The Effect of Links and Excerpts on Internet News Consumption

Jason M.T. Roos* Carl F. Mela† Ron Shachar‡

24 April 2020

Abstract

Internet news and search sites often excerpt content from and link to competing news outlets. On the one hand, providing outbound links can make the linking site more attractive, even to the point of stealing traffic from the linked sites. Regulatory policy, such as the European Union's Copyright Directive Article 15 taxing links, is predicated in part on this idea. On the other hand, receiving inbound links can increase a linked site's audience by informing readers about its news content that day. To explore these opposing perspectives, the authors develop a dynamic learning model and fit it to browsing and link data from celebrity news sites. They then simulate how banning links affects consumer browsing, and find that linking increases celebrity news consumption, especially among consumers who browse the least. On average, linking benefits both the linking and linked sites. The authors estimate that exposure to a link increases the likelihood of visiting the linked site by .14%. This increase is approximately three times the commonly reported click through rate for paid display advertisements.

Keywords: News consumption, Hyperlinking, Structural models, Learning models, Dynamic programming, Bayesian estimation

*Jason M.T. Roos is Associate Professor of Marketing, Rotterdam School of Management, Erasmus University, Netherlands (email: `jroos@rsm.nl`).

†Carl F. Mela is T. Austin Finch Foundation Professor of Marketing, Fuqua School of Business, Duke University, USA (email: `mela@duke.edu`).

‡Ron Shachar is Professor of Marketing and Economics, Arison School of Business, Interdisciplinary Center (IDC) Herzliya, Israel (email: `ronshachar@idc.ac.il`).

The authors would like to thank comScore for the data used in this study as well as Peter Arcidiacono, Andrew Ching, Andres Musalem, Ken Wilbur, Marshall Van Alstyne, Hema Yoganarasimhan, and seminar participants at Duke (Fuqua, Dept. of Economics), Erasmus (RSM, ESE), Ohio State University, Yale University, Washington University in St. Louis, Georgia Tech, Tilburg University, University of North Carolina Chapel Hill, University of San Diego, University of Michigan, INSEAD, University of Frankfurt, University of Houston, University of Pennsylvania, University of Toronto, University of Rochester, the 34th ISMS Marketing Science Conference, the 2012 HEC Marketing Camp, the 11th ZEW Conference on the Economics of ICT, the 2015 Marketing in Israel Conference, the 13th Marketing Dynamics Conference, and the 2016 Summer Institute in Competitive Strategy for their comments. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. This article stems from the first author's dissertation.

On October 2, 2009, celebrity news website *The Superficial* reported about homophobic comments rapper 50 Cent had made about Kanye West. *The Superficial's* article excerpted content from and linked back to another news article—also about West and 50 Cent—that was published at *Celebuzz,* another celebrity news site. Thus, anyone who read *The Superficial's* article but had not yet visited *Celebuzz* would have learned something about *Celebuzz*'s content that day. Importantly, this knowledge might have affected what these readers chose to do next. Fans of rap music, for example, might have been more likely to visit *Celebuzz* that day, whereas others might have been less likely.[1] Although this example comes from celebrity news (the empirical context for this article), linking among news sites is a key feature of internet news—of all types—that sets it apart from print news. Links to other news sites provide information about the linked sites' content that consumers would otherwise not observe. Thus, links play an important role in online news consumption by helping readers locate interesting content (and avoid uninteresting content) more efficiently.

The purpose of this study is to gain a better understanding of how linking among internet news sites affects demand for online news. We aim to measure how much the likelihood of visiting a linked celebrity news site changes after encountering a particular link, while accounting for the possibility that consumers anticipate and value outbound links when choosing which sites to visit. Motivated by recent regulatory initiatives, such as Article 15 of the European Union (EU) Copyright Directive—the so-called "link tax"—and legislation that led to the withdrawal of *Google News* from Spain, we consider the implications of a policy of banning links on the consumption of online news. Implicit in these regulations is the belief that links and excerpts are mostly harmful to news publishers. The idea is that by appropriating content from linked sites, linking sites steal audience share from the sites they link to, thereby decreasing the linked sites' traffic and advertising revenues. However, because excerpts inform readers about the linked sites' content, they can potentially increase the linked site's audience and revenues. Google estimates that news excerpts in search results drive 8 billion clicks per month to European publishers (Gingras 2019). We consider both perspectives about the effect of links on traffic and find evidence that links to news sites can be more beneficial to the linked sites than harmful. Quantification of these link effects is an important first step in measuring the welfare effects of link taxes.

Throughout this study, we make a distinction between two effects of linking on consumer demand for online news. One effect arises after a consumer encounters a specific link and learns about the linked site's content on that particular day. We refer to this learning as the *within-session* effect of linking, as its influence on the consumer's choices is confined to the remainder of that day's browsing session. The other effect arises when a consumer, before visiting a site, assigns it a higher value because the site tends to provide useful links. We refer to this enhanced value as the *across-session* effect, because the higher value (1) depends on the consumer's knowledge of sites' long-run average content and linking behaviors, and (2) affects which sites

[1] Because excerpts are almost always accompanied by a hyperlink to the excerpted site, and as our empirical study relies on hyperlinks to indicate when excerpting has occurred, we use the terms *links* and *excerpts* interchangeably.

consumers tend to visit in the early steps of all browsing sessions.

We measure both of these effects and further assess the net impact of banning links on (1) total traffic at the linking and linked sites, (2) the frequency with which consumers browse for news, and (3) the number of sites consumers visit in each session. These insights are relevant to (1) content producers, who need to know how linking affects their traffic (and, thus, advertising revenue); (2) policy makers—such as the EU—who need to understand how excerpting affects consumer demand for news; and (3) advertisers, who need to know how changes in linking affect the reach and frequency of ads running on multiple sites.

We consider the effects of linking on consumers and news sites by developing and estimating a structural model of demand for online news. The structural approach enables us to assess a counterfactual policy of banning links prospectively, rather than waiting to observe such a policy in data. The structural approach also facilitates a decomposition of linking effects into the within- and across-session effects just described. The combination of these effects can be either positive or negative for the linked sites. Thus, this decomposition of link effects both motivates and enriches the counterfactual policy analysis.

At its core, our model describes sequential news consumption with learning among consumers with heterogeneous opportunity costs from browsing and horizontal tastes for news (the latter means, for example, that some readers might enjoy reading about Kanye West, but others might not). A consumer's utility from reading a site's content depends on the consumer's match with what the site published that day. Due to the nature of news, consumers are *ex ante* uncertain about what each site has published each day. Therefore, at the start of each browsing session, consumers are uncertain about their horizontal match with each site that day.

Each link provides a signal about consumers' (heterogeneous) horizontal match utilities with the linked site's content on that day. Because these links are informative about daily variation in horizontal match, their within-session effect can be to increase the likelihood of visiting the linked site for some consumers and *decrease* it for others. In both cases, encountering a link lowers uncertainty and, thus, (on average) leads to better browsing choices later in the session. We consider a model with forward-looking individuals and contrast this model with one in which consumers are not forward looking. Forward-looking consumers anticipate that encountering links will decrease their uncertainty about their daily match with the linked sites. This anticipation is the source of the across-session effect, whereby forward-looking consumers place a higher expected value on sites that frequently link to others. We show that this higher valuation can lead to higher traffic for the linking site, but either higher or lower traffic at the linked site.

Because the net impact of linking on site traffic depends crucially on the particulars of a news ecosystem, the question of whether links are beneficial or not is fundamentally empirical. We conduct such an empirical analysis using internet panel data describing browsing at five celebrity news sites, which we augment with data describing the daily news content and links published at those sites. Preliminary analysis of the raw browsing and link data shows that for more than half of the panelists, the likelihood of visiting a site is *lower* after encountering a link to that site. This outcome is consistent with our modeling framework, which allows the within-session effect of observing a link to either increase or a decrease this likelihood. When the baseline probability of visiting a site is already low, this probability can increase much more than it can decrease due to a floor effect. For this reason, the aggregate effect of encountering a link (averaged across panelists) in the raw data is positive.

To assess how banning links affects browsing, we first fit the data to our structural model, and subsequently use the estimates for counterfactual simulations. The model estimates provide a view into how these celebrity news sites differentiate from one another, both vertically and horizontally. The results also underscore the importance of links to the consumers who visit these news sites. In this empirical setting, encountering a link lowers consumers' uncertainty about their daily match with the excerpted site by approximately 6%. The results further show that consumers value this reduced uncertainty and thus find sites that provide outbound links more attractive. In total, linking raises the value of reading celebrity news.

Although the data reflect choices made on days when sites both linked and did not link to each other, browsing on days without links occurred in a world where linking was allowed (meaning consumers could anticipate the possibility of finding links). Estimating a structural model thus allows us to consider a counterfactual policy of banning links that is not observed in the data. These results show that among these news sites, the total effect of linking is positive for both consumers and the sites. Compared with a counterfactual without linking, the median consumer visits .54% more sites, and total traffic at the five sites is between .01% and .18% higher. The benefits from linking accrue to sites at different steps of consumers' browsing sessions. Due to the across-session effect, providing outbound links helps some sites gain visitors early in consumers' browsing sessions. Due to the within-session effect, receiving inbound links helps some sites gain visitors later in consumers' browsing sessions. When we consider only individuals who encountered links, we find exposure to those links adds .14% to the probability of visiting the linked site—a 2.3% increase over the baseline. This increase in visit probability may at first appear small. However, compared against a relevant baseline of click-through rates for display ads, which are often at or below .05% (Lambrecht and Tucker 2013; Lewis, Rao, and Reiley 2011; Chaffey 2017), the effect is substantial.[2]

The remainder of this study is organized as follows. First, we discuss our study's contribution in the context of the prior literature. Next, we present the structural model and discuss its main behavioral implications. We then describe our data, and discuss issues related to estimation and model identification. Finally, we present the model estimates and results of the counterfactual simulations, before summarizing the implications and limits of this study.

[2] Estimates of advertising cost per thousand impressions on the internet vary widely, but are typically upwards of \$.50 (Karlštrems 2019; Pratskevich 2018). Using \$.50 as a conservative estimate of the typical value a firm places on promoting its content, the relative value of an inbound link could be upwards of \$.50(.14/.05) = \$1.40 per thousand impressions at the linking site.

Contribution and Related Literature

This study builds on previous work in marketing and economics that has modeled internet browsing both at the aggregate (Danaher 2007; Park and Fader 2004) and individual levels (Goldfarb 2002; Johnson et al. 2004; Lee, Zufryden, and Drèze 2003). Among these studies, our model is most similar to that of Goldfarb (2002). We describe utility-maximizing individuals choosing which site to visit next, in consideration of their past browsing decisions and any outbound links they expect to encounter. In our model, however, encountering outbound links does not generate utility *per se*. Instead, such a link provides the consumer with a positive or negative signal about the linked site's content. Thus, in our model, outbound links make the linking site more attractive because they help consumers make better browsing choices later in the session. Moreover, the extent of this increased attractiveness varies across consumers depending on (1) the set of sites that are typically linked, and (2) consumers' average preferences for those sites.

Although our study focuses on the demand implications of linking, this study is also related to previous theoretical work that has considered the process by which sites link to one another (Katona and Sarvary 2008; Mayzlin and Yoganarasimhan 2012; Dellarocas, Katona, and Rand 2013; Jeon and Nasr 2016). This work shows how links can play an important role in helping uninformed consumers discover new sites and learn about their typical news content. In equilibrium, sites end up serving a mix of experienced consumers—those who already know about the sites' typical content and outbound linking decisions—and inexperienced consumers—those who are as of yet uninformed about these. We do not model such a process of site discovery. Rather, we condition on its outcome—the content and links observed in our data—and estimate their effects on experienced consumers' demand for news sites.

Our model is grounded in the consumer learning literature in marketing (Erdem and Keane 1996; Ching, Erdem, and Keane 2013; Ching, Erdem, and Keane 2017). In our model, consumers sequentially choose which site to visit next in the current browsing session. This choice is made with uncertainty about the news content at each site. However, with each site visit, there is the potential to observe outbound links, and thereby obtain signals about the consumption utility from *other* sites. Our study is therefore related to previous work that has modeled consumer learning about a good via advertising or information spillovers from consuming related goods. Because our empirical setting involves individuals consuming news content that changes every day, we observe, for each consumer, multiple repetitions of a learning process that starts with the same initial condition (not having read the news yet).

We make two methodological contributions to the Bayesian learning literature. The first of these is related to our model. Although the main focus of our model is on consumers learning about horizontally differentiated site content after observing outbound links—and the Bayesian learning model we use to capture these dynamics is standard in the literature (Ching, Erdem, and Keane 2013)—there are other dynamics, also relevant in a news consumption setting, that motivate a second Bayesian learning process. Specifically, different news outlets can publish the same basic news facts, but consumers only gain utility from their first encounter with those facts. Thus, if two sites publish many of the same news facts, the marginal utility from the second site's content will be lower after visiting the first. Because the utilities from the two sites' news content are correlated, after visiting the first site, there is the potential for Bayesian updating about the amount of unknown content that remains at the second site. Thus, we augment the main model (horizontal differentiation and linking) to include a vertical dimension of utility with Bayesian learning based on the daily volume of basic news facts published at each site. We first conceptualize these news facts as distinct bits, each of which represents a unique piece of information capable of generating utility when first encountered (Allen 1983; Allen 1986; Allen 1990). Using this foundation, we then consider how these bits are consumed under uncertainty. The resulting Bayesian learning model is novel to the consumer learning literature. We believe this approach could be a useful basis for studying consumption utility in the context of news and other information goods. Moreover, this approach is general enough to be used to study sequential choices over alternatives with correlated utilities in other settings (e.g., retail store visits).

The second methodological contribution pertains to our estimation procedure, which combines two advances from the econometrics and statistics literatures, and provides a template for efficient Bayesian estimation of single-agent dynamic discrete choice models. Our approach to estimation is based primarily on that of Imai, Jain, and Ching (2009, hereinafter IJC). Compared to the standard nested fixed point algorithm for estimating dynamic discrete choice models (Aguirregabiria and Mira 2010), IJC's method requires significantly fewer computational resources (Ching, Imai, et al. 2012). Although IJC's computational advantages are great, the method still produces samples that can be highly autocorrelated. Thus, we further improve efficiency by using Girolami and Calderhead's (2011) full manifold Metropolis adjusted Langevin algorithm (MMALA) to construct high-quality proposal distributions for the Metropolis-Hastings accept/reject steps in the IJC algorithm. This approach decreases autocorrelation in the resulting sample chains and improves the rate of convergence to the posterior distribution.

Finally, although limited in scope to the empirical setting of celebrity news, our findings contribute to an emerging empirical literature that seeks to understand how the internet affects news consumption (Gentzkow and Shapiro 2008; Gentzkow and Shapiro 2011; Gentzkow, Shapiro, and Sinkinson 2011; Flaxman, Goel, and Rao 2016). Some of this work has looked at how large news aggregators—*Google News* in particular—affect the amount of traffic going to linked news sites (George and Hogendorn 2019; Posada de la Concha, García, and Cobos 2015; Chiou and Tucker 2017; Athey, Mobius, and Pál 2017; Majó-Vázquez, Cardenal, and González-Bailón 2017; Calzada and Gil 2019). Studies in this literature typically exploit sudden changes in *Google News's* linking behavior due to market entry, copyright lawsuits, or legislation. These studies have shown that the aggregator's outbound links can increase traffic to smaller or more horizontally differentiated news publishers, while having a less positive, or possibly negative total effect on larger or more mainstream news sites. As a pure news aggregator, *Google News* does not create any news content of its own. By contrast, the sites we consider primarily publish original celebrity news, while also excerpting from and linking back to one another. Sites such as these generate a substantial portion, if not the majority of the links and excerpts most readers will encounter when consuming news. We contribute to this literature by considering the impact of links originating from sites that publish original news content, studying individual consumers rather than aggregate traffic, structurally modeling the entire news browsing sequence, assessing the separate effects of linking within and across sessions, assessing how links affect demand at different steps of the browsing session, and simulating a counterfactual policy of banning links.

