores than attention holders. When we ran the ANCOVAs with all five curatorial formats—see Table 1(b)—we found search engines to be significantly higher in mean transversality than content producers in all three years.

⁵ The three networks have the same number of nodes with matching weighted degree scores. Only the clustering coefficient increased in 2013, indicating the emergence of more closely connected communities.

**Table 1. Comparison of Means Based on Tukey Post Hoc Analysis
(a)(i) Junctionality (weighted degrees), pairwise mean differences***

	Search Engine	Social Network	Web Portal	Content Producer	E-Retail
Search Engine	X	–	–	2009, 2011, 2013	2009, 2013
Social Network		X	2011	2009, 2011, 2013	2011, 2013
Web Portal			X	2013	–
Content Producer				X	
E-Retail					X

(a)(ii) Base ANCOVA model statistics**

Model	2009			2011			2013		
	df	F	η sq	df	F	η sq	df	F	η sq
Base Category	4	8.06	.01	4	8.87	.01	4	16.91	.03
Visitors (Log)	1	1728.16	.66	1	1526.61	.63	1	1694.93	.65
Residual	843			844			844		

All F values are significant with $p < .0001$. The partial η sq are reported (calculated as sum of squares for factor/total sum of squares).

(b)(i) Transversality (current flow betweenness), pairwise mean differences*

	Search Engine	Social Network	Web Portal	Content Producer	E-Retail
Search Engine	X	–	2013	2009, 2011, 2013	2013
Social Network		X	2013	2011, 2013	2013
Web Portal			X	2013	–
Content Producer				X	
E-Retail					X

(b)(ii) Base ANCOVA model statistics**

Model	2009			2011			2013		
	df	F	η sq	df	F	η sq	df	F	η sq
Base Category	4	6.27	.01	4	8.85	.02	4	10.21	.02
Visitors (Log)	1	1158.52	.57	1	1236.82	.58	1	806.59	.48
Residual	843			844			844		

All F values are significant with $p < .0001$. The partial η sq is reported (calculated as sum of squares for factor/total sum of squares).

*A cell in this section contains the year if in that year, the mean of the format listed in the rows is significantly greater than the mean of the format listed in the columns. Only significant differences ($p < .01$) are reported.

**In each ANCOVA model, curatorial format significantly differentiated the mean of the DV after controlling for visitor numbers.

Finally, we ran three regression models each to examine the association between a site of UGC format and its transversality and junctionality (Table 2). After controlling for visitor numbers (which were highly correlated with transversality), we found no significant association between the generative dimension and its transversality in 2009, but these associations became significant and more sizable in 2011 and 2013. Thus, in 2011 and 2013, given equal popularity (which was a positive predictor of transversality), UGC websites were more transversal than Non-UGC websites, and this gap, although nonsignificant in 2009, became significant in 2011 and increased further in 2013. For junctionality, the converse was true. We found website junctionality to have a significant positive association with it being UGC-driven in 2009 and 2011, but no such association was significant in 2013.

Table 2. Regression Models With Transversality and Junctionality as Dependent Variable.

	Transversality			Junctionality		
	2009	2011	2013	2009	2011	2013
UGC	0.027	0.084**	0.111**	0.046*	0.131**	0.079
Visitor (Log)	0.965**	0.943**	0.897**	0.967**	0.947**	0.960**
Intercept	-22.419	-22.278	-21.487	3.980	4.31	4.01
R^2	.880	.855	.825	.890	.890	.732
N	849	850	850	849	850	850

** $p < .01$. * $p < .05$.

Discussion: Web Infrastructures and Online Attention Ecology

Our findings at first glance may suggest that transversality and junctionality at least empirically manifest rather similarly in the attention ecologies we have mapped. Particularly for the megapopular sites, we see higher correlations between these measures.⁶ However, we also observed that an increase in transversality is not associated with an increase in junctionality. As global Web usage diverges to become more clustered yet sparse, which we found in 2013 and also is consistent with existing research (e.g., Wu & Taneja, 2016), our empirical results show the increased divergence between these two measures compared with previous years. Specifically, in 2013, when the difference between the transversality of UGC and Non-UGC was highest, the difference in the junctionality of these two formats was insignificant. Thus, the growing salience of disparate attention clusters does accompany the emergence of bridging sites in between; under such circumstances, there has been an increase in transversal sites that are not highly junctional, as well as the converse. Based on our results, we can speculate that as the Web grows to include users from hitherto low penetration countries and regions, we will see a greater divergence between junctionality and transversality, rendering them even more useful in mapping the attention ecology.

