The Effect of Links and Excerpts on Internet News Consumption

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

April 10, 2018

Abstract

Internet news sites often excerpt content from and link to competing news outlets. On the one hand, links and excerpts can make the excerpting site more attractive, potentially stealing traffic from the sites it links to. On the other hand, excerpting can increase the linked sites' audience by informing readers about that day's news content. We develop a structural model to study the effects of links and excerpts on consumers' browsing behavior. Using data from celebrity news sites, we find that linking to competing news sites increases total news consumption, benefiting both the linking and linked sites. On average, exposures to excerpts and links increase the likelihood of visiting a linked site by .14%, roughly three times the commonly reported effect of an exposure to a display advertisement.

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

*Rotterdam School of Management and ERIM, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, Netherlands; email: roos@rsm.nl; phone: +1 206 317 1713, +31 10 408 2527.

†Fuqua School of Business, Duke University, 100 Fuqua Drive, Durham, North Carolina, 27708; email: mela@duke.edu; phone: +1 919 660 7767.

‡Arison School of Business, Interdisciplinary Center (IDC) Herzliya, Herzliya 46150, Israel; email: ronshachar@idc.ac.il; phone +972 09 960 2408.

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, Yale, Wash. U. in St. Louis, Georgia Tech, Tilburg, UNC Chapel Hill, U. of San Diego, U. of Michigan, INSEAD, U. of Frankfurt, U. of Houston, U. of Pennsylvania, U. of Toronto, U. 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 paper stems from the first author's dissertation.

1 Introduction

On April 18, 2016 Jonathan Adler published an article in *The Washington Post* titled "What will the chief justice do in the U.S. versus Texas (the immigration case)?" In the second sentence, he wrote "…Adam Liptak of *The New York Times* has an interesting article exploring how the chief justice might approach the case." The article linked to *The New York Times'* web site, and, through a two-sentence excerpt from Liptak's article, described Liptak's major theme. Hence, readers of this article not only gained knowledge about the pending immigration case, they also learned more about the information they might find elsewhere.

Prior to the advent of the commercial Internet, it would have been unimaginable for a (print) newspaper like *The Washington Post* to regularly promote content published by *The New York Times*. And yet, thanks to the widespread practice of excerpting and linking, readers of Internet news now encounter this type of content on a daily basis. Over the last two decades, the consumption of news has experienced a fundamental shift as readers have migrated to the Internet, and now more than 70% of U.S. adults consume news over the Internet on a regular basis (Pew Research Center 2016). The practice of excerpting content from and hyperlinking back to other online news sources, which we just described, is a key distinguishing characteristic of this new digital medium.[1]

Excerpts and links play an important role in news consumption because they provide information about other sites' content that individuals might otherwise not observe. The information provided by these excerpts can be especially valuable to consumers because it helps them locate interesting content more efficiently. Understanding how excerpts influence consumers' decisions is therefore central to the broader goal of understanding how the Internet has changed news consumption (Gentzkow and Shapiro 2008; Gentzkow et al. 2011). We therefore seek to assess the consumption of news in an environment enriched by links and excerpts.

The practice of linking among Internet news sites raises several questions pertaining to news consumption, including: 1) How does a site's propensity to link to another site over time affect long-run demand? and 2) What is the short-term effect of a *particular* link on any given day? In other words, we are interested both in cases where a site changes its long-run tendency to link to others, and in cases where, on a specific day, it adds (or drops) a link to another specific site. We wish to understand how both of these phenomena affect traffic (i.e., total number of visitors) at both the linking and linked sites, the number of sites visited, the frequency of browsing, and how these effects vary across individuals. Such insights are relevant to: 1) content producers, who need to know how linking affects their traffic (as this affects their advertising revenue); 2) policy makers,

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

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.

An emerging literature pertaining to how Internet news consumption is influenced by major news aggregators—*Google News* in particular—underscores the importance of excerpting (Athey and Mobius 2012; George and Hogendorn 2013; Concha et al. 2015; Chiou and Tucker 2015; Athey et al. 2017). *Pure* news aggregators do not create news content of their own. Rather, they excerpt from and link to sites that produce original news. In general, studies in this literature—which exploit sudden changes in *Google News's* linking policies due to market entry, lawsuits by *Associated Press* and *Agence-France Presse*, and legislation in Spain and Germany—indicate that aggregators' outbound links increase total 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.

Whereas prior work in this area has focused almost entirely on *Google News*, our interest lies in the more ubiquitous setting in which sites not only excerpt from and link to other news outlets, but also create and publish their own original news content. These *hybrid* news publishers occupy the middle ground between pure news aggregators who only link, and pure news publishers who never link. Hybrid news sites generate a substantial portion, if not the majority, of the links and excerpts most readers will encounter when consuming news. Our research further builds upon the prior literature by unravelling the browsing behavior of consumers in the context of links—specifically, considering *individual readers* rather than *aggregate traffic*, mapping the entire reader browsing sequence, and assessing links' impacts on the choice to visit both linking and linked sites at various stages of the browsing session. By focussing on individual choices, we can measure the causal effects of links on those who are exposed to them.

To achieve this aim, we develop a model of costly news consumption with learning. The individual's utility for news depends on the individual's *daily* match with the news reported at each site (i.e., the horizontal dimension). For example, a consumer may gain higher utility from reading news about certain topics, such as health care, hence their utility from horizontal match with a site's content will be higher on days when it publishes news about health care, and lower when it does not. Consumers, however, are ex-ante uncertain about the utility they will derive from the specific news *each day* (hence, the term *news*). The main theoretical advance in this model comes from formulating links and excerpts as providing noisy signals of the individual's match with excerpted sites' daily content. Importantly, because links provide information about consumers' daily horizontal match, any particular link has the potential to *decrease* the probability some consumers will visit the linked site.[2]

[2]Note however that while our model captures the consumption of news, it does not intend to solve firms' strategic

We consider a model with forward-looking individuals and contrast it with the case in which consumers are not forward-looking. All else equal, forward-looking consumers prefer sites with many outbound links because these consumers anticipate learning match information that improves the efficiency of subsequent browsing. Consumers who are not forward-looking can still learn about their daily horizontal match with excerpted sites, but they do not anticipate this effect when choosing which site to visit. It is only in the forward-looking case that consumers would be more likely to visit sites providing many outbound links (since links provide information that makes subsequent browsing more efficient). One theoretical implication of our forward-looking model of news consumption is that it is possible for excerpting to increase both the tendency of individuals to browse for news and the number of sites visited in each session. In other words, links and excerpts can lead to greater consumption of news, on average, because they improve the efficiency with which news can be obtained.

We bring the model to data using Internet panel data that describe browsing at five celebrity news sites, which we match with data describing the news content published at those sites. Our empirical analysis begins with a model-free investigation of the combined browsing and link data for the purpose of assessing whether an individual link between two sites can *decrease* the likelihood consumers visit the linked site that day (as allowed by the model). We conduct two such analyses. The first, at the consumer level, shows that after being exposed to a link, the likelihood of visiting the linked site is lower, on average, for more than half of the panelists. The second analysis, however, shows that when aggregated across all consumers, the average effect on the linked site's traffic is typically positive. The difference in the two effects is due to the low aggregate baseline probability of visiting a news site in the absence of a link: when the probability of visiting a site is close to zero, it can increase far more than it can decrease.