Modeling Framework

A defining characteristic of news, whether online or offline, is its uncertainty. Consumers do not know exactly what a news site has published until *after* they visit the site and see its content (otherwise, the content isn't *news* to the consumer). The importance of this point for understanding how links affect news consumption is illustrated by the previous example of the link to *Celebuzz*'s coverage of 50 Cent and Kanye West. A consumer who likes reading about rap artists might have *experienced* higher-than-normal utility from visiting *Celebuzz* that day, while a reader who dislikes rap artists might have *experienced* lower than normal utility. In either case, the consumer could not have *anticipated* this difference in utility unless they knew something about *Celebuzz*'s coverage ahead of time.

Links provide consumers with this type of knowledge. Anyone who saw *The Superficial's* article, which excerpted and linked back to *Celebuzz*, would have learned something about *Celebuzz*'s coverage that day. Thus, consumers who *like* reading about rap artists might have been *more* likely to visit *Celebuzz* after seeing the link. At the same time, not all consumers want to read the same type of news. Thus, consumers who *dislike* rap might have grown *less* likely to visit *Celebuzz* after seeing the link.

This example highlights the core of our model. Consumers have heterogeneous *horizontal* preferences for differentiated news content, meaning that consumers differ in the type of content they like to read. Every day, sites publish new content, leading to variation in and uncertainty about the horizontal utility their readers will receive from reading that content. At the start of each browsing session, consumers are initially unaware of what each site has published. Yet as consumers encounter links to sites they have not visited yet, that uncertainty is reduced. This learning process takes place over the course of a single browsing session, and repeats each day starting with the same initial information state (not knowing the news).

The model describes the choices made by experienced consumers of online news. These consumers are *certain* about the type of content sites tend to publish *on average*, but *uncertain* about what those sites publish *on any given day*. We assume that based on their past browsing, these experienced consumers already know the sites' stable, long-run average content behaviors. Specifically, they know which topics the sites typically cover, as well as the frequency with which the sites link to each other.

Previous studies have considered the processes by which sites arrive at these stable, average content and linking behaviors, and consumers come to know them (Katona and Sarvary 2008; Mayzlin and Yoganarasimhan 2012; Dellarocas, Katona, and Rand 2013). We do not model such a process but, rather, assume it has already taken place. We condition on sites' supply of content and links to model experienced consumers' demand for these (we detail our identification strategy later).

To focus attention on the role of links, we first present a model in which sites are only horizontally differentiated in their news coverage. One site, for example, might focus on news about the film industry, and another site on news about reality television. Consumers who like films but dislike reality TV would probably prefer the former over the latter, on average. We subsequently extend the model so that news sites are vertically differentiated according to the amount of news facts they publish.

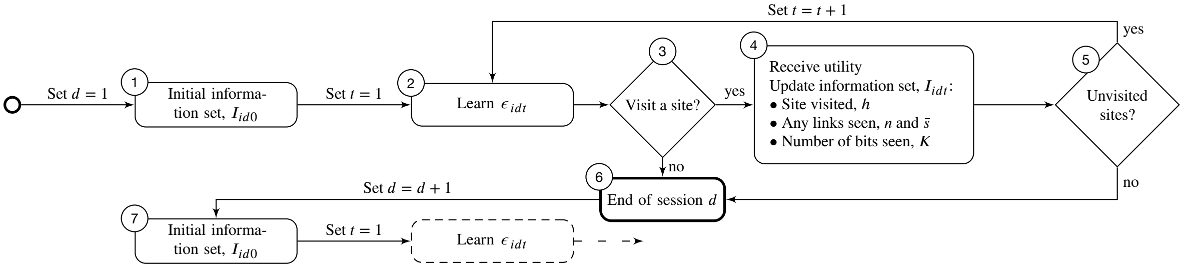
Notation, Timing, and Period Utility

We present the model from the perspective of a single consumer, bearing in mind that different consumers have different preferences for news. Every day, the consumer engages in a browsing session, which is indexed d. By a *browsing session*, we refer to the process of sequentially visiting zero or more sites within a day (visiting zero sites means not browsing that day). Figure 1 depicts the sequence of events within each browsing session (Figure 1 includes notation that we explain subsequently).

At each step of the browsing session, $t = 1, \dots, T_d$, the consumer decides which site to visit next, if any (Figure 1, °3). Visiting a site and viewing its content does two things: (1) it provides utility to the consumer and (2) it changes the consumer's information set (Figure 1, °4). Consumers are indexed with i and the sites with j. When discussing linking, we sometimes refer to the linking site by the index $j = L$, and the site receiving the link by the index $j = R$. The option $j = 0$ denotes the option to end the browsing session. The set $\mathcal{J}_{id1}$ contains the J sites under consideration, plus the option to end the session.

We follow the literature on sequential browsing online and assume the consumer sees all available content at each site visited, and therefore visits each at most once per session (Kim, Albuquerque, and Bronnenberg 2010). This assumption matches both the empirical context of celebrity news sites (whose home pages display all content posted each day), and choices observed in the estimation data (which we describe subsequently). Thus, the consumer's choice set, which is initially $\mathcal{J}_{idt}$, is reduced to $\mathcal{J}_{id\,t+1} = \mathcal{J}_{idt} \setminus j$ after visiting site j (Figure 1, °5). At each step t, the consumer must choose which previously unvisited site to visit next, or whether to end the session (Figure 1, °6). We denote by a_{idt} the index of the option j chosen by consumer i at step t of browsing session d.

Figure 1: Schematic Representation of Steps in Browsing Sessions



Notes: (1) Prior to browsing on day d, the consumer's information set is initialized to I_{id0}: none of the sites have been visited ($h = 0$), and thus no links or bits of news information have been observed ($n = 0$, $\bar{s} = 0$, and $K = 0$). (2) Before making a decision at each step t the consumer receives private shocks to utility, ϵ_{idt}. (3) If the present value of expected utility is high enough, a site is visited. (4) Visiting a site reveals its content and links. The information set is now I_{idt}, reflecting any links seen (n and $\bar{s}$), new bits encountered (K), and the site visit itself (h). (5) Unless all sites have been visited, the session advances to the next step $t = t + 1$. (6) If all sites have been visited or the present value of expected utility was too low at (3), the session ends for that day. (7) The next day $d = d + 1$, the consumer starts again with an initial information set, I_{id0}, and the process repeats.

The utility from visiting site j at step t of browsing session d comprises three parts.

$$U_{ijdt} = \mu_{ijd} - \gamma_{id} + \epsilon_{ijdt} \tag{1}$$

The first is μ_{ijd}, denoting the horizontal match utility consumer i gains from reading site j's content on day d. This component of utility is unknown to the consumer before visiting site j, as we discuss next. The second component, $\gamma_{id} > 0$, reflects the opportunity cost of forgoing the outside alternative (not browsing) in favor of reading celebrity news sites. We assume that γ_{id} is known to the consumer, and constant throughout the browsing session. In our empirical setting, we expect the incentive to browse for celebrity news might differ between weekdays and weekends or U.S. federal holidays (Columbus Day, Veterans Day, Thanksgiving, and Christmas). Thus, if day d falls on a weekend or holiday, $\gamma_{id} = \gamma_i \exp\left(\gamma_w\right)$, and otherwise $\gamma_{id} = \gamma_i$. The third component of utility is ϵ_{ijdt}, which is idiosyncratic to each consumer, site, and step of each browsing session. This utility shock is private information learned just *prior to* the decision at step t of the browsing session and is unobserved by the researcher (Figure 1, °2). Ending the session (or not starting a session in the first place) is an endogenous choice, yielding net utility of $U_{i0dt} = \epsilon_{i0dt}$.

Apart from γ_{id}, which affects the overall value of browsing relative to the outside option, the model does not include a day-level component of utility that is *a priori* known to all consumers. Common knowledge of such a component of utility, when the good consumed is news, is difficult to justify, as the day-level fixed effect would imply some foreknowledge of the day's news prior to visiting a news site to learn the day's news.

Horizontal Match Utility from Content

The horizontal component of utility, μ_{ijd}, arises from the match between the site's content on day d and the consumer's preferences. Because news sites post new content every day, the match utility provided to each consumer varies from session to session. A site that typically posts news about film actors, for example, might occasionally report on reality television. On these occasions, a consumer who prefers film over reality TV might experience lower utility from reading the site's content. Accordingly, the news events that take place each day, and which of those events site j chooses to report, influence the daily value of μ_{ijd}.

In these examples, as well as in our model, we make an important distinction between a site's long-run average match with the consumer, and daily deviations from that average. The consumer's long-run average match with a site depends on the type of content the site publishes on average. This long-run average is therefore the same for every browsing session. On the basis of a potentially long history of browsing, an experienced consumer knows their own long-run average match with each site. By contrast, daily deviations from that average arise due to news events and the sites' choices about what to publish. These daily deviations are therefore unknown at the start of each session (Figure 1, °1 and °7). We model the horizontal match utility consumer i receives from site j's content on day d as a function of (1) site j's long-run average position in a horizontal attribute space, z_j; (2) site j's deviation from this average position on day d, ν_{jd}; and (3) the consumer's preferences, υ_i.

$$\mu_{ijd} = \left(z_j + \nu_{jd}\right) \upsilon_i \tag{2}$$

This formulation implies that consumers, on average, prefer sites for which sign $\left(z_j\right)$ = sign $\left(\upsilon_i\right)$.[3] We model the ν_{jd}'s as coming from the following distribution.

$$\nu_{jd} \sim N\left(0, \tau_{\nu}^{-1}\right) \tag{3}$$

Like the z_j's, we assume that consumers know the value of τ_{ν}^{-1} on the basis of their prior browsing.

Signals of Horizontal Match Utility from Links

If the consumer visits site L during session d, and if site L has linked to site R that day, then the consumer will learn something about their horizontal match with site R that day. Site R, for example, might rarely report on rap artists. Site L's link to R's coverage of Kanye West thus signals that R's coverage leans more in the direction of rap artists that day. If site L links to site R on day d, we say $\ell_{\vec{LR}d} = 1$ (and $\ell_{\vec{LR}d} = 0$ otherwise). We denote the signal contained in this link as $s_{\vec{LR}d}$, and model these signals as noisy, but unbiased reflections of sites' true horizontal positions each day.

$$s_{\vec{LR}d} | \ell_{\vec{LR}d} = 1, \nu_{Rd} \sim N\left(z_R + \nu_{Rd}, \tau_s^{-1}\right) \tag{4}$$

Although horizontal position, $z_R + \nu_{Rd}$, is a characteristic particular to site R on each day d, two links to R from two sites L and L' can signal two different aspects of R's horizontal position. Thus, signals are indexed by both the receiving site R *and* the linking site L. The extent to which links accurately signal sites'

[3] Equation 2 allows sites to differentiate along multiple horizontal attributes. In our empirical application, we estimate a single horizontal dimension to parsimoniously capture the main component of horizontal variation. If the true number of dimensions is greater than 1, than the estimated z_j's will be a lower-dimension projection of the sites' positions in this higher-dimensional attribute space. In this case, the z_j's are weighted sums of unobserved attributes, with weights reflecting the attributes' importances for browsing.

daily horizontal match positions is known to consumers and denoted τ_s.

This setup highlights the informative role of linking among news sites. Links help consumers ascertain whether a site's content is more or less congruent with their preferences that day. Importantly, because different values of ν_{jd} imply higher or lower levels of match utility (relative to the site's long-run average), links can signal *lower* than average match (in which case, they make the consumer *less* likely to visit the linked site). Furthermore, because consumers have different horizontal preferences, υ_i, the same link from L to R might make some of L's readers more likely to visit R, and others less likely.

Updated Beliefs About Horizontal Match Utility

Let $n_{ijd\,t-1}$ denote the number of links to site j that consumer i has seen prior to choosing what to do at step t of session d, and $\bar{s}_{ijd\,t-1}$ denote these links' average signal value.

$$n_{ijd\,t-1} = \sum_{a \in \{a_{id1},\ldots,a_{id\,t-1}\}} \ell_{\vec{a}jd} \tag{5}$$

$$\bar{s}_{ijd\,t-1} = \begin{cases} \frac{1}{n_{ijd\,t-1}} \sum_{a \in \{a_{id1},\ldots,a_{id\,t-1}\}} s_{\vec{a}jd} & \text{if } n_{ijd\,t-1} > 0 \\ 0 & \text{if } n_{ijd\,t-1} = 0 \end{cases} \tag{6}$$

Before the consumer sees any links ($n_{ijd\,t-1} = 0$), expected horizontal match utility is simply equal to its long-run average, $\mathbb{E}\left(\mu_{ijd} | n_{ijd\,t-1} = 0\right) = z_j \upsilon_i$. But if the consumer has seen one or more links ($n_{ijd\,t-1} > 0$), their beliefs about j's horizontal match utility change. Standard Bayesian updating for conjugate normal distributions yields an expression for expected horizontal match utility, after seeing $n_{ijd\,t-1}$ links to site j (West and Harrison 1999).

$$\mathbb{E}\left(\mu_{ijd} | n_{ijd\,t-1}, \bar{s}_{ijd\,t-1}\right) = z_j \upsilon_i + \left(\frac{\tau_s n_{ijd\,t-1}}{\tau_s n_{ijd\,t-1} + \tau_\nu}\right)\left(\bar{s}_{ijd\,t-1} - z_j\right) \upsilon_i \tag{7}$$

Expected match utility at site j is thus a weighted average of the consumers's long-run average match, $z_j \upsilon_i$, and the match signaled by previously seen links to j, $\bar{s}_{ijd\,t-1} \upsilon_i$. The weight given to each of these depends on (1) the variability of sites' daily horizontal positions, τ_ν^{-1}; (2) how informative links are about those horizontal positions, τ_s; and (3) the number links the consumer has seen, $n_{ijd\,t-1}$. Equation 7 therefore illustrates the value of links to the consumer: on average, they can shift expectations about horizontal match utility away from their typical long-run values, and toward their true day-specific values. The left side of Table 1 summarizes the variables involved in this within-session, Bayesian updating of expected horizontal match.

Value Function for Present and Future Browsing

When consumers visit a site, they not only gain utility, but also may update their beliefs about match utility at other sites if they see outbound links (Figure 1, °4). As we show subsequently, forward-looking consumers anticipate this updating and thus face the standard exploitation-exploration trade-off when choosing which

Table 1: Summary of Bayesian Updating for μ_{ijd} and β_{ijdt}

Horizontal Utility, μ_{ijd}		Vertical Utility, β_{ijdt}	
Received		Received	
$\left(z_j + \nu_{jd}\right) \upsilon_i$	Horizontal utility from site j on day d	$\left(K_{idt}^{+j} - K_{id\,t-1}\right) \lambda_i$	Vertical utility from site j at step t on day d
υ_i	Horizontal preference	λ_i	Vertical preference
z_j	Long-run average horizontal position	$K_{id\,t-1}$	Bits seen prior to choice at step t of day d
ν_{jd}	Deviation in horizontal position on day d, unknown to consumers	K_{idt}^{+j}	Bits seen after visiting site j at step t of day d, unknown to consumers
Expected		Expected	
$z_j \upsilon_i$	Expected match utility prior to seeing links	$\left(\frac{\alpha_j}{1+\alpha_j}\right) N \lambda_i$	Expected vertical utility before seeing bits
$z_j \upsilon_i + \left(\frac{\tau_s n_{ijd\,t-1}}{\tau_s n_{ijd\,t-1} + \tau_\nu}\right) \left(\bar{s}_{ijd\,t-1} - z_j\right) \upsilon_i$	Expected match utility after seeing links	$\left(\frac{\alpha_j}{1+A\left(h_{id\,t-1}\right)+\alpha_j}\right) \left(N - K_{id\,t-1}\right) \lambda_i$	Expected vertical utility after seeing bits
τ_ν^{-1}	Variance of unknown horizontal position	N	Maximum total bits each day
τ_s	Precision of link signals	α_j	Long-run average daily bits
$n_{ijd\,t-1}$	Number of links seen prior to step t choice	$h_{id\,t-1}$	Sites visited prior to step t choice
$\bar{s}_{ijd\,t-1}$	Average signal value of links seen	$A\left(h_{id\,t-1}\right)$	Sum of α_j's for sites already visited

site to visit next. A consumer, for example, might decide to visit a site that frequently links to many others, expecting any links encountered to increase (decrease) their chance of visiting (avoiding) sites with higher (lower) daily match. For such consumers, sites that provide many outbound links provide value, in part, by raising the expected utility of the remainder of the browsing session.