Along the curatorial dimension, we found that among popular websites, the occurrence of attention holders—Web Portals, Content Producers, and E-Retailer—is between 4 and 5 times that of search engines and

⁶ We ran a number of correlations between junctionality and transversality separately for each year by format along each dimension and separately for megapopular and other sites. We found that these two measures were highly convergent in 2009 and 2011, but diverged in 2013. Further, in each year, the divergence was greater for sites outside the megapopular sites and for Non-UGC sites.

social networks. Even though, on average, they are not as popular as attention curators, taken as a whole, each of these formats is quite prevalent on the World Wide Web. The generous presence of e-retail among the world's most popular sites calls for a reconsideration of the common assumption of the Internet as new "media," a space facilitating the exchange of symbolic meanings. Our analysis also shows that the Social Network format has more presence in the megapopular range but less among the generally popular sites (i.e., our annual samples). The concentration of such curatorial power in fewer hands is especially alarming, because the presence of a handful of social media sites monopolizing the reconfiguration of public attention has crucial consequences for the news industry and the democratic process at large.

Beyond the popularity measure, the network measures we employed reveal the positions of attention curators to be more central in the attention ecology. Specifically, we found both Social Network and Search Engine, the two attention-curating formats, to be highly transversal and junctional, scoring significantly higher on these measures than Content Producer (in all three years), E-Retail (in at least two of the three years), and Web Portal (in 2013). This suggests that the average user visits (the smaller number of) attention-curating sites, regardless of what other sites she visits, and it is reasonable to assume that these sites lie on her paths of online navigation. These findings support the premise that the algorithms' role in influencing patterns of online attention is not merely technical in nature, but also institutional (Napoli, 2014).

Especially alarming in these results is the weak position of content-producing websites, the only group of sites that supply original content for public consumption. These (along with e-retailers) seem to acquire scattered fragments of attention organized around attention-curating sites. Further, content producers are relatively underrepresented among megapopular sites, and the few that are tend to be driven by UGC. Web portals, once exceedingly dominant destination sites on the Internet, fare somewhat better than content producers, yet are relatively peripheral in the present curator-dominated attention ecology. In 2013, Web Portal's junctionality is higher than Content Producer's, while its transversality is lower than that of Social Network and Search Engine. This suggests that portals provide a common ground for people with similar attention patterns, however, they fail to act as sites that bridge disparate attention patterns. The largely country-specific geographic focus of Web portals could explain this pattern.

Along the generative dimension, our findings show that the UGC sites are a minority in terms of incidence among the world's top 850 sites (websites with professionally produced content outnumber UGC sites by 3 to 4 times in our samples). Notwithstanding their low incidence, UGC sites are highly influential in the attention ecology. On average, a UGC site is more popular than a non-UGC site, and the format's presence is disproportionately high among the megapopular sites. When examined through a relational lens, the prominence of UGC becomes even clearer. Over time UGC, websites tend to be less junctional but more transversal in the attention ecology. Between 2009 and 2013, as UGC sites increased in popularity, their transversality relative to Non-UGC sites also increased. In contrast, their junctionality in 2013 was not (significantly) higher than Non-UGC sites. Combined with the general increase in the clustering of the network, this suggests the bridging role that UGC sites have gradually occupied on the Web. We can also infer that among equally popular sites, UGC sites are more likely to take bridging positions on the Web than Non-UGC. Indeed, the reliance on this content generation model not only capitalizes on the contribution of ordinary users but also is likely to help a website accumulate cross-cultural appeal.

Conclusion

Scholars including Henry Jenkins (2001) and Yochai Benkler (2006) have long advocated ecological approaches for understanding and theorizing “today’s interconnected, globalized, and diverse, and complex communications system” (Wahl-Jorgensen, 2016, p. 17). Viewed through an ecological perspective, the global attention patterns online result from user preferences coevolving with the Web’s massive infrastructural and technological developments (Sandvig, 2015). Our investigation takes an ecological approach by conceptualizing and empirically demonstrating the associations between attention patterns and Web infrastructures in a holistic manner.