Although the model-free analysis shows that exposure to links can increase or decrease the probability of visiting the linked site, it cannot be used to determine how consumers' forward-looking expectations about links they might encounter influence their choices. Hence, following the model-free evidence, we adopt a structural estimation approach in order to assess the impact of excerpts on news consumption, and to conduct counterfactual policy simulations.

Whereas previous studies have relied on exogenous changes in the long-run linking behavior of a single site (*Google News*) to estimate the impact of linking on consumers, our basic identification strategy exploits variation in the realization of individual links to other sites each day, whose existence and content are only learned by consumers after their choice to visit the linking site, and are thus exogenous to consumers' choices. We estimate our model by combining two advances from the econometrics (Imai et al. 2009) and statistics (Girolami and Calderhead 2011) literatures, and

choice of when and to whom to link (Mayzlin and Yoganarasimhan 2012; Dellarocas et al. 2013).

our approach provides a template for more efficient Bayesian estimation of single-agent dynamic discrete choice models.[3]

The model estimates provide a view into how news sites differentiate from one another by providing horizontally differentiated content or a higher news volumes. They also underscore the importance of links to consumers: In our setting, observing just one excerpt reduces consumers' uncertainty about their daily match with the excerpted site's content by about 6%.

To measure the overall impact of excerpting on site traffic—and more specifically, to assess the extent to which excerpts increase or decrease traffic at linked sites—we conduct counterfactual simulations. These simulations measure how browsing would have differed had sites not linked to others (as observed). This procedure allows us to quantify the impact of excerpting in terms of site traffic, consumers' propensity to browse, the order of site visits, and other metrics.

We estimate that the total effect of linking is positive for consumers and sites. The median consumer visits .54% more sites, compared to a counterfactual without linking, and traffic to each of the five sites increases between .01% and .18%. The benefits from linking accrue to sites at different stages of consumers browsing sessions, with some sites gaining more visitors at the start of their browsing sessions (due to consumers anticipating the value of seeing outbound links) and others gaining more visitors at later steps (due to consumers who have already seen inbound links). When we consider the latter—choices by individuals who have been exposed to a link—we find that exposure to the link adds .14% to the probability of visiting the linked site, a 2.3% increase. This increase in visit probability compares favorably with standard click-through rates for display ads, which are often at or below .05% (Lambrecht and Tucker 2013; Lewis et al. 2011; Chaffey 2017).

Our results contribute to the previously mentioned empirical literature on Internet news and to the theoretical literature on the impact of strategic linking (Mayzlin and Yoganarasimhan 2012; Dellarocas et al. 2013). We extend this research in part by allowing that exposure to a particular link might not always increase an individual's likelihood of visiting the linked site, an empirical regularity consistent with our data. Our work also adds to the consumer learning literature in marketing. Although ours is a model of costly consumption with learning, it differs from standard models in the vein of Erdem and Keane (1996) in important ways. Most notably, in the standard setting for these models (e.g., consumer packaged goods), consuming a product (e.g., Danon yogurt) provides a signal to the consumer about the true quality of the chosen good (i.e., Danon yogurt). But in our setting, consuming the news at one site (e.g. *dailykos.com*) also signals the characteristics of

[3] Yet another approach to measure the effect of links would be to develop an experiment coordinating links across multiple sites. However, to estimate the impact of linking on traffic at the five news sites we study here, we would need to coordinate linking among these five sites in 256 different experimental cells, rendering this approach impracticable.

the news published at other sites (e.g. *politico.com* and *drudgereport.com*). Notably, for each consumer, our data describes multiple repetitions of the same learning process (i.e., discovering what the day's news is) taking place on different days.

This study is also related to previous work in marketing and economics that has modeled Internet browsing at both the aggregate (e.g. Danaher 2007; Park and Fader 2004) and individual (e.g., Johnson et al. 2004; Lee et al. 2003) levels. The most similar of these models to ours is that of Goldfarb (2002), which also describes utility-maximizing individuals choosing which site to visit next, in consideration of their past browsing decisions and any outbound links they expect to encounter. A key difference is that in our model, excerpts make linking sites attractive because they improve the efficiency of browsing later in the session, not because they generate their own utility.

The remainder of this paper is structured as follows. First we present our model of news consumption in the presence of excerpts, and discuss its theoretical implications. We then describe our data and present preliminary analysis indicating excerpts may signal either positive and negative match. After describing our empirical specification and discussing issues relevant to estimation, we present the structural parameter estimates. Finally, we describe the counterfactual procedure and present its results before concluding with the main insights from this study.

2 A Theoretical Model of Consumer News Consumption

A defining characteristic of both online and offline news is its uncertainty—consumers do not know the news until after they encounter it, otherwise, it isn't news. To illustrate the importance of this point for understanding how excerpts and links affect news consumption, recall the earlier example of *The Washington Post* excerpting from and linking to *The New York Times'* coverage of a federal immigration case.

A person who read the *Times'* article and was also particularly interested in immigration policy might have received higher than normal utility by visiting the *Times* that day. But unless that reader knew something about the *Times'* coverage ahead of time, they would have been no more likely to visit the *Times* than they would have on any other day.

Because *The Washington Post* article excerpted from and briefly discussed the *Times'* coverage, *Post* readers (in particular those who saw the *Post* link and had not already visited the *Times*) would have gained information relevant to their decision of whether to visit the *Times* next. Readers who care about immigration policy might have been more likely to visit the *Times*, whereas those who do not might have been less likely.

This example highlights the core of our model: Consumers have heterogeneous preferences for different types of news coverage, which sites publish anew each day. Consumers are initially unaware of what has been published, but this uncertainty is reduced if consumers encounter excerpts

and links to sites they haven't already visited.

To focus attention on how excerpts help consumers locate the content best matching their individual preferences, we initially present a model in which sites are horizontally (but not vertically) differentiated, assuming each site publishes a unique set of news items each day. We relax these assumptions when describing the empirical model in Section 4, where we account for vertical differentiation and redundant news coverage. To focus the initial discussion, we first present the model from the perspective of a single consumer, keeping in mind that different consumers have different preferences for news.

2.1 Notation, Timing, and Period Utility

Every day, the consumer engages in a browsing session, which we index d. By a *browsing session*, we refer to the process of sequentially visiting zero or more sites within a day (hence not visiting any sites is an option). At each step of the browsing session, indexed $t = 1, \ldots, T_d$, the consumer must decide which (if any) site to visit next. We index consumers with i, and the sites they may visit with j (with $j = 0$ denoting the option of not visiting a site).

Following the literature on sequential browsing in an online setting (e.g., Kim et al. 2010) and matching our empirical setting of celebrity news sites (whose home pages are formatted as blogs, as thus include all content posted each day), we assume the consumer sees all available content at each site visited, and therefore visits each at most once per session. Hence the consumer's choice set, which is initially $\mathcal{J}_{i,d,t}$, becomes $\mathcal{J}_{i,d,t+1} = \mathcal{J}_{i,d,t} \setminus j$ after visiting site j. The consumer's decision can therefore be viewed as the choice of which previously unvisited site to visit next in the current session. We denote by $a_{i,d,t}$ the index j of the option chosen by consumer i at step t of browsing session d.