The following value function corresponds with consumer i's utility function and beliefs about match utility at step t of session d.

$$(8) \quad V\left(I_{id\,t-1}, \epsilon_{idt}\right) = \max\left(\epsilon_{i0dt}, \max_{0<j\in\mathcal{J}_{idt}} \left\{\mathbb{E}\left(U_{ijdt} | I_{id\,t-1}\right) + \delta \int_{I',\epsilon'} V\left(I', \epsilon'\right) f\left(I' | I_{id\,t-1}, j\right) g\left(\epsilon'\right)\right\}\right)$$

Equation 8 introduces the following notation:

- δ determines how much the consumer discounts the future expected utility from browsing,
- $g\left(\epsilon\right)$ is the distribution of the i.i.d. idiosyncratic shocks to utility at each step,
- $I_{id\,t-1}$ indicates consumer i's information state prior to step t of browsing session d, and
- $f\left(I' | I_{id\,t-1}, j\right)$ denotes a transition density reflecting the consumer's beliefs about how this information state evolves (conditional on a visit to some site j).

We describe the latter two bulleted terms next.

Consumer Information Set

The consumer's information set, I_{idt}, includes three variables.[4] The first two, the number of links to each site encountered, $\left\{n_{i1dt}, n_{i2dt}, \ldots, n_{iJdt}\right\}$, and the average signal value of those links, $\left\{\bar{s}_{i1dt}, \bar{s}_{i2dt}, \ldots, \bar{s}_{iJdt}\right\}$, determine the level of expected match (per Equation 7). The third variable, the set of sites that have been visited, is represented as a binary vector, $h_{idt} \in \{0,1\}^J$, and determines the consumer's choice set. The transition function, $f\left(I' | I_{id\,t-1}, j\right)$, reflects the consumer's beliefs about how each of these three variables, $I_{idt} \equiv \left\{n_{idt}, \bar{s}_{idt}, h_{idt}\right\}$, will evolve if site j is visited at step t.

First, and most simply, given the choice to visit site j, the set of sites visited, h_{idt}, will evolve deterministically to reflect this choice. Second, if site j has not linked to any other sites, then neither n_{idt} nor $\bar{s}_{idt}$ change.[5] However, if site j has linked to another site k (i.e., $\ell_{\vec{jk}d} = 1$), then $n_{ikdt} = n_{ikt\,t-1} + 1$, and the consumer anticipates a new value of $\bar{s}_{ikdt}$ from the following posterior predictive distribution (West and Harrison 1999):

$$(9) \qquad \bar{s}_{ikdt} | \ell_{\vec{jk}d} = 1, n_{ikd\,t-1}, \bar{s}_{ikd\,t-1} \sim N\left(z_k + \frac{\tau_s n_{ikd\,t-1}\left[\bar{s}_{ikd\,t-1} - z_k\right]}{\tau_s n_{ikd\,t-1} + \tau_\nu}, \tau_s^{-1} + \left[\tau_s n_{ikd\,t-1} + \tau_\nu\right]^{-1}\right)$$

Prior to visiting site j, however, the consumer does not know if site j has linked to any other sites. Because links are *a priori* unobserved by consumers, the transition function also reflects consumer i's uncertainty about whether they will see links to other sites. We assume these probabilistic beliefs are rational, to the extent that experienced consumers know the average frequency with which each site j links to every other site k. We denote this long-run average linking frequency $\omega_{\vec{jk}}$. From the perspective of each consumer i, given their knowledge of $\omega_{\vec{jk}}$, encountering a link to site k after arriving at site j is a random event with the following i.i.d. probability:

$$(10) \qquad \Pr{}_i\left[\ell_{\vec{jk}d} = 1 | \omega_{\vec{jk}}\right] = \omega_{\vec{jk}}$$

This link probability does *not* imply that site j links to site k at random. Importantly, *why* site j chose to link to site k on day d does not matter as long as the *consumer's* expectations about links are based on their knowledge of $\omega_{\vec{jk}}$. We elaborate on this point when discussing model identification.

Effects of Linking on Choice

The value function in Equation 8 encapsulates two routes through which linking can affect choice. The first is the *within-session effect* described in Equations 2–7. This effect operates through Bayesian updating of expected horizontal match utility as consumers are exposed to the set of available links, $\ell_{\vec{jk}d}$, over the course of each session. Upon encountering a link, the consumer may become more or less likely to visit the linked

[4] When we extend the model to allow vertical differentiation on the basis of news volume, we update the definitions of U_{ijdt} (Equation 1); the consumer's information state, I_{idt}; and its state transition function, $f\left(I' | I_{idt}, j\right)$. However, the definition of the value function in Equation 8 remains the same.

[5] Previous work has considered how the absence of a link can provide a negative signal about the quality of the (not-)linked site (e.g., Mayzlin and Yoganarasimhan 2012; Dellarocas, Katona, and Rand 2013). Here we assume that the absence of a link is not informative *per se*.

site, depending on the information in the link. This change in the likelihood of a visit, however, only affects choices within the *current session*.

The second route through which linking can affect choice is the *across-session effect*. This effect is a consequence of consumers' forward-looking behavior (Equations 8–10), and in particular, consumers' rational expectations about the sites' long-run average link frequencies, $\omega_{\vec{jk}}$. The consumer knows that encountering links improves the precision of predicted match utilities. Seeing a link thus increases the overall expected value of subsequent browsing. Consequently, sites that typically provide many outbound links may be especially attractive to visit in *any session*, in particular when visited in the early steps. Importantly, this higher valuation due to linking exists in expectation. A site visit therefore does not depend on whether the site has actually made any links that day. If consumers are myopic—that is, if $\delta = 0$—then they do not attend to the benefit of seeing links and *a priori* do not assign extra value to sites that (on average) provide many outbound links. In this model, the across-session effect of linking only makes the linking site more attractive if consumers are forward looking.

To illustrate these implications, we use our model to simulate browsing in a stylized setting with two news sites and one consumer. Site L sometimes links to site R, but R never links to L (thus, $\omega_{\vec{LR}} > 0$, whereas $\omega_{\vec{RL}} = 0$). Because links to R provide unbiased signals of R's horizontal match utility each day, half of L's links signal higher-than-average match with R, and half signal lower-than-average match. To simplify the illustration, we assume that both sites provide the same average match utility to the consumer (i.e., their z_j's are the same), and set the consumer's opportunity cost of browsing high enough that each site has less than 50% chance of being visited each day. The simulations illustrate how linking affects browsing decisions in ways that can be either beneficial or detrimental to the linked site. We report further details and full results in the Web Appendix. Next, we summarize three main insights.

Linking can increase traffic to the linked site through the within-session effect. If the probability of visiting R is below 50% at the start of the session—as is typical for most sites consumers visit—then links from L to R increase the number of L's visitors who subsequently visit R within the same session. Moreover, this increase arises even though half of L's links signal *lower*-than-average match at R. This increase in visits is due to a floor effect on the likelihood of visiting R. If the chance of visiting R is already low, a signal indicating lower than normal match utility can do little to lower the visit likelihood further. By contrast, a signal indicating higher match can raise the chance of visiting the linked site considerably. Importantly, the increase in R's traffic arises in cases when the consumer, if not for the link, would have ended the session after visiting L. Thus, exposure to links increases overall news consumption on average. Note also that the increase in traffic at site R is due to the consumer's exposure to *specific realizations* of links, $\ell_{\vec{LR}d}$, from site L. This increase is an *ex post* effect that arises as a consequence of the consumer of having seen a link. The increase therefore occurs whether or not the consumer is forward looking.

Providing links can increase the linking site's traffic at the start of the session through the across-session effect. Recall that in this example, the z_j's for sites L and R are the same and that L may link to R with probability $\omega_{\vec{LR}} > 0$, but R never links to L. At the start of each session d, before any links have been seen, both sites provide the same expected match utility. But unlike a visit to R, a visit to L might reveal a link to R. If there is in fact a link at L (i.e., $\ell_{\vec{LR}d} = 1$), and if that link signals higher-than-average match utility at R, then the consumer might benefit by visiting R next. Or, if instead the link signals lower-than-average match, then the consumer might also benefit by ending the session without visiting R. Consequently, the possibility of $\ell_{\vec{LR}d} = 1$ causes the expected utility from the entire browsing session to be higher if L is visited before R. A forward-looking consumer anticipates this possibility and thus finds L more attractive at the start of any browsing session.

The increased attractiveness of L has two effects on browsing. First, it makes the option of not browsing relatively less attractive, thus increasing the number of browsing sessions. Second, the increased attractiveness of L means R is relatively less attractive at the start of the session. Thus, by linking to R, site L may end up "stealing" traffic that would have otherwise gone to R (Dellarocas, Katona, and Rand 2013; Jeon and Nasr 2016). Importantly, both of these effects depend on the forward-looking behavior of the consumer, as a consumer who discounts the future completely ($\delta = 0$) does not consider the benefits of site L's links when choosing where to visit.

The combined within- and across-session effects can either increase or decrease traffic at the linked site. The within-session effect increases the number of L's visitors who might subsequently visit R. This positive effect is further amplified by the across-session effect—if L attracts more visitors early in the session, there will be more people seeing its links to R. But linking to R can also lower R's traffic. The across-session effect allows L to attract visitors who, in the absence of linking, would otherwise have visited R. Depending on the size of this effect, R may lose more traffic to L at the start of the session than R gains from L's links later in the session.

Whether the total impact of linking is positive or negative for the linked site is thus an empirical question. The sign depends on a variety of factors, including (1) how often sites link to each other, (2) how informative links are, (3) the extent of horizontal differentiation among the linking sites, (4) the overall popularity of the sites, and (5) the extent to which future benefits from browsing affect previous decisions.

Vertical Differentiation in News Volume

The period utility function in Equation 1 includes a horizontal match utility term, μ_{ijd}. This term varies by site and session, depending on what gets published, and across consumers, depending on what they like to read. As consumers encounter links to other sites, they update their beliefs about their match with the linked site that day. Together, these components provide a rich specification of horizontal site differentiation and consumer heterogeneity.

In most empirical contexts, including ours, sites are also differentiated vertically. In a news setting, vertical differentiation can be based on the volume of basic news facts sites publish, and consumers may differ in the value they place on greater news coverage. Extending the model to include a vertical dimension of utility has two implications for how it can rationalize browsing data. First, differences in news volume can help explain why some sites are more popular among all consumers. Second, in a setting where there is redundancy in news coverage across sites, differences in news volume can also help to explain why we rarely observe consumers visiting more than a few news sites in the same browsing session. In other words, vertical quality also helps to explain session length.

To illustrate the connection between news volume and session length, consider two sites that partially overlap in their coverage of basic news facts—such as the fact that an actor has been admitted to a drug rehab program, or the fact that a singer has released a new music video. After the consumer has visited the first site, some of the news facts at the second site will no longer be *news*, as they will already be known to the consumer. In the extreme, if two sites published every available news fact each day, their coverage would necessarily be identical. In such a case, a reader could obtain all of the day's news by visiting one site or the other, leaving nothing remaining at the second. Publishing a higher volume of news facts thus implies a higher degree of redundancy with other high-volume news sites.

From the perspective of modeling vertical differentiation in news volume, the main implication is that the vertical utility provided by a news site not only depends on how much news the site publishes, but also on (1) which sites were visited previously in the same session, and (2) how much news those sites published. The vertical component of utility is thus state dependent in this setting, as both expected and experienced vertical utility change after each site visit.

Utility from Vertically Differentiated News Sites

To account for vertical differentiation in the volume of news facts sites publish, we update the consumer's utility function (Equation 1) to the following:

$$U_{ijdt} = \mu_{ijd} + \beta_{ijdt} - \gamma_{id} + \epsilon_{ijdt} \tag{11}$$

The term β_{ijdt} represents a vertical component of utility. As a vertical component of utility, all consumers value it in absolute terms, albeit to varying degrees. Because the value of gaining factual information serves as the canonical example of such a vertical dimension, we normalize the utility from no news (i.e., being uninformed about the day's events) to 0 and assume that $\beta_{ijdt} \geq 0$.

We follow Allen (1983; 1986; 1990) by representing news facts as a collection of unique and indivisible *bits* that are observed by consumers, but not by the researcher. These bits correspond with the smallest units of news content that can generate a vertical component of utility (e.g., "Actor X will star in movie Y"). A new set of no more than N bits is available each day. Some of these bits are distributed heterogeneously across sites. A bit might appear at more than one site, or the bit might appear at none of the sites. If some bit b appears at site j on day d, we write $\imath_{bjd} = 1$ (and $\imath_{bjd} = 0$ otherwise). We assume only the first encounter with a bit generates utility, and that thereafter the bit becomes part of the consumer's state of knowledge. The utility from seeing a news bit for the first time is heterogeneous across consumers and is denoted by the parameter $\lambda_i > 0$.[6]

The number of distinct bits that have been seen, prior to choosing what to do at step t, is denoted $K_{id\,t-1}$. We use the notation K_{idt}^{+j} to indicate the number of distinct bits that will have been seen if site j is visited next. Thus, $K_{idt}^{+j} - K_{id\,t-1}$ is the number of remaining (i.e., unseen) bits that will generate utility if j is visited next.[7] We express the vertical utility consumer i gains in terms of this quantity.

$$\beta_{ijdt} = \left(K_{idt}^{+j} - K_{id\,t-1}\right)\lambda_i \tag{12}$$

The consumer knows $K_{id\,t-1}$ before visiting site j at step t. However, due to the nature of news, K_{idt}^{+j}, which indicates a future state of knowledge, is not known ahead of time.

Distribution of Bits and Consumer Learning

The number of bits at each site is obtained from a stylized model of information availability. On each day, there are at most N bits that can be published. The probability that bit b is available at site j is

$$\Pr\left[\imath_{bjd} = 1 | \alpha_j, \pi_b\right] = 1 - \left(1 - \pi_b\right)^{\alpha_j} \tag{13}$$

$$\pi_b \sim U\left(0, 1\right) \tag{14}$$

The parameter $\alpha_j \in (0, 1)$ determines the extent of site j's news coverage, with higher values of α_j indicating more extensive coverage. Bits with higher values of π_b are more likely to be published at all sites. Sites with higher values of α_j are more likely to publish all bits (some may be unique to site j, and others available at many sites). In this way, the α_j's also determine the extent to which sites tend to publish the same bits. Bits are distributed jointly across sites with correlations determined by the α_j's.

Given their past browsing, experienced consumers know sites' long-run, average daily number of bits—that is, they know the sites' α_j's. On any given day, however, consumers are uncertain about *which* bits are in the news ecosystem. Thus, consumers' choices are only affected by their expectations about the *number* of bits they haven't already seen. Accordingly, we augment the consumer's information set, I_{idt}, so it includes $K_{id\,t-1}$, and extend its transition function, $f\left(I' | I_{id\,t-1}, j\right)$, to reflect the consumer's beliefs about likely values

[6] All bits thus generate the same amount of utility for each consumer, and sites are vertically differentiated in the quantity of these bits published each day. Previous versions of this article also considered the case in which different bits produced different amounts of utility, and the average utility from bits varied on different days.

[7] Formally, the relationship between K_{idt}^{+j}, $K_{id\,t-1}$, and the $\imath_{bjd}$'s is the following. Let $\kappa_{idt} \in \{0, 1\}^N$ indicate at step t of day d which bits have already been seen: $K_{id\,t-1} = \sum_{b=1}^{N} \kappa_{bid\,t-1}$ and $K_{idt}^{+j} - K_{id\,t-1} = \sum_{b=1}^{N} \imath_{bjd}\left(1 - \kappa_{bid\,t-1}\right)$.

of $K_{idt}^{+j} - K_{id\,t-1}$.

We assume consumers' prior beliefs about the availability of bits at each site are consistent with Equations 13 and 14). An application of Bayes' rule then leads to the following (binomial) posterior distribution for $K_{idt}^{+j} - K_{id\,t-1}$ (see the Appendix for the derivation).

$$K_{idt}^{+j} - K_{id\,t-1} | K_{id\,t-1}, h_{id\,t-1} \sim B\Big(N - K_{id\,t-1}, \frac{\alpha_j}{1 + A\left(h_{id\,t-1}\right) + \alpha_j}\Big) \tag{15}$$

$$A\left(h\right) = \sum_{k=1}^{J} h_k \alpha_k \tag{16}$$

Recall that the state variable $h_{id\,t-1}$ is a binary vector indicating which sites have already been visited in the session. Thus, $A\left(h\right)$ is the sum of the α_k's for all previously visited sites k. The term $N - K_{id\,t-1}$ represents the maximum number of unseen bits that might yet be seen at one of the remaining news sites that day. The term $\alpha_j \Big/ \left(1 + A\left(h_{id\,t-1}\right) + \alpha_j\right)$ is the consumer's expected probability of finding any of those unseen bits if site j is visited next.