Many prior studies have professed the power or ability of various website characteristics to affect user behavior. These have often focused on a few well-known sites, such as Facebook, YouTube, Twitter, Google Search, and Wikipedia. Although many have used diverse empirical research methods, including content analysis, surveys, and on-site data mining, few have provided a systematic examination of how Web infrastructural traits shape online attention patterns. To move this research agenda forward, we abstract from existing literature two types of website-level infrastructures, or what we call website formats. The first type is curatorial, which can be located along the continuum of open versus closed digital media structures. The second type, generative website formats, concerns whether ordinary users or professional creators drive their content generation.

Moreover, to interrogate findings in existing literature about roles that website infrastructures play in shaping usage, we examine the attention ecology holistically and from a relational lens. For this, we advocate two network analytic measures to determine each website’s position in this ecology. These measures capture junctionality and transversality, two distinct aspects of attention that differ from websites’ popularity.

We then apply this analytical scheme connecting user attention and website formats to a large sample comprising of the world’s most popular sites at three recent time points and empirically demonstrate how website formats fare on these measures. Our findings show that traditional broadcast-based mechanisms of attracting an audience and holding their attention are not driving websites to the center of the attention ecology; it is the Internet’s native attention-curating formats that dominate. Furthermore, sites driven by UGC are especially likely to also bridge disparate user communities defined by shared usage of websites.

Despite the fast replacement rate of Web offerings, we identified a set of specific website formats that proved to be stable between 2009 and 2013. This typology illuminated the dynamic character of the Web infrastructure that organized, and developed in response to, the global attention landscape. In the following years, especially after the prevalence of mobile access, the Internet has evolved into a complex “technological hybrid” without essential characteristics (Sandvig, 2015, p. 236). However, the curatorial and the generative dimensions along which we conceptualized these formats will stay relatively fixed. Future studies would benefit from adapting these conceptual dimensions as they would from drawing on our empirical operationalization of the attention ecology. Our study hence provides a useful beginning for a continuing empirical research program to examine linkages between Web infrastructures and patterns of online attention.

References

- Ballatore, A. (2015). Google chemtrails: A methodology to analyze topic representation in search engine results. *First Monday*, 20(7). doi:10.5210/fm.v20i7.5597
- Bell, E. (2016, March 7). Facebook is eating the world. *Columbia Journalism Review*. Retrieved from http://www.cjr.org/analysis/facebook_and_media.php
- Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. New Haven, CT: Yale University Press.
- Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and beyond: From production to produsage*. New York, NY: Peter Lang.
- Carpentier, N. (2011). New configurations of the audience? The challenges of user-generated content for audience theory and media participation. In V. Nightingale (Ed.), *The handbook of media audiences* (pp. 190–212). Chichester, UK: Wiley-Blackwell.
- Davenport, T. H., & Beck, J. C. (2013). *The attention economy: Understanding the new currency of business*. Boston, MA: Harvard Business Press.
- DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, 39, 761–770. doi:10.1016/j.telpol.2015.04.003
- Devitt, A. J. (2004). *Writing genres*. Carbondale, IL: Southern Illinois University Press.
- Dylko, I. B. (2015). How technology encourages political selective exposure. *Communication Theory*, 26(4), 389–409. doi:10.1111/comt.12089
- Franck, G. (1999, December 7). The economy of attention: Decline of material wealth. *Telepolis*. Retrieved from <http://www.heise.de/tp/artikel/5/5567/1.html>
- Gillespie, T. (2010). The politics of "platforms." *New Media & Society*, 12, 347–364. doi:10.1177/1461444809342738
- Halavais, A. (2013). *Search engine society*. Malden, MA: Polity.
- Hindman, M. S. (2009). *The myth of digital democracy*. Princeton, NJ: Princeton University Press.
- Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, 42, 248–263. doi:10.1016/j.ipm.2004.10.007