The utility from visiting site j at step t of browsing session d comprises three parts. The first is $\mu_{i,j,d}$, which denotes the (horizontal) match utility consumer i receives from reading site j's content. We discuss this component of utility in detail below. The second component reflects the notion that viewing sites is costly in terms of time and effort. We denote this cost by $\gamma_i > 0$ and assume it is known to the consumer, and constant over the duration of the browsing session. Alternatively, one can view this as the opportunity cost of foregoing the outside alternative. The third component of utility is $\epsilon_{i,j,d,t}$, which is idiosyncratic and particular to each site at each step of the browsing session. This shock is private information learned just prior to the decision at step t, but is not observed by the researcher.

The period utility individual i gains by visiting site j at step t on day d is

$$U_{i,j,d,t} = \mu_{i,j,d} - \gamma_i + \epsilon_{i,j,d,t} \tag{1}$$

Ending the session (or not starting a session in the first place) is an endogenous choice, yielding

net utility of $U_{i,0,d,t} = \epsilon_{i,0,d,t}$.

2.2 Match Utility from Content

Match utility, denoted by $\mu_{i,j,d}$ in Equation (1), arises from the (horizontal) match between the site's editorial position and the views of the consumer. For example, a liberal consumer might receive higher match utility from topics covered by a Democratic-leaning news blog, such as *dailykos.com*, than from those covered by a Republican-leaning one, such as *drudgereport.com*. Because Internet news sites frequently update their content, the level of match varies from session to session. For example, it is possible that site j usually covers business and finance, but on some days reports on labor and healthcare. Accordingly, the news topics of each day will influence the daily value of $\mu_{i,j,d}$.

We model the match utility consumer i receives from visiting site j in session d as a function of 1) the site's long-run editorial position, z_j, 2) a session-specific, idiosyncratic deviation from this average, $\nu_{j,d}$, and 3) the consumer's horizontal taste, v_i. Specifically,

$$\mu_{i,j,d} = \left(z_j + \nu_{j,d}\right) v_i \tag{2}$$

This formulation implies consumers prefer sites for which sign $z_j = \text{sign}\, v_i$. For instance, a politically conservative consumer with $v_i = -1$ would prefer (on average) news published at a conservative site with $z_j = -1$ over a liberal site with $z_j = 1$.

Based on a potentially long history of browsing, the consumer knows site j's long-run editorial position, z_j. The daily deviation from this average, $\nu_{j,d}$, depends on whatever site j happens to publish, hence the quantity $\nu_{j,d}$ is not observed by the consumer until *after* visiting site j on day d. We assume the daily deviations from the long-run position are independent of the private idiosyncratic shocks, $\epsilon_{i,j,d,t}$, and have the following distribution.

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

In the absence of excerpts from other sites, the consumer has no *ex ante* information about these daily deviations, although we assume τ_ν^{-1} is known from prior browsing.

2.3 Links and Excerpts

While visiting site ℓ during session d, the consumer receives a signal of $\nu_{j,d}$ for some other site j only if site ℓ has excerpted from site j that day. The excerpt might signal, for example, that the editorial position of site j that day is more focused on foreign policy than average. We denote these signals $s_{j,\ell,d}$ and assume they are noisy, but unbiased reflections of sites' true match positions each day.

$$s_{j,\ell,d} | \nu_{j,d} \sim N\left(z_j + \nu_{j,d}, \tau_s^{-1}\right) \tag{4}$$

The notation $s_{j,\ell,d}$ indicates that the signal describing site j (the excerpted site) was observed while visiting site ℓ (the linking site) on day d.

Although the true editorial position ($z_j + \nu_{j,d}$) is a characteristic of site j, the signals provided at each linking site ℓ vary. Hence, even in the unlikely case that two sites excerpt identical content from the same site j, each will signal differ things about j's content, due to differences in the context in which the excerpts are embedded. The extent to which excerpts accurately describe daily deviations from sites' long-run positions (i.e., their level of noise), is denoted τ_s^{-1}, and is constant across sites and known to the consumer.

This setup highlights the informative role of excerpts in helping consumers learn whether the site's daily position is more or less congruent with their preferences. Importantly, because $\nu_{j,d}$ can increase or decrease the consumer's match utility (relative to the site's long-run average) excerpts can signal *lower* than average match, making the consumer *less* likely to visit the linked site.

Learning in our model occurs incrementally at every step t within a single session (day) d. Because new news content is published each day, consumers beliefs revert to their long-run expectations at the start of each session. Accordingly, our model describes consumers who already know their long-run expected match with each site (z_j) but who learn about their daily deviation from that average ($\nu_{j,d}$) over the course of each browsing session. Notably, in this setting, we observe the same consumers repeatedly (i.e., across many sessions) engaging in a learning process that starts from the same initial state of knowledge and proceeds with each step of the browsing session.

Finally, because sites excerpt from each other with asymmetric frequencies, we denote by $\omega_{\ell,j}$ the probability that site ℓ excerpts from site j, and allow $\omega_{\ell,j}$ and $\omega_{j,\ell}$ to differ. This linking strategy is common knowledge in the model, although consumers do not know a priori which links will appear each day: The ω's are known, but their specific realizations (actual links from one site to another) are unknown prior to visiting a linking site.

2.4 Updated Beliefs About Match Utility

The resulting posterior belief about expected match utility on each day arises from a standard application of conjugate normal distributions (West and Harrison 1999):

$$\mathbb{E}\left(\mu_{i,j,d} | I_{i,d,t}\right) = z_j v_i + \left(\frac{\tau_s n_{i,j,d,t}}{\tau_s n_{i,j,d,t} + \tau_v}\right)\left(\overline{s}_{i,j,d,t} - z_j\right) v_i \tag{5}$$

$$I_{i,d,t} \equiv \left\{n_{i,d,t}, \bar{s}_{i,d,t}, h_{i,d,t}\right\} \tag{6}$$

where at step t on day d,

- $n_{i,j,d,t}$ is a state variable indicating the number of sites excerpting j that were visited prior to step t,
- $\overline{s}_{i,j,d,t}$ is a state variable indicating the average match position signaled by all observed ex-

cerpts from site j,

- $h_{i,j,d,t}$ is a binary state variable indicating whether site j has already been visited, and
- $I_{i,d,t} = \left\{n_{i,d,t}, \bar{s}_{i,d,t}, h_{i,d,t}\right\}$ is the collection of all state vectors representing the consumer's information set as of the t^{th} step of the browsing session.

Expected match utility is thus a weighted average of the long-run match ($z_j v_i$) and the average match signaled by any excerpts encountered prior to step t ($\bar{s}_{i,j,d,t} v_i$)—the weights are determined by the signaling precision of the excerpts (τ_s), the variability of match across days (τ_v^{-1}), and the number of excerpting sites visited ($n_{i,j,d,t}$). When the individual starts a new browsing session, $n_{i,j,d,t} = 0$ for every site j, and thus expected match utility is exactly equal to the long-run value ($z_j v_i$).[4] Hence Equation (5) illustrates the value of excerpts to the consumer—on average they shift expectations about match utility away from their long-run average, and toward their actual day-specific values.