It follows from Equation 15 that the expected vertical utility from the next site j is

$$\mathbb{E}\left(\beta_{ijdt} | K_{id\,t-1}, h_{id\,t-1}\right) = \Big[\Big(\frac{\alpha_j}{1 + A\left(h_{id\,t-1}\right) + \alpha_j}\Big)\left(N - K_{id\,t-1}\right)\Big]\lambda_i \tag{17}$$

The expected level of vertical utility in Equation 17 is the expected number of new bits found at site j from Equation 15, multiplied by the consumer's preference for them, λ_i. Before visiting any sites, $A\left(h_{id0}\right) = 0$ and $K_{id0} = 0$. Thus, $\mathbb{E}\left[\beta_{id1} | K_{id0}, h_{id0}\right] = \left[N\alpha_j / \left(1 + \alpha_j\right)\right]\lambda_i$ at the start of the session. The right side of Table 1 summarizes the within-session, Bayesian updating process for the expected vertical component of utility.

Implications for Browsing

Here we briefly comment on the implications of this part of the model for browsing. Equation 17 reflects how the expected vertical utility is (1) higher at sites that publish more news facts on average, α_j; (2) higher for consumers who receive the most utility from news facts, λ_i; but (3) lower if a large amount of news information has already been obtained, $K_{id\,t-1}$. Moreover, due to the term $A\left(h\right)$ in Equation 17, expected vertical utility is lower if many sites have already been visited—and lower still if the sites that were visited had large values of α_j. The intuition is that any bits that were not already found at a high α site are unlikely to be available from a low α site—whereas the reverse is not true (bits not found at a low α site might yet be available from a high α site). All else equal, a consumer will, on average, prefer to visit sites with higher α's earlier in the browsing session, and sites with lower α's later.

Because each bit can potentially be published by more than one site, the volume of news published each day is correlated across sites. This correlation in news volume affects the consumers' browsing choices at steps $t > 1$ of each browsing session—that is, after the consumer has visited one or more sites and learned something about the day's news coverage. The number of bits at one site thus provides the basis for learning about the number of bits at other sites. We assume that the presence or absence of links is not *directly* informative about the number of bits at the linked site. The presence of links, however, can be *indirectly* informative about bits. This is because links can affect the order of visits, and the ordering of site visits determines which bits are encountered. In this way, links can affect the consumer's beliefs about the bits available at the remaining sites.[8]

This stylized specification of the vertical component of utility achieves two main objectives. First, the specification captures a long-run average component of utility that all consumers value in absolute terms, and thus it cannot be reflected in the model of horizontal match utility. Second, and relatedly, the specification accounts for the impact of daily variation in news volumes and redundant coverage on traffic to news sites.

Data

We estimate the model using data that describe browsing and content at five celebrity news sites between October 1, 2009 and December 31, 2009—a period of 92 days. We assemble these data from two sources: (1) comScore panel data describing consumers' browsing at the URL level and (2) links and content scraped from the sites. We describe both of these data sources before concluding with preliminary evidence that links can either encourage or discourage visits to linked sites—a central feature of this model.

Consumer Data

The browsing data were provided by comScore as part of a larger data set describing visits by a rolling panel of U.S. consumers to more than 3,000 sites (all of which are members of the same blog-oriented advertising network). We focus on celebrity news sites in this study because (1) these sites cover a limited range of news items each day, (2) they frequently excerpt from each other, and (3) they format their home pages like blogs (i.e., as scrolling lists of news stories). We limit our attention to the five most visited celebrity news sites among the panel: *Celebuzz*, *Dlisted*, *Egotastic!*, *Perez Hilton*, and *The Superficial*.[9]

Panelists. Most panelists visit only a fraction of the total available sites, and therefore are largely inconsequential for assessing the impact of links on traffic. We thus limit attention to the most active panelists (Flaxman, Goel, and Rao 2016). These are panelists who (1) visited one or more of the 3,000 sites on at least 16 occasions in Q4 2009, (2) had at least five of those visits occur in each of the three calendar months, and (3)

[8] There are a few reasons not to model two separate, direct effects of links on the horizontal and vertical components of utility. First, because both components are latent and additive in the utility function, it may be difficult to empirically disentangle the marginal impact of a link's presence across the two. Second, we show in the Web Appendix that, in our empirical setting, the presence of one or more inbound links to a site is not meaningfully correlated with the amount of information it publishes. This observation also suggests that identification of the informativeness of links for vertical quality might be difficult to achieve. Third, the signaling effect of links on the vertical component of utility is an interesting question in its own right but tangential to our article's counterfactual goals. In total, the marginal benefit of a more complex model is limited relative to its costs.

[9] We first chose the celebrity news category, then ranked the sites according to the number of unique daily visits from high-frequency readers (those browsing 15 days or more per month to any site in the archive). We then chose the top five sites that provided exclusively celebrity news.

visited at least two of the five sites used for this study. Browsing and demographic data for the 127 consumers who fit this profile make up the estimation panel. In Q4 2009, these 127 consumers comprised 10.8% of browsing sessions involving any of the five sites, and 13.3% of those sites' traffic, even though they represent about 1% of the unique visitors to these sites. The sample thus comprises individuals who are relatively experienced and frequent readers of celebrity news, who would plausibly know (1) the long-run average horizontal position of the five sites, z_j; (2) the typical news volume for each site, α_j; (3) the extent to which horizontal match varies across days, τ_ν; (4) the informativeness of links as match signals, τ_s; and (5) the average frequency with which the five sites link to one another, $\omega_{\vec{L}R}$. Using less restrictive thresholds when defining the panel leads to the inclusion of consumers who do not browse as often, and thus may be less familiar with the sites' average match locations and linking frequencies. In the Web Appendix, we show that our main results are qualitatively insensitive to the cutoffs used to construct the estimation sample.

Most consumers in the estimation panel are female (65%), with the majority (60%) between 25 and 55 years of age (35% are younger, 5% older). Income is reported categorically, with a median of \$55,000–\$65,000 per year. Most panelists have children living with them (57%), and the average household size is 2.7 people. Five panelists listed their race as African American. We code binary variables as $\{-.5, .5\}$, scale the seven income categories between 0 and 1 using the center of the category range, and scale household size by subtracting the median (two people) and dividing by two standard deviations (2.89). We denote by $\boldsymbol{D}_i$ the row vector of demographic variables for consumer i. In the Web Appendix, we contrast the demographics of the estimation sample with a larger set of comScore panelists. Compared with the larger comScore panel, the estimation sample has a higher proportion of consumers who are female, are aged 25–55 years, and have higher incomes.

Browsing data. A consumer's browsing session includes all of their site visits occurring on the same day (as celebrity news sites operate under the same 24-hour news cycle as other media; Leskovec, Backstrom, and Kleinberg 2009). Thus, for each panelist, we compile the order in which any of the five sites were visited each day (the step *t* choices, a_{idt}, in the model). During Q4 2009, the 127 panelists in our estimation sample made 19,130 such choices over the course of 5,757 browsing sessions (where a session might comprise the choice not to browse that day).

Recall that visiting the same site more than once within the same browsing session is not feasible in our model framework. For the median consumer in our sample, 96.9% of sessions generated data consistent with this no-revisit assumption (for a graphical depiction of this distribution, see the Web Appendix). Furthermore, 96.9% might be a lower bound, because internet panel data contain false positives for site/page visits due to web browsers refreshing pages in open tabs (without any action taken by the consumer). Modeling revisits would add significant computational burden in exchange for limited insights. Thus, consistent with the online browsing literature (e.g. Kim, Albuquerque, and Bronnenberg 2010), we do not model revisits. The a_{idt}'s thus reflect the daily rank order of the earliest page request for each of the five sites.

Table 2: Summary of Browsing Behavior by Site and Gender

	Visitors per Day			Step in Session		
Site	Male	Female	All	Male	Female	All
Celebuzz	2.5	7.9	10.3	1.37	1.46	1.44
	(2.2, 2.7)	(7.3, 8.4)	(9.7, 11)	(1.29, 1.46)	(1.41, 1.51)	(1.4, 1.49)
Dlisted	3.4	9.0	12.4	1.42	1.28	1.32
	(3.1, 3.7)	(8.5, 9.6)	(11.8, 13)	(1.35, 1.5)	(1.25, 1.32)	(1.29, 1.35)
Egotastic!	6.8	2.9	9.5	1.23	1.57	1.33
	(6.3, 7.2)	(2.6, 3.2)	(8.8, 10.1)	(1.18, 1.27)	(1.44, 1.7)	(1.27, 1.37)
Perez Hilton	12.3	28.5	40.8	1.19	1.14	1.15
	(11.8, 12.8)	(27.4, 29.5)	(39.6, 41.9)	(1.17, 1.21)	(1.12, 1.16)	(1.14, 1.17)
The Superficial	4.6	3.6	8.0	1.38	1.94	1.62
	(4.3, 4.8)	(3.3, 3.9)	(7.5, 8.5)	(1.32, 1.46)	(1.86, 2.01)	(1.57, 1.67)

Notes: Means and bootstrapped 95% CI's based on 19,130 observed choices over the course of 5,757 browsing sessions. There are 127 consumers in the estimation panel (45 male and 82 female). "Visitors per Day" indicates the average number of male or female panelists visiting each site per day. "Step in Session" indicates the average time index t across visits; thus, lower values indicate visits that occurred earlier in the browsing session.

Panelists differ in the subset of sites they visited most, as well as in the typical order of those visits within the session. Table 2 shows that *Perez Hilton* was the most popular site among both male and female consumers, and was visited earliest in the session on average. Although panelists vary in the order of site visits across sessions, their typical ordering is stable over time (i.e., they are not learning which site is their favorite on average). The audiences of the other four sites differ noticeably by gender: male panelists visited *Egotastic!* and *The Superficial* relatively more, and female panelists visited *Dlisted* and *Celebuzz* relatively more.

Men make up 35% of the panel, but browsed more often than women. The median man browsed on 46 (out of 92) days, averaging 1.12 site visits per session, and the median woman browsed on 44.5 days, averaging 1.05 site visits per session. Variation within each group exceeds these cross-group differences.

Website Data

We created an automated web crawler to collect the full text from all news posts published at each of the five sites in Q4 2009. We use the text scraped from each site to determine, for each day, which other sites the linking site linked to and how many words each site published. We describe each of these next.

Link data. Links that appear within the text of posts are typically accompanied by an excerpt from the linked site, or a brief description of the linked content (Dellarocas, Katona, and Rand 2013). Thus, even though we use the shorter term "link" to refer to both the link and excerpt, the excerpted content, and not the link per se, signals consumers' match with the linked site. We therefore ignore static sidebar links that may be part of a site's navigation, but are never accompanied by an excerpt. After determining which (if any) of the other sites were linked each day (the $\ell_{\vec{L}Rd}$'s), we use the browsing data to infer the number of links to each site consumers

Table 3: Empirical Link Frequencies (%)

	Link Target				
Linking Site	*Celebuzz*	*Dlisted*	*Egotastic!*	*Perez Hilton*	*The Superficial*
Celebuzz	-	6.5	0	1.1	9.8
Dlisted	69.6	-	68.5	2.2	2.2
Egotastic!	0	65.2	-	0	0
Perez Hilton	7.6	0	0	-	0
The Superficial	63	0	0	0	-

Notes: Links were embedded in news articles. We ignore static or sidebar links, as well as links to a site's own content.

would have already seen at each step of the browsing session ($n_{ijd\,t-1}$ in Equation 7).[10]

We derive the $\omega_{\vec{LR}}$'s—the average frequencies with which each site linked to every other site—by averaging over observed links during the 92 days in Q4 2009, and treat them as data during estimation. These frequencies appear in Table 3. As many sites never linked to each other, half of the entries in Table 3 contain zeros (and thus half of the $\omega_{\vec{LR}}$'s are zero). By contrast, *Dlisted* and *Egotastic!* linked to each other about 67% of the time during Q4 2009.

The model assumes consumers know these average link frequencies, $\omega_{\vec{LR}}$, but not on which days those links will appear, $\ell_{\vec{LR}d}$. One implication of this assumption is that the choice to visit a site at the start of a session should not be related to the site's inbound or outbound links that day. To determine whether the data contradict this assumption, we conducted an analysis of the first site consumers visited on days with different numbers of in- and outbound links. Results, which are reported in the Web Appendix, show that browsing sessions are equally likely to start at sites, regardless of how many in- or outbound links they have that day. This result is consistent with the assumption that the decision to browse on day d is independent of the set of links appearing that day, $\ell_{\vec{LR}d}$ (conditional on the consumer's knowledge of the average linking frequencies among sites, $\omega_{\vec{LR}}$). The distinction between consumers' knowledge of average link frequencies, $\omega_{\vec{LR}}$, and their uncertainty about daily link realizations, $\ell_{\vec{LR}d}$, is central to our strategy for identifying link effects on browsing. We elaborate on this point when discussing model identification.

Word counts. Recall that the vertical component of utility is motivated in part by potentially large differences in the amount of news facts the sites publish each day. We use the number of words that sites have published each day as a measure for the unobserved quantity of news facts. As discussed in the model section, a consumer who has just visited a site with a large amount of news facts that are potentially redundant with content at the remaining sites might be more likely to end the session (and vice versa after visiting a site with

[10] Our model assumes that excerpts link to content published on the same day as the excerpt. Linking to older news is possible, but (1) sites have an incentive to appear ahead of their audience with respect to their coverage of the news by linking to fresh content, and (2) we have found exceptions to this to be rare. Accordingly, we make a simplifying assumption that encompasses the majority of cases.

Table 4: Summary of Daily Word Counts by Site

Site	Min	25%	Median	Mean	75%	Max
Celebuzz	0	1,140	1,923	1,912	2,873	4,076
Dlisted	1,746	6,627	11,013	11,072	14,137	33,461
Egotastic!	0	0	463	604	727	2,872
Perez Hilton	0	2,113	4,906	4,482	6,336	9,002
The Superficial	0	280	928	755	1,068	1,769

Notes: Counts include all words in the headline and body text of all posts published on a given day.

very little news information). By using word counts as proxies for sites' unobserved daily quantities of news information, we can measure more precisely the extent of this state dependence due to redundancy. For each site, we calculate the number of words in all posts published that day (including the text of hyperlinks to other sites, if any). Sites' word counts are summarized in Table 4. We transform the daily word counts to define $w_{jd} \propto \log\left(1 + words_{jd}\right)$, and consider w_{jd} to be an indirect measure of the total news volume at site j on day d.