- Jenkins, H. (2001). Convergence? I diverge. *Technology Review*, 104, 93. Retrieved from http://phase1.nccr-trade.org/images/stories/jenkins_convergence_optional.pdf
- Jenkins, H. (2006). *Convergence culture*. New York, NY: New York University Press.
- Jiang, M. (2014). Search concentration, bias, and parochialism: A comparative study of Google, Baidu, and Jike's search results from China. *Journal of Communication*, 64, 1088–1110. doi:10.1111/jcom.12126
- Jin, D. Y. (2013). The construction of platform imperialism in the globalization era. *Communication, Capitalism & Critique*, 11(1), 145–172. doi:10.31269/triplec.v11i1.458
- Kim, Y., & Chen, H.-T. (2015). Discussion network heterogeneity matters: Examining a moderated mediation model of social media use and civic engagement. *International Journal of Communication*, 9, 2344–2365. Retrieved from <https://ijoc.org/index.php/ijoc/article/view/3254>
- Lee, J. K., Choi, J., Kim, C., & Kim, Y. (2014). Social media, network heterogeneity, and opinion polarization. *Journal of Communication*, 64, 702–722. doi:10.1111/jcom.12077
- Murthy, D. (2012). Towards a sociological understanding of social media: Theorizing Twitter. *Sociology*, 46, 1059–1073. doi:10.1177/0038038511422553
- Napoli, P. M. (2003). *Audience economics*. New York, NY: Columbia University Press.
- Napoli, P. M. (2014). Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication Theory*, 24, 340–360. doi:10.1111/comt.12039
- Platt, E. L., Bhargava, R., & Zuckerman, E. (2015, April). The international affiliation network of YouTube trends. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (pp. 318–326). Oxford, UK.
- Read, J. (2014, December 18). Distracted by attention. *The New Inquiry*. Retrieved from <http://thenewinquiry.com/essays/distracted-by-attention/>
- Sandvig, C. (2013). The Internet as infrastructure. In W. Dutton (Ed.), *The Oxford handbook of Internet studies* (pp. 86–106). Oxford, UK: Oxford University Press.
- Sandvig, C. (2015). The Internet as the anti-television: Distribution infrastructure as culture and power. In L. Parks & N. Starosielski (Eds.), *Signal traffic: Critical studies of media infrastructures* (pp. 246–278). Chicago, IL: University of Illinois Press.
- Segev, E. (2010). *Google and the digital divide*. Oxford, UK: Chandos.

- Siles, I. (2011). From online filter to Web format: Articulating materiality and meaning in the early history of blogs. *Social Studies of Science, 41*, 737–758. doi:10.1177/0306312711420190
- Sterne, J. (2012). *MP3: The meaning of a format*. Durham, NC: Duke University Press.
- Taneja, H., & Wu, A. X. (2014). Does the great firewall really isolate the Chinese? Integrating access blockage with cultural factors to explain Web user behavior. *The Information Society, 30*, 297–309. doi:10.1080/01972243.2014.944728
- Tatnall, A. (2005). *Web portals: The new gateways to Internet information and services*. Hershey, PA: Idea Group.
- Thorson, K., & Wells, C. (2015). Curated flows: A framework for mapping media exposure in the digital age. *Communication Theory, 26*(3), 309–328. doi:10.1111/comt.12087
- Vaidhyathan, S. (2011). *The Googlization of everything*. Berkeley, CA: University of California Press.
- Wahl-Jorgensen, K. (2016). The Chicago school and ecology: A reappraisal for the digital era. *American Behavioral Scientist, 60*, 8–23. doi:10.1177/0002764215601709
- Waisbord, S. (2004). McTV: Understanding the global popularity of television formats. *Television & New Media, 5*, 359–383. doi:10.1177/1527476404268922
- Webster, J. G. (2011). The duality of media: A structural theory of public attention. *Communication Theory, 21*, 43–66. doi:10.1111/j.1468-2885.2010.01375.x
- Webster, J. G. (2014). *The marketplace of attention*. Cambridge, MA: MIT Press.
- Webster, J. G., & Ksiazek, T. B. (2012). The dynamics of audience fragmentation: Public attention in an age of digital media. *Journal of Communication, 62*(1), 39–56. doi:10.1111/j.1460-2466.2011.01616.x
- Williams, R. (1974). *Television: Technology and cultural form*. London, UK: Routledge.
- Wu, A. X., & Taneja, H. (2016). Reimagining Internet geographies: A user-centric ethnological mapping of the World Wide Web. *Journal of Computer-Mediated Communication, 21*(3), 230–246. doi:10.1111/jcc4.12157
- Zuckerman, E. (2007). Meet the bridgebloggers. *Public Choice, 134*, 47–65. doi:10.1007/s11127-007-9200-y