2.5 Value Function

When consumers read a site's content, they not only gain current period utility, they also update their beliefs about match utility at other sites. Forward-looking consumers anticipate this updating, and therefore face the standard exploitation-exploration trade off when deciding which site to visit next. For example, a consumer might choose to visit a site that frequently excerpts from many other sites, in the expectation that these excerpts will increase (decrease) the chance of subsequently visiting a site with high (low) match. By choosing sites that are informative about other sites (i.e., those that typically publish excerpts), consumers can increase the value of the rest of their browsing session.

Dropping the i and d subscripts for clarity, the following value function corresponds with the consumer's utility function and beliefs about match utility:

$$V\left(I_t, \epsilon_t\right) = \max\left(\epsilon_{0,t},\ \max_{j \in \mathcal{J}_t \setminus 0}\left\{\mathbb{E}\left(\mu_j \mid I_t\right) - \gamma + \epsilon_{j,t} + \delta \int V\left(I', \epsilon'\right) f\left(I' \mid I_t, j\right) g\left(\epsilon'\right) dI' d\epsilon'\right\}\right) \tag{7}$$

where at step t,

- δ is the rate at which the consumer discounts future utility from browsing,
- $f\left(I' \mid I_t, j\right)$ is the distribution of the next information set given the current information set I_t and choice j, and
- $g\left(\epsilon\right)$ is the distribution of ϵ.

Although the value function (7) is standard, a brief description of the state transition function, $f\left(I' | I_t, j\right)$, is useful here (a complete characterization is given in Appendix B). Recall that $I_t \equiv \left\{n_t, \bar{s}_t, h_t\right\}$ represents the consumers' state of knowledge at step t. First, the state variable h_t,

[4] Thus choices at step $t = 1$ vary only as a result of differences in realizations of $\epsilon_{i,j,d,1}$ across sessions, as is common in discrete choice models.

which indicates which sites were previously visited, evolves deterministically by whichever site j is chosen. Hence h' is always h_t with the addition of an indicator for site j. Second, n_t and $\bar{s}_t$, which indicate the number and average signal value of any previously seen excerpts, evolve stochastically conditional on: 1) which site j is chosen next, 2) which other sites it links to that day, and 3) the signal values contained in those excerpts. Recall that links from site ℓ to site j are observed if the consumer visits site ℓ and site ℓ linked to site j that day, and that the latter occurs with probability $\omega_{\ell,j}$; if no new excerpts are observed at the next site j, then the values of n_t and $\bar{s}_t$ do not change.

The state transition function $f\left(I'|I_t, j\right)$ reflects the effect of choosing to visit site j on the rest of the browsing session. Any excerpts seen at site j improve the precision of the consumer's predicted match utility at the other unvisited sites. For this reason, sites with many outbound links are especially attractive early in the browsing session when the information set of the individual is quite empty. Later in the session, excerpts are most informative when they point to sites that have not already been linked to.

Finally, consumers discount the future value of browsing within the same session at a rate of δ. Browsing decisions on one day do not affect decisions on future days, however, because we model consumers who have already learned sites' steady state characteristics (e.g., their average horizontal position). If consumers are myopic, they do not anticipate seeing excerpts, and thus do not place any additional value on sites that frequently provide excerpts to many other sites, whereas forward-looking consumers do.

2.6 Theoretical Implications

To illustrate the implications of excerpting for consumer behavior, we next consider a stylized setting in which a single consumer chooses among two news sites. The results we report here are based on numerical simulations, with further details and results reported in the Online Appendix.

In this example, one of the two sites (site L) regularly links to the other (site R), but the reverse never happens. Recall that according to our model, excerpts provide unbiased signals, so on average half of these excerpts will signal higher than average (and the other half lower than average) match with site R. To further clarify the discussion and isolate the effect of links, we assume that both sites provide the same average level of match utility (i.e., their z_j's are the same). We further assume the consumer's browsing cost is high enough that each site has less than 50% chance of being visited each day. Even under this highly stylized setup, excerpting plays an important role in the consumer's choice of which site to visit next, and can be either beneficial or detrimental to the linked site. Below we highlight three key results from this analysis:

For sites that are visited infrequently, getting excerpted increases the share of traffic coming from the excerpting site. If the consumer visits site L and sees an excerpt indicating site R's content

is more appealing than usual, then the chance of visiting *R* might increase substantially (remember the baseline probability of visiting site *R* is already low). If instead the excerpt signals *R*'s content is less attractive than usual, then the chance of visiting site *R* might be lower—but not by as much because it is already low to begin with. In the real world, most news sites are not visited very often (George and Hogendorn 2013), hence there is a *floor effect* allowing excerpts to have a positive impact on the linked site (even though half of them on average signal lower match).

The increase in traffic at the excerpted site (*R*) comes from the consumer visiting the linking site (*L*), seeing an excerpt, and choosing to visit site ***R*** rather than ending the browsing session (which might have happened in the absence of the excerpt). For this reason, excerpts increase the average number of sites the consumer visits in each browsing session—that is, total media consumption is higher when sites excerpt.

Because excerpts allow the forward-looking consumer to browse more efficiently, excerpting increases a site's popularity at the start of the browsing session. Consider the consumer's beliefs before visiting any sites. A visit to site *L* will reveal a signal about *R*'s content. If the excerpt signals site *R* is more attractive than usual, the consumer can benefit by subsequently visiting *R*; but if instead it signals *R*'s content is less attractive than usual, then the consumer can benefit by avoiding *R* and ending the browsing session. Hence, even though both sites provide the same level of match utility in expectation at the start of the session (by assumption in this example), starting the session at site *L* leads to higher total expected utility from the entire browsing session.

The increased attractiveness of site *L* has two effects on browsing. First, it increases the number of browsing sessions because the option not to browse is relatively less attractive. This theoretical effect is consistent with empirical results reported in Athey and Mobius (2012) and George and Hogendorn (2013), whereby people who typically started their sessions at *Google News* did so more frequently after the site expanded its news coverage. Second, the increased attractiveness of site *L* leads it to steal some traffic from *R,* because some sessions that might otherwise start at site *R* will start at *L* instead. This is the typical claim made by those seeking to curtail or monetize excerpting (e.g., in the AP and AFP contract disputes with *Google News*). Both of these effects depend on the forward-looking behavior of the consumer, since a myopic consumer who discounts the future at a rate of $\delta = 0$ would not consider the positive effect of site *L*'s excepts when choosing which site to visit first.

The overall effect of excerpting on traffic at excerpted sites can be positive or negative. The preceding discussion implies there are two ways excerpting can increase traffic at the excerpted site. First, excerpting can increase the flow of traffic from the linking site to the excerpted site. Second, it can increase the popularity of the linking site, which, by increasing the total amount of browsing, further amplifies the flow of traffic to the excerpted site. There is a countervailing effect,

however, which can lead to an overall decrease in traffic at the excerpted site: If excerpting makes the linking site popular enough and browsing is costly, then the linking site may end up stealing more traffic from the sites it links to than it provides.