Recall that the vertical component of utility described in the previous section is defined in terms of the amount of news facts (bits) available at each site, but not the number of words. We relate the two as follows. First, after visiting the first site in any session, the consumer will have seen all of the bits published at that site. Thus, the state variable for the number of bits seen after visiting the first site, K_{id1}, is equal to the number of bits that site published on day d. The estimation strategy is thus to functionally equate the values of w_{jd} with daily realizations of the state variables K_{id1}, so $K_{id1} \sim B\left(N, q\left(w_{jd}\right)\right)$. We provide the derivation of $q(\cdot)$ in the Appendix.[11]

Preliminary Analysis

Recall that an excerpt in our model can signal either higher or lower horizontal match utility, thereby increasing *or decreasing* the likelihood of visiting the linked site. To understand whether variation in the data is consistent with the model, we conduct a preliminary analysis at the level of individual consumers. We first define two empirical choice probabilities, for each consumer i, at each site j. The first is the probability that consumer i visits site j after seeing one or more links to j at a previous site:

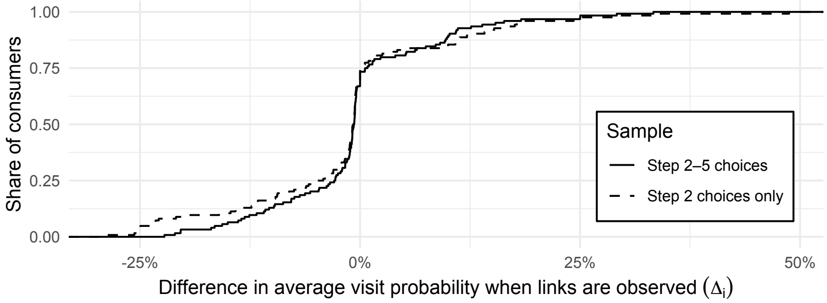
$$\widehat{\Pr}_i\left(a = j | n_{ij} > 0\right) = \frac{\sum_d \sum_t \mathbb{1}\left(a_{idt} = j \text{ and } n_{ijd\,t-1} > 0\right)}{\sum_d \sum_t \mathbb{1}\left(n_{ijd\,t-1} > 0\right)} \tag{18}$$

The second is the probability that consumer i visits j without previously seeing a link to site j:

$$\widehat{\Pr}_i\left(a = j | n_{ij} = 0\right) = \frac{\sum_d \sum_t \mathbb{1}\left(a_{idt} = j \text{ and } n_{ijd\,t-1} = 0\right)}{\sum_d \sum_t \mathbb{1}\left(n_{ijd\,t-1} = 0\right)} \tag{19}$$

[11] Recall that we do not model a direct effect of links on beliefs about the quantity of news information at the linked site. An analysis of links and word counts in the Web Appendix shows that in this setting, there is not a meaningful relationship between the the two.

Figure 2: Average Effect of Exposure to Links on Consumers' Probability of Visiting the Linked Site



Notes: The difference in probability (x-axis) indicates a consumer's frequency-weighted average probability of visiting a site after seeing a link, minus the probability of visiting that same site in the absence of a link, denoted Δ_i in the text.

We next calculate, for each consumer i, the frequency-weighted average of each of these probabilities (i.e., averaging across all five sites). Thus, $\widehat{\Pr}_i\left(a > 0 | n_a > 0\right)$ and $\widehat{\Pr}_i\left(a > 0 | n_a = 0\right)$ denote the probability that consumer i visits *any* site a, given prior exposure to either $n_a > 0$ or $n_a = 0$ links to that particular site. Finally, we calculate the difference between these two probabilities: $\Delta_i = \widehat{\Pr}_i\left(a > 0 | n_a > 0\right) - \widehat{\Pr}_i\left(a > 0 | n_a = 0\right)$. If links tend to encourage consumer i to visit (avoid) linked sites, then we expect $\Delta_i > 0$ ($\Delta_i < 0$); if links have no average effect on browsing, then we expect $\Delta_i \approx 0$.

Because observed links only affect choices at steps $t = 2$ and later (they are seen only after visiting a site), we compute these statistics using a subset of the full sample that excludes choices at step $t = 1$. We also repeat the analysis using only the subset of choices made at step $t = 2$. Figure 2 plots the empirical cumulative distribution of the difference in visit probabilities with and without links, Δ_i, for both subsets. Individuals with negative values of Δ_i are most prevalent. The left tail in Figure 2 corresponds with the majority of consumers who were less likely to visit a linked site after seeing the link. The right tail corresponds with the remaining minority who were more likely to visit a linked site.[12]

This analysis provides preliminary support for our modeling approach, whereby individual links encountered during a browsing session can either increase or decrease the chance of visiting the linked site. The relative prevalence of negative values of Δ_i in the estimation sample shows that links can potentially discourage individuals from visiting the linked site, and that this discouragement might occur to a meaningful extent.

Although this result indicates that for many individuals encountering links might be detrimental to the linked site, recall that the average effect (over all consumers) on site traffic might still be positive, because there is a floor on the probability of visiting the excerpted site. We see evidence for this positive outcome in Figure 2, as the magnitudes of increases in choice probability (the right tail) are greater than the magnitudes

[12] To verify the numerical robustness of this analysis, we repeat it for each subset of consumers who saw a total of at least n links, for $n = 1, \dots, 50$. The share of consumers with $\Delta_i < 0$ ranges between 68% and 83%, and the share with $\Delta_i > 0$ ranges between 17% and 32%.

of decreases (the left tail). To understand if the overall effect is positive, we also calculate frequency-weighted averages of the probabilities in Equations 18 and 19 for each site—meaning we average across choice occasions to derive $\widehat{\Pr}\left(a=j|n_j>0\right)$ and $\widehat{\Pr}\left(a=j|n_j=0\right)$ for each site j. We then again calculate the difference in site-level visit probabilities with and without inbound links. The average effects are positive for four of the sites (ranging, in the $t>1$ subset, from a .3% increase at *The Superficial* to a 3.6% increase at *Egotastic!*), and negative for *Perez Hilton* (-3.7%).

The preliminary analysis exploits variation in consumer choices on days when sites linked or did not link to one other. However, this data is not enough to determine what would be the result of a policy that bans links completely, because the choices recorded in the data occurred in a world in which linking, as a practice, was allowed. Consumers did not know which links would appear at any given site, but they knew the appearance of a link was possible (and might occur with a known average frequency, $\omega_{\vec{jk}}$). Thus, to assess how linking, as a practice, affects browsing, we fit our structural model to the data, and use the estimates to simulate a counterfactual policy of banning links. We next present details relevant to estimation, and then present the model estimates and results of the counterfactual simulations.

Model Specification, Identification, and Estimation

Here we complete the empirical model and describe alternative specifications, model identification, and our MCMC sampling procedure.

Model Specification

Consumer parameters. Consumers are heterogeneous with respect to horizontal match preferences, v_i, their vertical utility from bits of news information, λ_i, and their opportunity costs from browsing, γ_i. We model this heterogeneity using consumers' observed demographic variables, D_i, through the following prior distributions:

$$(20)\qquad v_i \sim N\left(\eta_v + D_i\phi_v, \zeta_v^2\right), \qquad \log\lambda_i \sim N\left(\eta_\lambda + D_i\phi_\lambda, \zeta_\lambda^2\right), \qquad \log\gamma_i \sim N\left(\eta_\gamma + D_i\phi_\gamma, \zeta_\gamma^2\right)$$

Although this prior distribution assumes conditional independence among these parameters, they may be dependent in the joint posterior distribution.

Model likelihood. Following the literature on single agent, dynamic discrete choice models, we assume that the unobserved utility shocks, ϵ_{idjt}, follow an i.i.d. $EV(0,1)$ distribution (Aguirregabiria and Mira 2010). Conditional on the state variables $I_{id\,t-1}$, the value of visiting site j is $V_j\left(I_{id\,t-1}\right)+\epsilon_{idjt}$, where $V_j\left(I_{id\,t-1}\right)$

denotes the choice-specific value function:

$$V_j\left(I_{id\,t-1}\right) = \mathbb{E}\left(\mu_{ijd}|I_{id\,t-1}\right)+\mathbb{E}\left(\beta_{ijdt}|I_{id\,t-1}\right)-\gamma_{id}+\delta\int\log\sum_{k\in\mathcal{J}_{idt}\setminus j}\exp\left[V_k\left(I'\right)\right]f\left(I'|I_{id\,t-1},k\right)dI' \tag{21}$$

The choice-specific value function comprises two parts: (1) the expected period utility from visiting site j at step t, and (2) the expected maximum utility from the remainder of the session, after visiting site j. The latter is an expectation taken with respect to the consumer's information set, $I \equiv \{n, \bar{s}, K, h\}$, which evolves differently depending on which site (if any) is visited after j. Dropping subscripts, the transition function for consumer i's information set is

$$f\left(I'|I,j\right) = p\left(\bar{s}'|n',n,\bar{s},j\right)p\left(n'|n,j\right)p\left(K'|K,h,j\right)p\left(h'|h,j\right) \tag{22}$$

where (1) $p\left(h'|h,j\right)$ denotes the deterministic update of the set of visited sites to include the visit to j; (2) $p\left(K'|K,h,j\right)$ is given by Equation 15; (3) $p\left(n'|n,j\right)$ is such that, per Equation 10, for every site $k \neq j$, $n'_k = n_k + 1$ with probability $\omega_{\vec{jk}}$, and $n'_k = n_k$ otherwise; and $p\left(\bar{s}'|n',n,\bar{s},j\right)$ is given by Equation 9.

Integrating over the unobserved utility shocks, ϵ_{idjt}, leads to the likelihood of the model parameters, θ, conditional on (1) observed browsing choices, a_{idt}; (2) state variables, I_{idt}; (3) average site linking frequencies, $\omega_{\vec{LR}}$; and (4) word counts, w_{jd}:

$$L\left(\theta|a,I,\omega,w\right) \propto \prod_i\prod_d\prod_t^{T_{id}}\prod_{j\in\mathcal{J}_{idt}}\left\{\frac{\exp\left[V_j\left(I_{id\,t-1}|\theta\right)\right]}{1+\sum_{k\in\mathcal{J}_{idt}}\exp\left[V_k\left(I_{id\,t-1}|\theta\right)\right]}\right\}^{1\left(a_{idt}=j\right)} \tag{23}$$

Parameter normalizations. Several parameter normalizations are necessary for estimation. First, recall the term N in Equation 17 represents an upper limit on values of K_{idt} (the number of bits seen). Model fit is insensitive to this value, as the λ_i's can scale up or down at different values of N. We set $N = 30$ during estimation. Second, we set $\tau_\nu = 1$ because the link data can only identify the ratio τ_s/τ_ν. Third, average horizontal match locations, z_j, are latent; thus we normalize them with respect to consumer's horizontal match preferences, υ_i, by setting the mean of the z_j's to be zero. Finally, to avoid a degenerate posterior density for the υ_i's, we set the prior intercept and scale of the υ_i's to $\eta_\upsilon = 0$ and $\zeta_\upsilon = 1$, respectively (Roos and Shachar 2014). The parameters to be estimated are summarized in Table 5.

Bayesian posterior distribution. We assume the following prior distributions for the model parameters:

$$\begin{gathered}\operatorname{logit}\alpha_j \sim N\left(0,1\right), \quad z_j \sim N\left(0,1\right), \quad \tau_s^{-1/2} \sim Ga\left(.4,5\right) \Rightarrow \mathbb{E}\left(\tau_s^{-1/2}\right) = 2, \\ \eta_\lambda \sim N\left(1,.5\right), \quad \eta_\gamma \sim N\left(-1,.5\right), \quad \phi|\zeta \sim N\left(0,\zeta^2\right), \quad \phi_\upsilon \sim N\left(0,1\right), \\ \zeta^2 \sim \text{Sc-Inv-}\chi^2\left(10,.4\right), \quad \gamma_w \sim N\left(0,1\right), \quad \delta \sim U\left(0,1\right)\end{gathered} \tag{24}$$

The likelihood function in Equation 23 depends on the state variables, I_{idt}. The state variables n (number of observed links) and h (sites previously visited) are observed by the researcher, whereas K (number of

Table 5: Summary of Estimated Parameters

Parameter	Dimension	Description
(z_j, α_j)	5×2	Sites' long-run average (horizontal) match locations and (vertical) bit quantities
$(\phi_\upsilon, \phi_\lambda, \phi_\gamma)$	7×3	Demographic coefficients for horizontal match preferences (υ_i), vertical utility (λ_i), and the opportunity cost of browsing (γ_i)
$(\eta_\lambda, \eta_\gamma)$	1×2	Intercepts for vertical utility and the opportunity cost of browsing
$(\zeta_\lambda, \zeta_\gamma)$	1×2	Prior scales for vertical utility and the opportunity cost of browsing
γ_w	1×1	Incremental (multiplicative) opportunity cost of browsing on weekends and holidays
τ_s	1×1	Precision of link signals
δ	1×1	Discount rate for future browsing

Notes: Parameters that are integrated out of the posterior distribution through data augmentation are not listed.

bits seen) and $\bar{s}$ (average signal value of observed links) are not observed. To obtain the marginal likelihood $L(\theta | a, n, h, \omega, w)$, we integrate over the posterior distribution of the unobserved state variables K and $\bar{s}$ using the standard Bayesian approach of data augmentation (Tanner and Wong 1987; Rossi, Allenby, and McCulloch 2005). To improve the efficiency of our sampling procedure, we transform the $\bar{s}$'s. First, we define $s^*_{\vec{L}Rd} \equiv \left(s_{\vec{L}Rd} - z_R - \nu_{Rd}\right) \tau_s^{-1/2}$, so that $s^*_{\vec{L}Rd}$ follows a standard normal distribution independent of z_R and ν_{Rd}. Second, we enforce the identifying restrictions $\mathbb{E}\left(s^*_{\vec{L}Rd}\right) = 0$ and $\mathbb{V}\left(s^*_{\vec{L}Rd}\right) = 1$ through pairwise sampling of the s^*'s, using the method of Musalem, Bradlow, and Raju (2009). We use a parallel strategy to sample the data-augmented ν_{jd}'s.

Alternative specifications. We compare the full model specification to two nested specifications. The first restricts the discount parameter δ to be zero, which means consumers are insensitive to the value of future browsing when choosing which sites to visit next. We refer to this specification as *myopic*. Comparing the myopic and full specifications provides a view into how much consumers' choices depend on their forward-looking beliefs about linking. The second nested model further restricts the informativeness of links, τ_s, to be zero. We refer to this specification as *no signals*. Comparing this specification with the myopic one provides a view into how much consumers' choices are affected by the links they have encountered. Results from these alternative specifications, plus the full model estimated with a larger number of panelists, are in the Web Appendix.

Identification

Model parameters. The model includes multiple parameters defined at the level of individual consumers and sites. These are separately identified due to observing sequential choices within a single browsing session, many sessions over time, and different realizations of state variables for both of these. For many parameters, identification arguments are analogous to those for standard choice models using panel data, in which a person's choices are observed over multiple periods. Consumers' cost parameters, γ_i and γ_w, are like household intercepts in a standard choice model, and are identified from the total amount of browsing (i.e., choices $j = 0$ at step $t = 1$). Consumers' average horizontal match utilities with each site, $z_j \upsilon_i$, are like heterogeneous brand intercepts, and are identified from average choice shares at the start of the browsing session (i.e., choices $j > 0$ at step $t = 1$).

Identification of the remaining structural parameters arises due to differences in choice shares between step $t = 1$ and subsequent steps $t > 1$ of browsing sessions. The covariance of these differences in choice share, when different numbers of links have been encountered, identifies the link informativeness parameter, τ_s (we discuss identification of linking effects more generally below). Similarly, the covariance of the choice share differences, when different numbers of word counts have been observed, identifies the components of vertical utility (sites' α_j's and consumers' λ_i). Conditional on the functional specification for utility, identification of the discount parameter, δ, depends on covariation between choice shares and the average linking frequencies between sites, $\omega_{\vec{LR}}$'s, as well as the exclusion restriction that encountering a link affects choice by altering the consumer's expectations, and not providing utility of its own.

Linking effects. Links do not have a direct effect on utility but, instead, affect browsing through consumers' information sets and expectations. Accordingly, there are several parameters that govern the effect of linking on site traffic. These parameters are defined at the daily and step level (observed links, n, and their average signals, $\bar{s}$), the individual level (horizontal match preferences, υ), the site level (site horizontal locations, z, and average link frequencies, ω), and globally (variation in horizontal location, τ_ν^{-1}, and informativeness of links, τ_s). The within-session effect due to exposure to a link is prototypically reflected in the data when, after seeing the link, a group of consumers with relatively similar horizontal preferences visits the linked site, and a different group—with preferences dissimilar to those of the first group—*do not* visit the linked site.

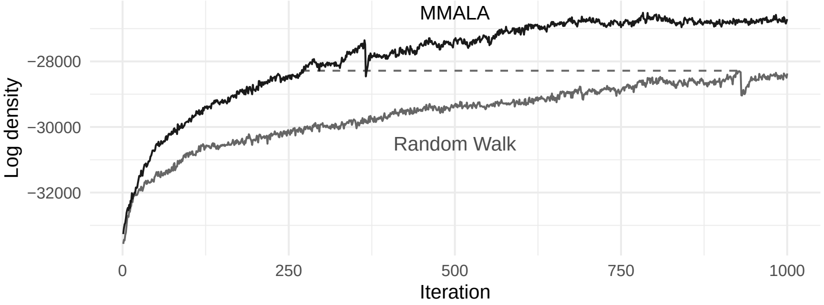
Sites may link to each other for many reasons. However, even if linking is strategic, estimates of the model primitives are statistically consistent. First, consider the choice to visit a potentially linking site L. In this or any other news setting, the consumer does not see site L's links on day d until *after* visiting site L.[13] Thus, even though the consumer knows how often L links to R on average, L's daily outbound links are *a priori* unknown in the current session, and thus exogenous to the choice to visit L.[14]

Next, consider the choice to visit site R after encountering a link to it at site L. Sites obviously do not link to each other at random each day. In particular, site L might choose to link (or not link) to site R on day d for reasons that we do not observe in the data. For example, say that site L only links to R on days when R's content is a better match with L's audience's preferences. Due to selection, the unobserved horizontal match signaled

[13] This is somewhat analogous to how the decision to run ads during the Super Bowl is made prior to the start of the football season (Hartmann and Klapper 2017), thus reducing the possibility of selective exposure to advertising.

[14] Even if a consumer could obtain partial information about L's links prior to a visit, the potential for bias is minimal because other consumers would typically not have access to the same information.

Figure 3: Illustration of IJC's Faster Rate of Convergence with MMALA Proposal Distributions



Notes: Compares the first 1000 draws, during which step size parameters are being tuned to the same target acceptance rates. The maximum log density for the first 1000 draws with random walk proposals is exceeded with less than a third the number of draws with MMALA proposals.

by the link from L to R, $s_{\vec{LR}d}$, would be correlated with the existence or nonexistence of the link, $\ell_{\vec{LR}d} = 1$ or $\ell_{\vec{LR}d} = 0$. If we were to assume independence between the two, we would get inconsistent estimates of link effects. Accordingly, we need to account for this potential correlation in the estimation procedure. More specifically, when integrating over the unobserved match values and link signals, ν_{Rd} and $s_{\vec{LR}d}$, we should account for selected exposure to these unobservables due to (1) the existence of the link on day d, $\ell_{\vec{LR}d}$; and (2) the sequence of choices determining which consumers observe them, a_{itd} (Anand and Shachar 2011). Thus, we follow the standard approach for Bayesian models with correlated unobservables by integrating over the *posterior* distribution of the data-augmented daily match values and link signals, as the posterior distribution of these data-augmented unobservables accounts for both sources of selection.

Estimation

We use the IJC method to sample from the data-augmented posterior distribution of the model parameters. This method is based on the random walk Metropolis-Hastings (MH) algorithm, but augmented with a method for approximating the forward-looking component of the choice-specific value function (Equation 21). The computational gains from IJC are substantial, but may still produce sample chains with high autocorrelation. We alleviate some of this autocorrelation by using Girolami and Calderhead's (2011) MMALA procedure to construct high-quality MH proposal distributions. These proposal distributions have two important features. First, they are centered over points lying in the direction of higher density regions of the parameter space (relative to the current parameter vector). Second, the covariance of the proposal distribution approximates the local curvature of the target distribution.

These features greatly improve the rate of convergence and reduce autocorrelation. Figure 3 illustrates this improved efficiency. The figure plots the first 1,000 draws from the full model estimated with random walk and MMALA proposal distributions. The MMALA chain reaches the maximum log density attained by the random walk chain with less than a third as many iterations. The benefits of MMALA proposal distributions

Table 6: Model Fit Statistics

Model	Parameter Restrictions	MAPE of Site Visits	Expected Deviance	Unrestricted Parameters	Observations	AIC	BIC
Full	-	22.8	$51,205.7$	38	19,130	$51,271.7$	$51,580.3$
Myopic	$\delta = 0$	26.2	$51,277.8$	37	19,130	$51,351.8$	$51,642.5$
No signals	$\delta = 0$, $\tau_s = 0$	32.4	$51,329.4$	36	19,130	$51,401.4$	$51,684.3$

Notes: MAPE = the median absolute percentage error of posterior predicted browsing at each of the five sites; AIC = Akaike information criterion; BIC = Bayesian information criterion.

are not limited to models estimated using IJC. When we estimate the myopic model using MMALA, lag-1, -5, and -50 autocorrelations are 19%, 36%, and 56% lower, respectively, compared with random walk, and effective sample sizes are 13 times higher.

Constructing the MMALA proposal distribution requires the first, second, and third partial derivatives of the target log-density function with respect to the parameters. For single-agent dynamic discrete choice models, these derivatives are not available in closed form. Thus, we obtain their values through automatic differentiation (AD).[15] The MH proposal distributions we construct are based on derivatives of the model posterior distribution while ignoring the IJC approximation to the forward-looking component of the value function. Performing AD on the IJC approximation reduces the numerical stability of the derivative calculations and increases the computational expense. Estimation code is written in MATLAB and C++ using the CppAD library for automatic differentiation (Bell 2007), and is tested using the method of posterior quantiles (Cook, Gelman, and Rubin 2006). The Web Appendix provides additional details about the sampling algorithm.

Results

We first compare the alternative specifications. We then present and discuss parameter estimates from the full model.

Model Fit

We compare model specifications using two measures of fit. First is the median absolute percentage error (MAPE) of the posterior predictive distribution of the total visits to each site across all consumers and days in the sample, which provides a broad measure of fit with the sample data. Second is the expected deviance, which provides a measure of predictive accuracy (Gelman et al. 2004). Table 6 shows the full specification performs best for both measures, and the differences in fit are substantial. Because models with more parameters

[15] Automatic differentiation is a procedure for automatically augmenting computer code so that evaluating an arbitrary function $f(x)$ also yields its derivatives $f'(x)$, $f''(x)$, and so on. The augmented program accomplishes differentiation by algorithmically applying the chain rule corresponding with the primitive operations (addition, multiplication, etc.) comprising the original function (Griewank, Juedes, and Utke 1996; Su and Judd 2012).

Table 7: MAPE of Posterior Predictive Distribution of Total Site Visits

Model	Parameter Restrictions	*Celebuzz*	*Dlisted*	*Egotastic!*	*Perez Hilton*	*The Superficial*
Full	-	24.5	26.1	2.3	5.2	36.8
Myopic	$\delta = 0$	26.4	31.8	2.6	4.1	35.2
No signals	$\delta = 0$, $\tau_s = 0$	35.5	35.0	9.5	2.3	36.1

Table 8: Horizontal (z) and Vertical (α) Quality Parameter Estimates by Site

Parameter	*Celebuzz*	*Dlisted*	*Egotastic!*	*Perez Hilton*	*The Superficial*
z_j	.95	$-.89$	-2.72	2.20	.45
	(.04)	(.07)	(.06)	(.02)	(.07)
α_j	.030	.222	.010	.082	.014
	(.003)	(.007)	(.002)	(.003)	(.002)

Notes: Estimates are posterior means with standard deviations in parentheses.

may have lower expected deviance due to overfitting, Table 6 also presents the Akaike and Bayesian information criteria. These penalize expected deviance by $2p$ and $\ln(O)p$ respectively (for p equal to the number of unrestricted parameters, and O equal to the number of observations).

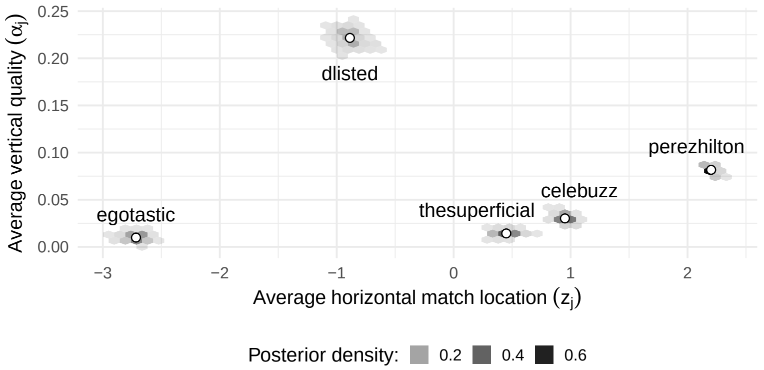
The model comparisons suggest not only that links provide useful information that enables consumers to find better matching content, but also that consumers anticipate these benefits and use them in the way suggested by the full model. First, accounting for exposure to links to other sites in the model helps rationalize consumers' choices within the current session. Second, accounting for the anticipated value of links to other sites helps to rationalize consumers' choices across all sessions.

Table 7 reports the MAPE of total traffic across all consumers and days in the sample for each site. All specifications fit total traffic at *Perez Hilton* and *Egotastic!* better than at the other three sites. The full model performs relatively worse at *Perez Hilton* in terms of prediction error (perhaps because *Perez Hilton* does not provide or receive as many outbound and inbound links as the other four sites). In terms of overall fit, the full model has the lowest prediction error, and we present results from this specification next.

Parameter Estimates

Horizontal match utility. Recall that average horizontal match utility is factored into a site-specific location, z_j, and consumer-specific preference, v_i. Posterior means and standard deviations for sites' average match locations (z_j) are shown in Table 8, and posterior densities are depicted along the x-axis in Figure 4. Drawing on an informal sampling of site content, our post hoc, qualitative interpretation is that sites are horizontally differentiated depending on whether they emphasize content that is more or less sexual (e.g., pictorials of attractive, female entertainers and models). Two sites have negative values of z_j, *Egotastic!* (-2.72) and

Figure 4: Joint Posterior Distribution of Sites' Average Vertical Quality (α_j) and Horizontal Match Location (z_j)



Notes: White points indicate locations of posterior means.

Dlisted (−.89); the other three sites are positive: *The Superficial* (.45), *Celebuzz* (.95) and *Perez Hilton* (2.20). This ordering is consistent with what we see as a relatively greater amount of salacious content at *Egotastic!*, *Dlisted*, and *The Superficial*. Although *Celebuzz* and *Perez Hilton* also publish sexually-oriented content, they do so less frequently and feature attractive male celebrities more than the other three sites. Moreover, reporting at *Celebuzz* and *Perez Hilton* aligns more closely with traditional tabloid celebrity gossip compared with the other three sites.

Consumers' horizontal match preferences (υ_i) are highly heterogeneous, as shown in column 1 of Table 9. This heterogeneity is partly explained by two demographic variables. The most important of these is gender. The match preference coefficient for gender is positive with a 95% Bayesian credible interval (CI) that excludes zero. This implies a higher preference among men (on average) for sites with $z_j < 0$ (i.e. the more salacious content of *Egotastic!* and *Dlisted*), and higher preference among women for sites with $z_j > 0$ (i.e., the more gossipy content of *Celebuzz* and *Perez Hilton*). The other demographic match preference coefficient with a 95% CI excluding zero is African American: these consumers receive higher match utility at *Egotastic!* and *Dlisted*, although we note that this estimate reflects the preferences of just five panelists. In total, demographic variables account for 12.7% of the heterogeneity in horizontal match preferences.

Link informativeness. Horizontal match utility from each site varies each day, and links signal these deviations to consumers. The informativeness of links, relative to daily variation in horizontal match, is reflected in the parameter τ_s in Equation 7. The marginal posterior distribution of τ_s has a 95% CI of (.01, .22) with a mean of .06. The inverse root of this parameter, $\tau_s^{-1/2}$, is the ratio of the standard deviations of signals and daily match deviations. Its posterior mean is 5.2 with a 95% CI of $(2.1, 10.3)$. Figure 5 illustrates the informativeness of links by showing the reduction in uncertainty about a site's match utility after observing increasingly more links. The first link reduces uncertainty about match utility by approximately 6%, and the

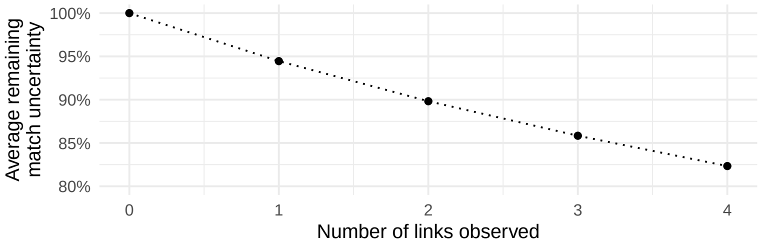
Table 9: Consumer Heterogeneity Parameter Estimates

	Horizontal Match Location (υ)	Vertical Quality Preference ($\log \lambda$)	Opportunity Cost of Browsing ($\log \gamma$)
Observed factors			
Female	.92*	.79*	.21*
	(.18)	(.26)	(.10)
Age<25	.03	−.56*	.03
	(.19)	(.25)	(.10)
Age>55	.08	−1.03*	−.25
	(.32)	(.50)	(.17)
Income	.37	.41	.17
	(.24)	(.37)	(.14)
Children	.10	.02	−.13
	(.24)	(.34)	(.12)
Household Size	.09	−.21	.08
	(.24)	(.33)	(.12)
African American	−1.38*	.86	.09
	(.36)	(.52)	(.20)
Intercept (η)	.00	−1.94*	1.29*
	-	(.36)	(.14)
Unobserved factors			
Prior variance (ζ^2)	1.00	1.30	.51
	-	(.10)	(.04)
Posterior variance	2.21	1.34	.22
Total heterogeneity			
Posterior variance	2.55	1.66	.24
Explained by observed factors	12.7%	16.0%	6.0%

* For observed heterogeneity parameters, indicates that the estimates with 95% CIs exclude zero.

Notes: Estimates are posterior means with standard deviations in parentheses.

Figure 5: Links Reduce Uncertainty about Match Utility



Notes: Match uncertainty remaining (y-axis) is the ratio of the posterior and prior variance of match utility after observing $n = 0, \ldots, 4$ links (x-axis), $\text{var}(\mu | n = 0, \ldots, 4) \big/ \text{var}(\mu | n = 0)$, estimated as the posterior mean of $\left(\tau_s n + 1\right)^{-1}$.

second link by another 4%.

Overall, we find compelling evidence that links provide informative signals about content at other sites, inasmuch as after seeing a link, consumers are less uncertain about match utility at the linked site. The counterfactual simulations further demonstrate that the information provided by links can have a meaningful impact on browsing.

Vertical utility from news volume. Sites are differentiated according to the volume of news published on average. Posterior means and standard deviations for sites' vertical qualities (α_j) are also shown in Table 8, and posterior densities are depicted along the y-axis in Figure 4. *Dlisted* is estimated to provide the highest average level of vertical utility; *Egotastic!* and *The Superficial* the lowest. These estimates reflect consumers' browsing habits, as well as differences in the average number of words published each day.

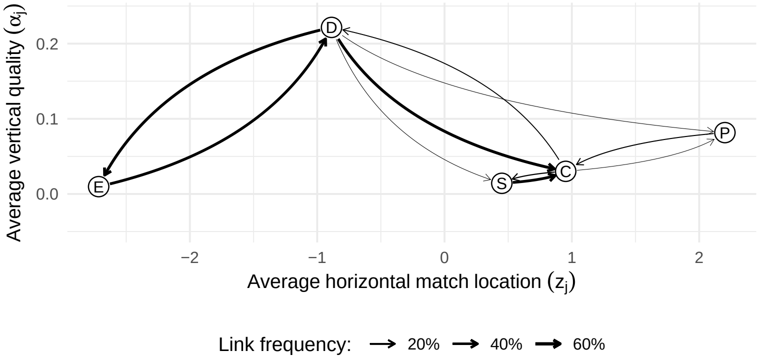
Consumers are heterogeneous in their preference for this vertical component of utility, λ_i, column 2 of Table 9 shows. Demographic variables explain 16% of this heterogeneity, with female consumers and those aged 25–55 years receiving the greatest amount of this vertical component of utility.

Opportunity costs of browsing . The opportunity cost of browsing, γ_i, also varies by gender. Column 3 of Table 9 shows that female consumers have higher costs than male consumers. Together, demographic variables explain 6% of the variation in $\log \gamma_i$. As expected, consumers with the highest opportunity costs tended to visit the fewest number of sites. Furthermore, they were more likely to choose *Egotastic!* and *Perez Hilton* (both sources of high match utility) at the start of their sessions. The estimate for γ_w indicates browsing costs are approximately 7.2% (SD = .02%) higher on weekends. Because the opportunity cost of browsing is measured relative to the value of the outside option, this result is consistent with an outside option that is more valuable on weekends (Ahn, Duan, and Mela 2015).