In light of these results, it is evident that the impact of excerpting on linked sites and consumers is an empirical question that depends on a variety of factors, including: 1) the linking frequency among sites, 2) the informativeness of match signals, 3) the relative level of match utility provided by each site, 4) the average frequency of visits to the sites, and 5) the discount applied to future benefits from browsing. In an empirical setting, any of these forces may come to dominate. In other words, the overall effects of linking is a measurement issue ideally suited for a structural model. In the remaining sections, we describe the data, empirical model, and estimation procedure, before turning to parameter estimates and counterfactual results.

3 Data

We estimate our model using data that describe reader 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 via an automated web crawling procedure. We describe both of these data sources before concluding with preliminary evidence that links can either encourage or discourage visits to linked sites.

3.1 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.com*, *dlisted.com*, *egotastic.com*, *perezhilton.com*, and *thesuperficial.com*.

3.1.1 Sample Selection and Consumer Characteristics

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 therefore limit attention to the most active readers, which we define as anyone who 1) visited one or more of the 3,000 sites on at least 16 occasions in Q4 2009, 2) had at least 5 of those visits occur in each of the 3 calendar months, and 3) visited at least 2 of the 5 sites used for this study. Browsing and demographic data for the 127 consumers who fit this profile make up the estimation panel. Using less restrictive thresholds when

Table 1: 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)
perezhilton	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)
thesuperficial	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, hence lower values indicate visits that occurred earlier in the browsing session.

defining the panel leads to the inclusion of consumers who do not browse very often, and thus are probably less familiar with the average match locations and linking frequencies for these sites. In Appendix F, we show that our main results are qualitatively insensitive to these cutoffs.

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 \$55–65k per year. Most panelists have children living with them (57%), and the average household size is 2.7. Five panelists listed their race as African American. We code binary variables as $\{-.5, .5\}$, scale the 7 income categories between 0 and 1 using the center of the category range, and scale household size by subtracting the median (2) and dividing by two standard deviations (2.89). We denote by D_i the row vector of demographic variables for consumer i.

3.1.2 Browsing Data

As mentioned in Section 2, we define the length of a browsing session to be one day, since celebrity news sites operate under the same 24-hour news cycle as other media (Leskovec et al. 2009). For each panelist, we observe the order $t = 1, \ldots, T_d$ in which any of the five sites were visited for the first time each day (the choices $a_{i,d,t}$ in our model).[5] During Q4 2009, the 127 panelists in our estimation sample made 19,130 such choices over the course of 5,757 browsing sessions.

Panelists varied considerably in the subset of sites visited, as well as the order in which sites were typically visited. Table 1 shows that *perezhilton* was by far the most popular site among both

[5]Recall that our model assumes consumers do not return to sites they have already visited. In the estimation sample, we calculate the upper limit on site re-visits to be 3.1% of sessions for the median consumer, and no more than 12% of all sessions. We consider this to be an upper limit because web browsing apps may request pages (e.g., those in open tabs) that have already been read.

Table 2: Empirical Link Frequencies (%)

	Link Target				
Linking Site	*celebuzz*	*dlisted*	*egotastic*	*perezhilton*	*thesuperficial*
celebuzz	-	6.5	0	1.1	9.8
dlisted	69.6	-	68.5	2.2	2.2
egotastic	0	65.2	-	0	0
perezhilton	7.6	0	0	-	0
thesuperficial	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.

male and female consumers, and was visited earliest on average. Although there is variation in the order of sites visited each day, panelists in the estimation sample appear to be stable over time with respect to their average ordering of site visits (i.e., they are not learning which site is their favorite on average). Preference for visiting the other sites differs by gender: male panelists with relatively higher preference for *egotastic* and *thesuperficial*, and female panelists with relatively higher preference for *dlisted* and *celebuzz*.

Males comprise just 35% of the panel, but browsed more often than females. The median male browsed on 46 (out of 92) days, averaging 1.12 sites per session. The median female browsed on 44.5 days, averaging 1.05 sites per session. Although the group averages differ, variation across individuals within each group far exceeds the variation across groups.

3.2 Web Site 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. For each of those days, we use the text scraped from each site to determine 1) which other sites it linked to, and 2) how many words it published. We describe each of these next.

3.2.1 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. Hence, even though we use the shorter term "link" to refer to both the link and excerpt, it is the excerpted content, and not the link per se, that signals consumers' match with the linked site (for this reason, we ignore so-called sidebar, or static links that may appear as part of a site's navigation, but are never accompanied by an excerpt, and we do not consider so-called "around the web" display ads, as these were not used by the sites in our sample). We extract any links that appeared in the body of a news post. Then, using the sequence of sites visited by consumer i and the set of observed links between sites on each day d, we infer the number of match signals to each site that were received at each step of the browsing session

Table 3: 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
perezhilton	0	2,113	4,906	4,482	6,336	9,002
thesuperficial	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.

($n_{i,j,d,t}$ from Equation (5)).[6]

The frequencies with which sites linked to each other (ω in our model) are shown in Table 2. As many sites never linked to each other, half of the entries in Table 2 contain zeros. By contrast, *dlisted* and *egotastic* linked to each other about 67% of the time during Q4 2009. As we explain below when discussing model identification, this variation in realizations of links between sites each day allows us to measure the impact of links and excerpts on browsing.

3.2.2 Word Count Data

In Section 4.1, we describe our empirical strategy for accounting for vertical differentiation among sites on the basis of the volume of news content they publish. Daily word counts at each site (summarized in Table 3) provide an indirect measure of the level of vertical quality achieved by each site each day in terms of its total news volume. Typical word counts vary considerably across sites, and these differences might relate to some sites providing greater utility from vertical quality, on average.

3.3 Preliminary Analysis

Recall from our model that an excerpt can signal either higher or lower match, thereby increasing or decreasing the likelihood of visiting the linked site. Because this aspect of our model runs counter to the standard assumption in the theoretical literature, wherein observing a *particular* link to another site never makes the reader less likely to visit the linked site (Dellarocas et al. 2013; Mayzlin and Yoganarasimhan 2012), we conduct preliminary analysis with the goal of understanding whether consumers in our estimation sample were more or less likely to visit linked sites.[7] Accordingly, we conduct this analysis at the level of individual consumers. We define two empirical choice

[6] 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.

[7] In this theoretical literature, linking can lead to equilibrium outcomes under which consumers browse less on average, but at the *individual* level, seeing a link to another site is assumed to have a weakly positive impact on the likelihood of visiting the linked site.