Discount rate. The parameter δ determines the rate at which future browsing is discounted. This parameter is estimated in the full model, and has a posterior mean (median) of .256 (.253) and a 95% CI of (.001, .645). Although the posterior distribution includes values very close to zero, model fit is significantly improved when this parameter is estimated (rather than set to zero).

The discount rate is high when compared to those estimated from purchase data. In models of consumer purchases, utilities are measured relative to money costs, which permits a monetary interpretation of the discount rate. Here, utilities are measured relative to the value of an outside option that corresponds with "not browsing." This lack of a dollar metric limits our ability to interpret the magnitude of the discount rate. However, the fact that this parameter takes on a nonzero value suggests that the value of future browsing has an impact on the choice of which site to visit.

Figure 6: Summary of Link Frequencies and Site Heterogeneity



Notes: Link frequency indicates the empirical distribution of links as observed in the data (ω). Sites are located at their posterior means for z_j and α_j. C = *Celebuzz*; D = *Dlisted*; E = *Egotastic!*; P = *Perez Hilton*; S = *The Superficial*.

Average Linking Frequency and Site Differentiation

Next, we comment briefly on how the frequency of links between competing sites affects how sites are differentiated in the eyes of consumers. Sites are differentiated by their average horizontal match locations, z_j, average news volumes, α_j, and linking frequencies, $\omega_{\vec{jk}}$. Figure 6 depicts these characteristics spatially. Sites are indicated as points according to their horizontal match location along the x-axis and vertical quality along the y-axis. Link frequencies are overlaid as arcs of varying widths.

Figure 6 shows that sites tend to link to competitors with similar values of z_j (i.e., their closest neighbors along the x-axis).[16] Because links provide signals about daily match locations, sites that frequently link to their closest competitors provide value by informing their audiences about sites with similar levels of match utility. If instead, links tended to point to sites with very different match locations—if *Egotastic!* were to link to *Perez Hilton,* for example—then consumers would find links to be less useful, because the links would be telling consumers about sites they are highly unlikely to visit anyway. As we demonstrate next through counterfactual simulation, a meaningful portion of some sites' value to consumers stems from their tendency to link to other sites.

Counterfactual Analysis

How much do the within-session, across-session, and combined effects of linking affect consumer demand for online news? In this section, we use estimates from the structural model to answer that question in the empirical context we study—consumption of celebrity news in Q4 2009. We use estimates from the structural model because they enable us to compare browsing not only in the presence or absence of links (as in the preliminary analysis) but also in the presence or absence of consumers' expectations about links. This structural approach

[16] In the Web Appendix, we show that the estimates of z_j and α_j are similar when estimating models with and without the link data.

makes it possible to consider policies not reflected in the data, such as outright or *de facto* bans on linking to news sites (similar to ones previously enacted within the EU). We first describe the approach and then discuss the main insights.

Procedure

We measure the impact of linking in terms of the amount of browsing, the flow of traffic between sites, and total traffic at each site, by comparing demand simulated under two scenarios. In the baseline scenario, we simulate demand using all links that are observed in the data (see Table 3 and Figure 6). In the counterfactual scenario, we remove these links and update consumers expectations about linking frequencies. In other words, we set all of the $\ell_{\vec{j}kd}$'s and $\omega_{\vec{j}k}$'s to zero before simulating demand. This counterfactual scenario assumes there has been an external intervention prohibiting linking (e.g., an extreme version of a "link tax"), and that sites continue to produce the same type of content as before. Banning links might also induce sites to change their typical content, but because we do not model these decisions, we cannot consider such an outcome. These results should therefore be interpreted as conditional effects in light of the existing content strategies.

We simulate the full 92-day sequence of browsing S times for every consumer under the baseline and counterfactual scenarios. Each of the S simulations corresponds with a sample from the data-augmented posterior distribution of the model parameters. For each simulation, we calculate a quantity of interest (e.g., the change in a site's traffic particular site), then take the average over all S simulations (i.e., we integrate over the posterior distribution). Because consumers' expectations about the links they will encounter depend on the $\omega_{\vec{j}k}$'s, we re-estimate the value function for each of the S parameter draws. This re-estimation is computationally expensive, and thus we set $S = 500$. To account for simulation error, we calculate bootstrap confidence intervals for all estimates and focus attention on measured effects that are reliably different from zero. To facilitate intuition, we frame the results as changes from the counterfactual with no linking to the baseline with linking. Thus, when speaking of a quantity y as the expected percentage change from links, we mean $\mathbb{E}_\theta\left[\left(y^{baseline} - y^{counter}\right)/y^{counter}\right]$.

We present the results in two stages. First, we discuss the total effect of links on consumers and sites at the aggregate level. Second, we decompose this total effect into two theoretically distinct effects of linking on choice: (1) the within-session effect due to observing a particular link on a given day (as a result of the $\ell_{\vec{j}kd}$'s), and (2) the across-session effect due to the anticipation of outbound links (as a result of the $\omega_{\vec{j}k}$'s).

Total Effect of Links

Among these sites, links positively affect browsing, as shown in Table 10. When we compare the baseline with linking with a counterfactual without, the total number of browsing sessions increases .11%. For the median consumer, the number of browsing sessions increases by .59%, the number of sites visited per session by .14%, and the total number of site visits increases by .54%. The impact of linking on the median consumer

Table 10: Total Effects of Linking on Consumers and Sites

	Relative Change (%)	
	Median Consumer	All Consumers
Number of browsing sessions	.59*	.11*
Sites visited per browsing session	.14*	−.05
Total sites visited	.54*	.06
Visits to:		
Celebuzz		.10
Dlisted		.18
Egotastic!		.14
Perez Hilton		.09
The Superficial		.01

* Indicates that the 95% bootstrap CI around the estimate excludes 0.

Notes: Percentage changes are expressed relative to the counterfactual with no linking.

is more positive than the average effect across all consumers. This is because the increases in browsing and site visits due to linking are greatest among consumers who browse relatively less. Put another way, links provide less of an incentive to browse for those who would browse anyway, and more of an incentive for the marginal consumer. Turning to the site-specific browsing results, Table 10 shows that links also increase sites' traffic to varying degrees. The greatest gains in total visits are found at *Dlisted* (.18%) and *Egotastic!* (.14%), sites that give and receive relatively greater numbers of links.

The total effects reported in Table 10 reflect both the within- and across-session effects and are averaged across conditions in which links affect browsing decisions to different degrees. At the start of a browsing session, for instance, only forward-looking consumers' *expectations* about links influence choice. Later in the session, choices are also affected by the actual links they might encounter. Because sites do not always provide outbound links, and because consumers do not visit every site, we interpret the results in Table 10 as the total effect of a broader policy of allowing links, with the understanding that some consumers may see few or even none of those links. To understand how exposure to any *particular* link affects consumers' choices, we decompose the total effect into its constituent parts.

Decomposition of the Total Effect of Links

The across-session effect due to changes in beliefs about linking frequencies. Linking has a different effect on decisions at step $t = 1$ of a session, compared with later steps. At step $t = 1$ (prior to observing any links), only the across-session effect (due to $\omega_{\vec{jk}}$) matters. Choices are affected by forward-looking consumers' anticipation of links they might encounter but unaffected by specific realizations of links between sites (none have been encountered yet). The columns labeled "all consumers" in Table 11 show how linking changes aggregate site traffic differently at the first step of consumers' browsing sessions (when only expectations of links contribute to choice), compared with later steps (when both expectations and prior exposure to links

Table 11: Relative Change in Total Visitors by Step of Browsing Session (%)

	Arriving at Step $t = 1$			Arriving at Steps $t > 1$		
	Core Audience	Non-Core Audience	All Consumers	Core Audience	Non-Core Audience	All Consumers
Celebuzz	.42	.72	.21	.73	.68	.45
Dlisted	.13	1.18*	.21	−.20	2.76*	.42
Egotastic!	.07	1.66	.03	1.31*	5.50*	1.32*
Perez Hilton	.22*	.23	.21*	.01	−.47	−.55
The Superficial	.89	1.83*	.76	−.05	.82	−.14
All sites			.11*			−.07

* Indicates that the 95% bootstrap CI around the estimate excludes 0.

Notes: Percentage changes are expressed relative to the counterfactual with no linking. Consumers are included in either the core audience or noncore audience segments based on how many times they visited each site in the raw data. The top 30 consumers comprise the core audience and the remainder the noncore audience.

matter). The columns labeled "core audience" and "noncore audience" provide insight into the heterogeneity of linking effects, as the statistics in those columns are calculated using a different subset of consumers for each site. Specifically, consumers are defined to be part of a site's core audience if they are among the top 30 most frequent visitors to that site in the raw data. Otherwise, they are noncore.

The difference between the within-session and across-session effects for *Egotastic!* provide a useful illustration. *Egotastic!* gains little traffic at step $t = 1$ (.03% in total) due to the cross-session effect. This suggests that *Egotastic!'s* outbound links do not make it more attractive. But *Egotastic!* does gain substantially more traffic at later steps $t > 1$ (1.32% in total) due mostly to the within-session effect. This result shows how the across-session effect depends on both sites' horizontal positions and the other sites to which they link. Specifically, *Egotastic!* creates and receives a large number of links, but only in exchange with *Dlisted*. *Dlisted*, in contrast, links to all other sites. Because a substantial portion of *Egotastic!*'s audience frequently visits *Dlisted* as well (even in the absence of links), the information value of *Egotastic!*'s links is lower than *Dlisted*'s in expectation. Accordingly, when linking is allowed, *Egotastic!* actually loses a portion of its audience to *Dlisted* at step $t = 1$ (and *Dlisted* gains 1.18% in traffic from its noncore audience). The loss in some of *Egotastic!'s* traffic to *Dlisted* at step $t = 1$ thus offsets any gains that might have accrued to *Egotastic!* from its own outbound links. Accordingly, the increase in traffic at step $t = 1$ (due to the across-session effect) is greater for *Dlisted* (.21%) than for *Egotastic!* (.03%). At the same time, the within-session effect for *Egotastic!* is substantial. When linking is allowed, the number of visitors to *Egotastic!* at later browsing steps is 1.32% higher. Moreover, among *Egotastic!'s* noncore audience, the gain in total traffic is 5.5%.

The within-session effect due to exposure to a link. Although the increase in traffic at later stages is relatively large for *Egotastic!*, this increase is defined as an average over cases in which some consumers encounter a link and others do not. We are also interested in comparing consumers' choices when they have been exposed to a link against counterfactual choices in which the link has been removed. The challenge with

Table 12: Difference in Choice Probability by Prior Exposure to Link (%)

	After Exposure to Actual or Removed Links			With No Prior Exposure to Links		
	Core Audience	Non-Core Audience	All Consumers	Core Audience	Non-Core Audience	All Consumers
Celebuzz	.27	−.04	.04	−.01	.01	.01
Dlisted	.02	.42*	.26*	−.04	.00	−.02
Egotastic!	.27*	−.01	.08*	−.02	.01*	.00
Perez Hilton	.72	.25	.47	−.17	−.11*	−.14*
The Superficial	.26	.21	.27*	−.07	−.07	−.03
All sites			.14*			−.02*

* Indicates that the 95% bootstrap CI around the estimate excludes 0.

Notes: Differences are expressed relative to the counterfactual with no linking. Consumers are included in either the core audience and noncore audience segments based on how many times they visited each site in the raw data. The top 30 consumers comprise the core audience and the remainder the noncore audience.

making this comparison is that, as Figure 6 shows, sites often link to their closest neighbors in terms of match location. Thus, the propensity to visit a linked site, conditional on having already visited the linking site, is *a priori* high.

To deal with this challenge, we introduce the concept of a *removed link*, meaning a link that a consumer would have seen, had we not removed that link under the counterfactual of no linking.[17] We compare baseline choices, in which a link was observed, against counterfactual choices, in which a removed link would have been observed had the link not been deleted. Thus, the measured effect is almost entirely due to the exogenous presence or absence of the link itself. These differences in consumers' propensities to visit a linked site are somewhat analogous to click-through rates for (untargeted) internet ads, in the sense that their measurement is predicated on exposure to a particular link (or ad).

Table 12 summarizes the results of this analysis. The third column of Table 12 compares consumers' baseline choices after exposure to links with their counterfactual choices after "exposure" to removed links. For example, the probability of visiting *Dlisted* after exposure to a link is .26% higher than the visitation probability would be without the link. This result represents a 3.8% increase in the amount of *Dlisted*'s traffic originating from linking sites. The changes in visit probabilities differ in magnitude among sites' core and noncore audiences. Links to *Dlisted* from other sites have a greater impact on visits to *Dlisted* among its noncore audience than among its core audience. Links to *Dlisted* thus increase its traffic relatively more on the extensive margin. By contrast, links to *Egotastic!* (all of which come from *Dlisted*) have a greater impact on visits to *Egotastic!* among its core audience than among its noncore audience. Links to *Egotastic!* thus increase its traffic relatively more on the intensive margin.

The difference in how links affect traffic among *Dlisted* and *Egotastic!'s* core and noncore audiences is

[17] Specifically, if site L linked to site R on day d, any consumer visiting site L on day d under the counterfactual with no linking is said to be exposed to a removed link. That is, these consumers would have seen the link to site R ($\ell_{\vec{L}Rd} = 1$), if not for us removing it under the counterfactual ($\ell_{\vec{L}Rd} = 0$).

related to the order in which these sites are typically visited when linking is allowed or banned. When linking is allowed, a substantial portion of *Egotastic!'s* core audience is made up of individuals who prefer to visit *Dlisted* first because of its outbound links. This result thus demonstrates another way in which the across-session effect moderates the within-session effect—it determines in part which consumers are (or are not) exposed to links.

The overall frequency-weighted average increase in the probability of visiting a linked site due to prior exposure to a link is .14%, a 2.3% increase. A relevant baseline for comparison is paid forms of links, such as display advertising. These typically have click-through rates less than .05% (Lambrecht and Tucker 2013; Lewis, Rao, and Reiley 2011; Chaffey 2017), and the effect we measure is large by comparison.[18]

Contribution and Opportunities for Further Study

Linking between news sites is a distinguishing feature of internet news, and one with the potential to change the way individuals stay informed. By providing information to consumers of online news, links make it easier to seek out more interesting content, and to avoid less interesting content. When consumers value news links, sites that provide them become more attractive to consumers, even to the point that the linking site might be more popular than the sites it links to (e.g., *Google News*). The potential for a linking site to benefit more than the sites it links to has been the catalyst for both lawsuits (both the Associated Press and Agence France-Presse news services previously sued *Google News*) and regulatory actions (legislation in Germany, Spain, and the EU have narrowed the legal basis for linking to news sites). The reasoning behind these legal actions is often predicated on a presumption of harm to the news sites that receive inbound links. Yet, to date, there has been little academic work seeking to understand the impact of links on demand for news sites, and what work has been done has mostly been limited to the context of *Google News*.

We offer a new perspective on this issue by studying the impact of linking among news publishers—that is, sites that both publish original news, and link to other news sites. We show that it is possible to measure the effects of links on consumers and news sites, instead of presuming they are harmful. In the empirical setting we study, celebrity news, we find that links change the way experienced consumers browse for news to the benefit of both consumers and news sites. Compared with a counterfactual policy in which links are banned, consumers browse more under the baseline scenario with linking. In the sample we consider, the median consumer is both more likely to start browsing, and more likely to continue browsing, when linking is allowed. These differences are greatest among consumers who browse for news relatively less, which suggests that linking plays an important role in increasing news consumption at the extensive margin. Individuals who encounter links have, on average, a .14% higher probability of visiting the linked site, a 2.3% increase over the

[18] Table 12 also reports changes in traffic when no links were encountered in the baseline scenario with linking. These differences in choice probabilities are entirely due to the across-session effect, but only considering choices at steps $t > 1$.

counterfactual without links. The size of this effect is approximately three times larger than typically reported increases from display ads.

We also make several methodological contributions. First, we present a model in which links signal consumers' daily horizontal match with the linked site's news content, thus allowing the within-session effect of encountering a link to either increase or decrease the chance of a visit. We show that even when links signal lower than average match, the aggregate effect on the linked site can be positive. Second, we also develop a novel Bayesian learning model based on bits of news information that are redundantly distributed across multiple news sites. We developed this model in the context of studying internet news consumption, but the approach can be applied to study the consumption of offline news, or more generally, sequential choices in other settings where alternatives have correlated utilities due to redundancy. Third, we demonstrate the value of combining adaptive MMALA proposal distributions with IJC's method for sampling from dynamic discrete choice models. Compared with the standard IJC method using random walk proposal distributions, the gains from our approach are significant: autocorrelations using our approach are substantially lower, and effective sample sizes are many times larger.