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:

$$\widehat{\Pr}_i\left(a = j | n_{i,j} > 0\right) = \frac{\sum_d \sum_t 1\left(a_{i,d,t} = j \text{ and } n_{i,j,d,t} > 0\right)}{\sum_d \sum_t 1\left(n_{i,j,d,t} > 0\right)} \tag{8}$$

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

$$\widehat{\Pr}_i\left(a = j | n_{i,j} = 0\right) = \frac{\sum_d \sum_t 1\left(a_{i,d,t} = j \text{ and } n_{i,j,d,t} = 0\right)}{\sum_d \sum_t 1\left(n_{i,j,d,t} = 0\right)} \tag{9}$$

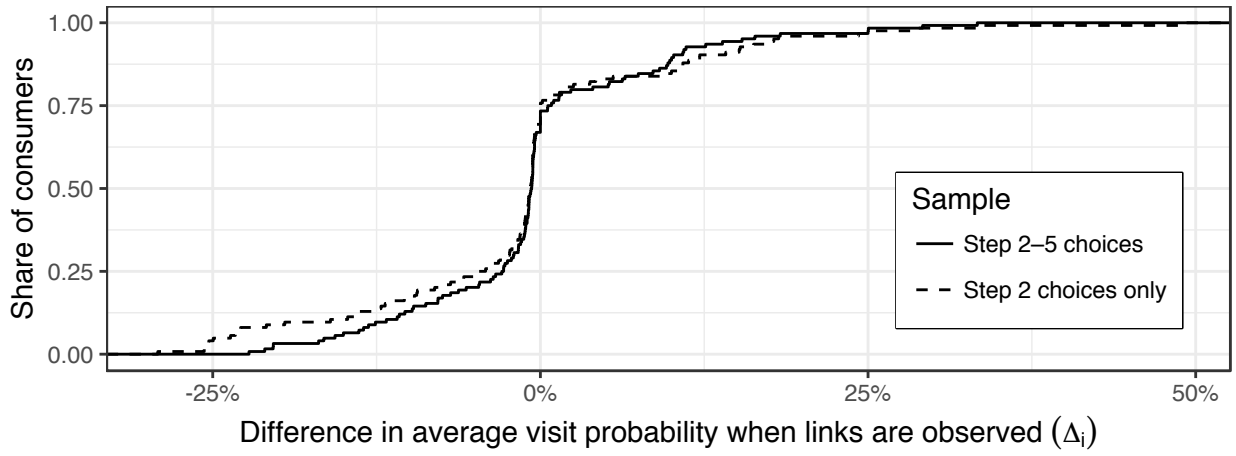
Next, we calculate for each consumer i the frequency-weighted average of each of these probabilities (i.e., averaging across all 5 sites). Thus $\widehat{\Pr}_i\left(a > 0 | n_a > 0\right)$ and $\widehat{\Pr}_i\left(a > 0 | n_a = 0\right)$ denote the probability consumer i visits *any* site a, given prior exposure to $n_a > 0$ links to that specific 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) the linked site, then we expect $\Delta_i > 0$ ($\Delta_i < 0$); if they have no effect, 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 $t = 1$, as well as a subset that includes only choices at step $t = 2$. Figure 1 plots, for both subsets, the empirical cumulative distribution of the difference between the two choice probabilities (Δ_i) across consumers. Individuals with lower visit probabilities after observing links are most prevalent: The left tail corresponds with the majority of consumers who were less likely on average to visit the linked site after seeing links ($\Delta_i < 0$), and the right tail corresponds with the remaining minority who were more likely on average to visit sites after seeing links to them ($\Delta_i > 0$).[8] This evidence provides preliminary support for our modeling approach, whereby links can either increase or decrease traffic to the linked site. Specifically, although previous studies have not taken into account the possibility that links might discourage individuals from visiting the linked site, the data demonstrate that this is indeed possible, and may occur in quite meaningful numbers.

Although this result indicates that excerpting might be detrimental to the linked site (for at least some of its audience), recall that in Section 2.6 we explained how the average effect across all consumers might still be positive under these conditions (due to a floor effect when the probability of visiting the excerpted site is already low). We see evidence for this in Figure 1, as the magnitudes of increases in choice probability (the right tail) are greater than the magnitudes of decreases (the left tail). To test the model implication that excerpts might have an overall positive effect in this setting, we also calculate frequency-weighted averages of the probabilities in Equations (8) and (9) for each site (i.e., $\widehat{\Pr}\left(a = j | n_j > 0\right)$ and $\widehat{\Pr}\left(a = j | n_j = 0\right)$), and find that the average effects are

[8] To verify the numerical robustness of this analysis, we repeat it for each subset of consumers who saw a total of at least ℓ links, for $\ell = 1, \ldots, 50$. 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%.

Figure 1: Average Effect of Exposure to Links on Consumers' Probability of Visiting the Linked Site



NOTE. 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.

positive for four of the sites (ranging, in the $t > 1$ sample, from a .3% increase at *thesuperficial* to a 3.6% increase at *egotastic*), and negative for *perezhilton* (−3.7%).

4 Empirical Model and Estimation

Here we discuss details related to our full empirical model, alternative specifications, model identification, and our MCMC sampling procedure.

4.1 Vertical Differentiation among Sites

Recall the period utility function specified in Equation (1), which includes terms representing horizontal match utility ($\mu_{i,j,d}$), browsing cost (γ_i), and an idiosyncratic shock ($\epsilon_{i,j,d,t}$). Match utility varies by site (j) and session/day (d) depending on whatever content sites publish. At each step t of the browsing session, consumers who observe excerpts and links update their beliefs about the excerpted site's match quality. Together, these components provide a rich specification of horizontal site differentiation and consumer heterogeneity.

However, in most empirical contexts, including ours, sites are also differentiated vertically by the volume of news content published, whereas consumers vary in the value they place on greater news coverage. Augmenting our model to account for this additional vertical dimension of site and consumer heterogeneity has two implications for how our model rationalizes the browsing data. The first is that differences in vertical quality help explain why some sites are more popular than others among all consumers (thus serving the same function as brand intercepts in standard consumer choice models). The second implication is that differences in news volume across sites help to explain why consumers are seldom seen visiting all of the sites in our sample in a single

session (the reason for this is related to the nature of news coverage, as we explain next). Failing to account for differences in vertical quality leads to a model that predicts far more browsing than is observed empirically.

4.1.1 Redundancy in Coverage

In order to model vertically differentiated news sites, we must also consider redundancy in news coverage across sites. To illustrate why these two ideas are related, consider two sites that partially overlap in their coverage of basic news facts (such as which candidate won an election, or when a singer's new album will be released). After visiting the first site, some of the facts presented at the second site will no longer be considered news, because those facts would have already been seen. In the extreme, if the two sites publish every available news item each day, their coverage of the news will be identical, and a consumer could obtain all of the day's news by visiting one or the other. Moreover, after visiting the first site, the amount of *news* items remaining at the second site would be zero, since from the consumer's perspective, none of the second site's coverage would be *new*. Holding total news fixed, higher news volumes imply higher degrees of redundancy with other high-volume news sites.

The main implication of this idea, in terms of modeling vertical differentiation, is that the utility from the vertical quality dimension depends not only on which site is chosen, but also which other sites were previously visited, and how much news they published that day. Utility from vertical quality is therefore state dependent in this setting, changing at each step t of a browsing session depending on which sites were previously visited.

4.1.2 Utility from Vertically Differentiated News Sites

To incorporate utility from vertical quality into the model, we augment Equation (1) with an additional term, denoted $\beta_{i,j,d,t}$.

$$U_{i,j,d,t} = \mu_{i,j,d} + \beta_{i,j,d,t} - \gamma_i + \epsilon_{i,j,d,t} \tag{10}$$

$\beta_{i,j,d,t}$ represents utility obtained from a vertical component of quality that all consumers value in absolute terms (albeit to varying degrees). We use the value gained from learning basic news facts as the canonical example of a vertical component of quality. Accordingly, we normalize the utility from observing no news (and therefore being uninformed about the day's events) to 0 and assume $\beta_{i,j,d,t} \geq 0$. We refer to $\beta_{i,j,d,t}$ as *vertical utility* to distinguish it from the *horizontal match utility* already described in the model by $\mu_{i,j,d}$.

We next describe the most important characteristics of this vertical component of utility. We start with the assumption that on each day d there is an upper limit on the amount of news that can be published. Following Allen (1983, 1986, 1990), we represent this news information as a collection of N unique and indivisible news items, which we call *bits*. These bits of news represent the

smallest unit of news content that can provide vertical utility to the consumer. Every day, a random number of new bits of news are distributed heterogeneously across sites, with the possibility that some bits will appear at more than one site (or none at all).

As described above, a consumer can encounter one of these news items at more than one site. To simplify matters, we assume only the first encounter generates utility, and that thereafter the bit becomes part of the consumer's prior state of knowledge. The utility obtained from seeing a news bit for the first time is heterogeneous across consumers, and indicated by the parameter $\lambda_i > 0$.[9] We denote by $K_{i,d,t}$ the number of unique bits that were seen prior to visiting site j, and by $K_{i,d,t}^{+j}$ the number of unique bits that will have been seen after visiting site j. That is, $K_{i,d,t}^{+j} - K_{i,d,t}$ is the number of previously unseen bits consumer i encounters for the first time by visiting site j at step t of the session. We express the vertical utility consumer i obtains after visiting site j at step t of browsing session d as a function of the number of new news items encountered at each site:

$$\beta_{i,j,d,t} = \lambda_i \left(K_{i,d,t}^{+j} - K_{i,d,t} \right). \tag{11}$$

The quantity $K_{i,d,t}$ is of course known to the consumer prior to visiting site j. However, due to the nature of news, $K_{i,d,t}^{+j}$, indicates a future state of knowledge, and is unknown to the consumer prior to visiting site j.

4.1.3 Distribution of News Items and Consumer Learning

The number of new utility-generating bits of news found at each site is derived from a model of information availability. In this model, each bit b appears at site j with probability $1 - \left(1 - \pi_b\right)^{\alpha_j}$. The probability $\pi_b \sim U(0,1)$ describes both the independent Bernoulli probability that bit b would be published by a (hypothetical) site providing maximum news coverage, and the consumers' prior beliefs. 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. Under this model, an application of Bayes' rule to consumers' beliefs about the amount of news available from site j at step t of session d, conditional on previous site visits, leads to the following binomial distribution for $K_{i,d,t}^{+j} - K_{i,d,t}$:

$$K_{i,d,t}^{+j} - K_{i,d,t} | K_{i,d,t}, h_{i,d,t} \sim Binom \left\{ N - K_{i,d,t}, \frac{\alpha_j}{1 + A\left(h_{i,d,t}\right) + \alpha_j} \right\} \tag{12}$$

$$A(h) = \sum_{k=1}^{J} h_k \alpha_k \tag{13}$$

As mentioned previously, the state variable $h_{i,d,t} \in \{0,1\}^J$ is a binary vector indicating which sites were previously visited, and the term $A(h)$ is the sum of the α_j's for any previously visited sites.

[9] All bits thus generate the same amount of utility for each consumer, and sites are vertically differentiated in the average quantity of these bits published each day. Earlier versions of this paper consider the case where bits differ in the amount of utility they produce.

The term $N - K_{i,d,t}$ represents the maximum amount of news that hasn't already been seen, and the term $\alpha_j \big/ \left(1 + A\left(h_{i,d,t}\right) + \alpha_j\right)$ represents the updated probability that any of this unseen news will be seen at site j if it is visited next. A derivation of this distribution, along with further details, can be found in Appendix A.

Finally, it follows from Equation (12) that the expected vertical utility from visiting site j at step t of browsing session d, is

$$\mathbb{E}\left(\beta_{i,j,d,t} | I_{i,d,t}\right) = \lambda_i \left[\left(\frac{\alpha_j}{1 + A\left(h_{i,d,t}\right) + \alpha_j}\right)\left(N - K_{i,d,t}\right)\right] \tag{14}$$

with $I_{i,d,t}$ now redefined to include the state variable, $K_{i,d,t}$:

$$I_{i,d,t} \equiv \left\{n_{i,d,t}, \bar{s}_{i,d,t}, h_{i,d,t}, K_{i,d,t}\right\} \tag{15}$$

The expected utility from vertical quality in Equation (14) is the expected amount of (previously unseen) news at site j from Equation (12) times the consumer's preference parameter, λ_i.

Equation (14) reflects how the expected vertical utility from news content is higher at sites that publish more news on average (α_j), and for consumers who receive more utility from each unit of news (λ_i), but lower when a large number of news items have already been seen ($K_{i,d,t}$). Perhaps less obvious is that the placement of the $A(h)$ term in Equation (14) means that expected vertical utility is also decreasing in the number of sites that were previously visited—and even more so if those sites have larger values of α_j. The intuition behind this is that any news items that were not already discovered at a site with a higher value of α_j are unlikely to be discovered at a different site with a relatively lower value of α_j, but the reverse is not true.

4.1.4 Incorporating Word Counts

Although we do not observe K directly, we do observe a related quantity: the number of words published at each site each day, $words_{j,d}$. We log-transform these daily word count to define $w_{j,d} \propto \log\left(1 + words_{j,d}\right)$ in order to account for diminishing marginal news content as a function of word count.

We treat these daily word counts ($w_{j,d}$) as noisy measures of the total amount of news content published at each site each day, and use them to draw data-augmented values of $K_{i,d,1}$ during estimation. In our model, the amount of news published on day d at site j is represented by the state variable $K_{i,d,1}$ for consumer i if site j is visited at the start of session d.[10] Hence our estimation strategy is to functionally relate the values of $w_{j,d}$ with realizations of the state variables $K_{i,d,1}$. Details regarding this functional relationship are given in Appendix C.

[10] Because consumer i obtains all available information from each site visited, the realization of the state variable $K_{i,d,1}$—the quantity of information obtained from the first site visited on day d—is also equal to the *total* amount of information available from that site (this is not the case when visiting sites at later steps of the session).

4.2 Consumer Parameters

Consumers are heterogeneous with respect to their values of match preference (υ_i), the utility they receive from each unit of news information (λ_i), and their browsing costs (γ_i). We model this heterogeneity using consumers' observed demographic variables (D_i) via the following prior distributions:

$$\upsilon_i \sim N\left(\eta_\upsilon + D_i\phi_\upsilon, \zeta_\upsilon^2\right), \qquad \log\lambda_i \sim N\left(\eta_\lambda + D_i\phi_\lambda, \zeta_\lambda^2\right), \qquad \log\gamma_i \sim N\left(\eta_\gamma + D_i\phi_\gamma, \zeta_\gamma^2\right) \tag{16}$$

Note that although these prior distributions assume independence among these parameters, this does not rule out any dependencies among their posterior distributions.