This study also has limitations that bear on how we interpret the results, but that also point in potentially useful directions for future studies. The first of these is related to our focus on steady-state demand among experienced consumers of celebrity news (as opposed to the more transient behaviors of new consumers). That is, we study within-session learning about daily variation in content among experienced consumers who encounter links to sites they are already familiar with. Owing to the complexity involved in understanding this within-session learning process, this study does not consider the process by which inexperienced consumers eventually become experienced over the course of many sessions. Yet understanding this process of site discovery is important, and links certainly play a role in how this process unfolds. The total value of linking depends not only on the value a site's outbound links provide for its readers but also on whether the site's links inform readers about the availability of new, and potentially superior alternatives. Previous work has considered how sites may choose their content and links strategically to attract and sustain interest among both types of consumers (e.g., Mayzlin and Yoganarasimhan 2012). Empirically studying how sites balance these two forces will be a critical step forward in our understanding of how linking affects news consumption in a competitive setting.

The inclusion of both horizontal and vertical dimensions of utility in our model allows it to flexibly represent how interactions between consumer preferences and site content determine consumers' choices. At the same time, the available data and the need to model forward-looking consumers constrain the degree to which these aspects can be further developed. We use word counts as proxies for the amount of news facts sites publish each day, and model the horizontal component of utility in a one-dimensional latent space. These word count and dimensionality choices allow for model estimation but limit our ability to extrapolate from the parameter estimates to other empirical settings (e.g., making out-of-sample predictions for traffic at other sites). By applying recent advances in image and text analysis, we might overcome these limitations, and further gain a clearer understanding of how news sites differentiate from one another. Such an approach might enable us to measure the signals embedded in individual links more precisely. Better measurement of link content would lead to richer models in which links could influence learning on both the horizontal and vertical dimensions (e.g., by allowing the absence of a link to signal lower vertical quality at the not-linked site). Previous studies have considered links as signals of sites' short- and long-run vertical qualities (Mayzlin and Yoganarasimhan 2012; Dellarocas, Katona, and Rand 2013). There is room for a unifying theory of linking in both the horizontal and vertical dimensions and for further empirical work in this area.

A final limitation pertains to the generalizability of the empirical results. The results we obtain apply to the limited context of celebrity news at a particular point in time, and are predicated on the behavior of more frequent (and knowledgeable) readers. These results are valid for this setting, but this setting may not be typical of most news consumption. Our modeling framework suggests that each news ecosystem might have a different answer to the question of how linking affects consumers and news sites. Accumulating more evidence about how links affect news consumption might lead to generalizations about the conditions under which linking is beneficial or harmful. This study is a step in this direction, and one that we hope stimulates further work on the important topic of news consumption.

References

Aguirregabiria, Victor and Pedro Mira (2010), "Dynamic discrete choice structural models: A survey," *Journal of Econometrics,* 156 (1), 38–67.

Ahn, Dae-Yong, Jason A. Duan, and Carl F. Mela (2015), "Managing user-generated content: A dynamic rational expectations equilibrium approach," *Marketing Science,* 35 (2), 284–303.

Allen, Beth (1983), "Neighboring information and distributions of agents' characteristics under uncertainty," *Journal of Mathematical Economics,* 12 (1), 63–101.

Allen, Beth (1986), "The demand for (differentiated) information," *The Review of Economic Studies,* 53 (3), 311.

Allen, Beth (1990), "Information as an economic commodity," *The American Economic Review,* 80 (2), 268–273.

Anand, Bharat N. and Ron Shachar (2011), "Advertising, the matchmaker," *The RAND Journal of Economics,* 42 (2), 205–245.

Athey, Susan, Markus Mobius, and Jenő Pál (2017), "The Impact of Aggregators on Internet News Consumption," https://ssrn.com/abstract=2897960.

Bell, Bradley M. (2007), "CppAD: A Package for Differentiation of C++ Algorithms," https://coin-or.github.io/CppAD.

Calzada, Joan and Ricard Gil (Jan. 10, 2019), “What Do News Aggregators Do? Evidence from Google News in Spain and Germany,” https://ssrn.com/abstract=2837553.

Chaffey, Dave (Mar. 8, 2017), “Display advertising clickthrough rates,” https://www.smartinsights.com/internet-advertising/internet-advertising-analytics/display-advertising-clickthrough-rates/ (captured by archive.org on 11/15/2017).

Ching, Andrew T., Tülin Erdem, and Michael P. Keane (2013), “Learning models: An assessment of progress, challenges, and new developments,” *Marketing Science,* 32 (6), 913–938.

Ching, Andrew T., Tülin Erdem, and Michael P. Keane (2017), “Empirical models of learning dynamics: A survey of recent developments,” *Handbook of Marketing Decision Models*. Cham: Springer, 223–257.

Ching, Andrew T., Susumu Imai, Masakazu Ishihara, and Neelam Jain (2012), “A practitioner’s guide to Bayesian estimation of discrete choice dynamic programming models,” *Quantitative Marketing and Economics,* 10 (2), 151–196.

Chiou, Lesley and Catherine Tucker (2017), “Content aggregation by platforms: The case of the news media,” *Journal of Economics & Management Strategy,* 26 (4), 782–805.

Cook, Samantha R., Andrew Gelman, and Donald B. Rubin (2006), “Validation of software for Bayesian models using posterior quantiles,” *Journal of Computational and Graphical Statistics,* 15 (3), 675–692.

Danaher, Peter J. (2007), “Modeling page views across multiple websites with an application to Internet reach and frequency prediction,” *Marketing Science,* 26 (3), 422–437.

Dellarocas, Chrysanthos, Zsolt Katona, and William Rand (2013), “Media, aggregators, and the link economy: Strategic hyperlink formation in content networks,” *Management Science,* 59 (10), 2360–2379.

Erdem, Tülin and Michael P. Keane (1996), “Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets,” *Marketing Science,* 15 (1), 1–20.

Flaxman, Seth, Sharad Goel, and Justin M. Rao (2016), “Filter bubbles, echo chambers, and online news consumption,” *Public Opinion Quarterly,* 80 (S1), 298–320.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (2004), *Bayesian Data Analysis,* 2nd ed. Boca Raton: Chapman & Hall/CRC.

Gentzkow, Matthew and Jesse M. Shapiro (2008), “Competition and truth in the market for news,” *The Journal of Economic Perspectives,* 22 (2), 133–154.

Gentzkow, Matthew and Jesse M. Shapiro (2011), “Ideological segregation online and offline,” *The Quarterly Journal of Economics,* 126 (4), 1799–1839.

Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson (2011), “The effect of newspaper entry and exit on electoral politics,” *The American Economic Review,* 101 (7), 2980–3018.

George, Lisa M. and Christiaan Hogendorn (2019), “Local news online: Aggregators, geo-targeting and the market for local news,” https://ssrn.com/abstract=2357586.

Gingras, Richard (Sept. 25, 2019), “How Google invests in news,” https://www.blog.google/perspectives/richard-gingras/how-google-invests-news/ (captured by archive.org on 12/15/2019).

Girolami, Mark and Ben Calderhead (2011), “Riemann manifold Langevin and Hamiltonian Monte Carlo methods,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 73 (2), 123–214.

Goldfarb, Avi (2002), “Analyzing website choice using clickstream data,” *Advances in Applied Microeconomics,* 11 209–230.

Griewank, Andreas, David Juedes, and Jean Utke (1996), “Algorithm 755: ADOL-C: A Package for the automatic differentiation of algorithms written in C/C++,” *ACM Transactions on Mathematical Software,* 22 (2), 131–167.

Hartmann, Wesley R. and Daniel Klapper (2017), “Super Bowl ads,” *Marketing Science,* 37 (1), 78–96.

Imai, Susumu, Neelam Jain, and Andrew T. Ching (2009), “Bayesian estimation of dynamic discrete choice models,” *Econometrica,* 77 (6), 1865–1899.

Jeon, Doh-Shin and Nikrooz Nasr (2016), “News aggregators and competition among newspapers on the Internet,” *American Economic Journal: Microeconomics,* 8 (4), 91–114.

Johnson, Eric J., Wendy W. Moe, Peter S. Fader, Steven Bellman, and Gerald L. Lohse (2004), “On the depth and dynamics of online search behavior,” *Management Science,* 50 (3), 299–308.

Karlštrems, Alvils (Aug. 26, 2019), “Google AdSense CPM Rates 2019,” https://www.bannertag.com/google-adsense-cpm-rates/ (captured by archive.org on 01/29/2020).

Katona, Zsolt and Miklos Sarvary (2008), “Network formation and the structure of the commercial World Wide Web,” *Marketing Science,* 27 (5), 764–778.

Kim, Jun B., Paulo Albuquerque, and Bart J. Bronnenberg (2010), “Online demand under limited consumer search,” *Marketing Science,* 29 (6), 1001–1023.

Lambrecht, Anja and Catherine Tucker (2013), “When does retargeting work? Information specificity in online advertising,” *Journal of Marketing Research,* 50 (5), 561–576.

Lee, Sukekyu, Fred Zufryden, and Xavier Drèze (2003), “A study of consumer switching behavior across Internet portal web sites,” *International Journal of Electronic Commerce,* 7 (3), 39–63.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg (2009), “Meme-tracking and the dynamics of the news cycle,” *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. New York, 497–506.

Lewis, Randall A., Justin M. Rao, and David H. Reiley (2011), “Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising,” *Proceedings of the 20th International Conference on World Wide Web*. ACM. New York, 157–166.

Majó-Vázquez, Sílvia, Ana S. Cardenal, and Sandra González-Bailón (2017), "Digital news consumption and copyright intervention: Evidence from Spain before and after the 2015 'Link Tax'," *Journal of Computer-Mediated Communication,* 22 (5), 284–301.

Mayzlin, Dina and Hema Yoganarasimhan (2012), "Link to success: How blogs build an audience by promoting rivals," *Management Science,* 58 (9), 1651–1668.

Musalem, Andrés, Eric T. Bradlow, and Jagmohan S. Raju (2009), "Bayesian estimation of random-coefficients choice models using aggregate data," *Journal of Applied Econometrics,* 24 (3), 490–516.

Park, Young-Hoon and Peter S. Fader (2004), "Modeling browsing behavior at multiple websites," *Marketing Science,* 280–303.

Posada de la Concha, Pedro, Alberto Gutiérrez García, and Hugo Hernández Cobos (2015), "Impacto del nuevo Artículo 32.2 de la Ley de Propiedad Intelectual," tech. rep. NERA Economic Consulting. Summarized in http://www.aeepp.com/noticia/2272/actividades/informe-economico-del-impacto-del-nuevo-articulo-32.2-de-la-lpi-nera-para-la-aeepp.html (captured by archive.org on 8/14/2015) and translated by Google Translate.

Pratskevich, Anya (June 13, 2018), "Google Display Ads CPM, CPC, & CTR Benchmarks in Q1 2018," https://blog.adstage.io/google-display-ads-cpm-cpc-ctr-benchmarks-in-q1-2018 (captured by archive.org on 12/14/2018).

Roos, Jason M. T. and Ron Shachar (2014), "When Kerry met Sally: Politics and perceptions in the demand for movies," *Management Science,* 60 (7), 1617–1631.

Rossi, Peter E., Greg M. Allenby, and Robert McCulloch (2005), *Bayesian Statistics and Marketing,* Chichester: John Wiley & Sons, Ltd.

Su, Che-Lin and Kenneth L. Judd (2012), "Constrained optimization approaches to estimation of structural models," *Econometrica,* 80 (5), 2213–2230.

Tanner, Martin A. and Wing Hung Wong (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association,* 82 (398), 528–540.

West, Mike and Jeff Harrison (1999), *Bayesian Forecasting and Dynamic Models,* 2nd ed. New York: Springer-Verlag.

Appendix: Vertical Differentiation from News Volume

Distribution of Unseen News Bits

Recall from Equations 13 and 14 that the probability of bit b being published at site j on day d is $\Pr\left[\imath_{bjd} = 1 | \alpha_j, \pi_b\right] = 1 - \left(1 - \pi_b\right)^{\alpha_j}$ with $\pi_b \sim U(0,1)$, and that consumers' beliefs are consistent with this model of bit availability. Recall as well that because the consumer does not know which bits are available each day, they need only predict the total number of bits at each site.

To derive the Bayesian updating equations, consider first the case when there is one bit available in the environment ($N = 1$), which the consumer has not yet seen ($K = 0$). Suppressing the i and d subscripts for clarity, the likelihood of *not* having seen bit b at one of the previous $t-1$ sites is $\left(1 - \pi_b\right)^{\alpha_{a_1}} \cdots \left(1 - \pi_b\right)^{\alpha_{a_{t-1}}} = \left(1 - \pi_b\right)^{A_t}$, where $A_t \equiv A\left(h_t\right)$ is the sum of α_j's for the $t-1$ sites that were previously visited, per Equation 16. Combining this likelihood with the uniform prior distribution for π_b leads to a beta posterior distribution for π_b.

$$(25) \qquad p\left(\pi_b | A_t, K_t = 0, N = 1\right) = \frac{\left(1 - \pi_b\right)^{A_t}}{\int_0^1 (1-\rho)^{A_t}\, d\rho} = \left(1 - \pi_b\right)^{A_t}\left(1 + A_t\right) = Beta\left(\pi_b \middle| 1, 1 + A_t\right)$$

The step-ahead forecast probability of finding bit b at the next site j is derived by integrating the probability $\Pr\left[\imath_{bj} = 1 | \alpha_j, \pi_b\right]$ over the posterior distribution $p\left(\pi_b | A_t, K_t = 0, N = 1\right)$:

$$(26) \quad \Pr\left[\imath_{bj} = 1 \middle| \alpha_j, A_t, K_t = 0, N = 1\right] = \int_0^1 \left(1 - \left(1 - \pi_b\right)^{\alpha_j}\right)\left(1 - \pi_b\right)^{A_t}\left(1 + A_t\right) d\pi_b = \frac{\alpha_j}{1 + A_t + \alpha_j}$$

Because unseen bits are exchangeable, the extension to more than one bit ($N > 1$) is straightforward: the number of new bits at site j is the result of $N - K_{t-1}$ Bernoulli draws with success probabilities $\alpha_j \big/ \left(1 + A_t + \alpha_j\right)$. This leads to the binomial distribution described in Equation 15.

Relating Bits to Word Counts

Per Equation 15, the state variable K_{id1}—the number of bits encountered at the first site $j = a_1$—is binomial with expected value $\mathbb{E}\left[K_{id1}\right] = N\alpha_j/(1 + \alpha_j)$. In the absence of the daily word counts, w_{jd}, we would sample data-augmented values of K_{id1} from this distribution during estimation. The daily word counts, however, provide a noisy measure of the total amount of news facts published at each site. Thus, we sample data-augmented values of K_{id1} from a binomial distribution with expected value $\mathbb{E}\left[K_{id1} \middle| w_{jd}\right] = N\, q\left(w_{jd}\right)$.

The function $q\left(w_{jd}\right)$ translates daily word counts to the appropriate scale, and is described next. First, note that we only sample values of K_{id1} from this distribution—values of K_{idt} for steps $t > 1$ are sampled from Equation 15. Second, the consumer's beliefs are always represented by Equation 15, even at step $t = 1$.

In choosing a function $q\left(w_{jd}\right)$, we face a constraint: the function $q\left(w_{jd}\right)$ must map w_{jd} to the interval $(0, 1/2)$, because the parameters α_j lie within the interval $(0, 1)$, and thus $\alpha_j \big/ \left(1 + \alpha_j\right) \in (0, 1/2)$. The following half-logit function satisfies this restriction:

$$(27) \qquad q\left(w_{jd}\right) = \frac{2}{1+\exp\left(-w_{jd}c\right)} - 1, \qquad c \equiv \frac{\log 3}{\max\left\{w_{jd}\right\}}$$

Equation 27 is such that if a site publishes zero words on day d, consumer i would see a quantity of news with expected value 0; if the site publishes $\max\left\{w_{jd}\right\}$ words, then consumer i would see a quantity of news with expected value $N/2$.