We anticipate the incentive to browse could be different on weekends and U.S. Federal holidays (Columbus, Veterans, Thanksgiving, and Christmas Days) due to differences in the value of consumers' time. The following specification allows consumers' γ_i's to differ systematically on these days:

$$\gamma_{i,d} = \begin{cases} \gamma_i \exp\left(\gamma_w\right) & \text{if } d \text{ is a weekend or holiday} \\ \gamma_i & \text{otherwise} \end{cases} \tag{17}$$

All else equal, a value of $\gamma_w > 0$ will lead to less browsing on weekends and holidays.

4.3 Model Likelihood and Bayesian Posterior Distribution

Following the literature on single agent, dynamic discrete choice models (Aguirregabiria and Mira 2010), we assume that the unobserved utility shocks ($\epsilon_{i,d,j,t}$'s) follow an i.i.d $EV\,(0,1)$ distribution. The value of visiting site j, conditional on the state variables $I_{i,d,t}$, is $V_j\left(I_{i,d,t}\right) + \epsilon_{i,d,j,t}$, with $V_j\left(I_{i,d,t}\right)$ denoting the choice-specific value function:

$$V_j\left(I_{i,d,t}\right) = \mathbb{E}\left(\mu_{i,d,j}|I_{i,d,t}\right) + \mathbb{E}\left(\beta_{i,d,j,t}|I_{i,d,t}\right) - \gamma_{i,d} + \delta\int \log\sum_{k\in\mathcal{J}_{i,d,t}\setminus j} \exp\left[V_k\left(I'\right)\right] f\left(I'|I_{i,d,t},k\right) dI' \tag{18}$$

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. Integrating over the unobserved utility shocks (the $\epsilon_{i,d,j,t}$'s) leads to the conditional likelihood of the model parameters, θ, given the observed browsing choices, $a = \left\{a_{i,d,t}\right\}$, state variables, $I = \left\{I_{i,d,t}\right\}$, and other data:

$$L\left(\theta|a, I, \omega, w\right) \propto \prod_i \prod_d \prod_t^{T_{i,d}} \prod_{j\in\mathcal{J}_{i,d,t}} \left\{\frac{\exp\left[V_j\left(I_{i,d,t}|\theta\right)\right]}{1+\sum_{j'\in\mathcal{J}_{i,d,t}}\exp\left[V_{j'}\left(I_{i,d,t}|\theta\right)\right]}\right\}^{1\left(a_{i,d,t}=j\right)} \tag{19}$$

The parameters to be estimated are summarized in Table 4.

The likelihood function depends on the state variables (I), of which K and $\overline{s}$ are not directly

Table 4: Summary of Estimated Parameters

Parameter	Dimension	Description
(z_j, α_j)	5×2	Match location and information quantity for each site
$(\phi_v, \phi_\lambda, \phi_\gamma)$	7×3	Demographic coefficients for horizontal match preferences (v_i), vertical utility (λ_i), and browsing cost (γ_i) parameters
$(\eta_\lambda, \eta_\gamma)$	1×2	Intercepts for vertical utility and browsing cost parameters
$(\zeta_\lambda, \zeta_\gamma)$	1×2	Prior scales for information utility and browsing cost
γ^w	1×1	Incremental browsing cost on weekends and holidays
τ_s	1×1	Precision of link signals
δ	1×1	Discount rate

NOTES: Parameters that are integrated out of the posterior distribution via data augmentation are not listed.

observed by the researcher. To obtain the marginal likelihood $L(\theta|a, \omega, n, w, h)$—where n indicates the observed links, ω the link empirical average link frequencies, w the log-transformed word counts, and h the set of sites previously visited within the current session—we integrate over the distribution of the unobserved state variables K and $\bar{s}$ using the standard Bayesian approach of data augmentation (Tanner and Wong 1987; Rossi et al. 2005). Appendix D lists prior distributions for the remaining model parameters and Appendix E describes the procedure for sampling from the posterior distribution of the model parameters.

4.4 Identification

Here we provide an overview of model identification. We first discuss which aspects of the data provide information about the structural parameters, then we discuss how the setting of online news consumption avoids many of the standard concerns about endogeneity bias, and finally we summarize important parameter normalizations. Additional technical details for all three can be found in Appendix D.

4.4.1 Identification of Structural Parameters

Parameter identification depends on two broad groups of data moments. The first group includes the probabilities that consumers choose to browse each day, and if they do browse, which sites they choose to visit *first*. The second group of moments includes the probabilities with which consumers visit sites at *later* steps of their browsing sessions, conditional on the number of links and amount of content previously seen.

The first group of moments is analogous to what is observed in discrete-choice settings with a single choice observation per period. Hence, choices of the first site visited each day identify average utilities at both the consumer- and consumer/site-level. Consumers' average utilities correspond with the cost parameters, γ_i, which determines the overall amount of browsing. The average consumer/site utilities are factored into the parameters z_j (sites' average horizontal locations) and

υ_i (consumers' horizontal preferences), which jointly determine the sites each consumer visits most often. As in other models that include consumer heterogeneity, we use consumer demographics to estimate parameters governing the distribution of consumers' parameters.

The second group of moments, which characterize subsequent choices on days with browsing, identify the remaining structural parameters. First, the link informativeness parameter, τ_s, is identified by differences in how often consumers choose to visit a particular site after having seen 0, 1, or more links to it. Second, the average utilities from vertical quality at each step are identified by differences in choice probabilities before and after visiting sites that have published more or less content each day. These average utilities are factored into α_j (sites' average vertical qualities) and λ_i (consumers' tastes for vertical quality). And third, the discount parameter, δ, is identified by the joint distribution of choice probabilities at all steps of the browsing session and the average link frequencies between sites. The latter is especially important for identification, as the exclusion of link frequencies from the period utility function means that links affect consumer's choices only through expectations about future browsing, and not by providing their own utility.

4.4.2 Unobserved Components of Utility

Consumers do not know the news until after they read it (indeed, it is consumers' lack of knowledge of the content they are about to consume that drives their demand for it). Hence, true product quality—the horizontal and vertical utilities from news content each day—is not only unobserved by researchers, it is also unobserved by consumers prior to choice. This stands in contrast with the typical setting in which unobserved product quality gives rise to concerns about endogeneity bias. In such cases, consumers know a product's true quality prior to choice, and thus model terms which are assumed to be independent are actually dependent, leading to biased parameter estimates.

In our setting, however, a link from site L to some other site R on a specific day is exogenous to any demand shifting variables that would affect the consumer's decision to visit L. Recall that a consumer who chooses to visit site L knows the long-run frequency with which site L links to site R, but not whether L has linked to R on that particular day. And if there is an excerpt from site R embedded at site L, it provides information about the news content published at site R on that particular day, which, as news, is unknown to the consumer. Consequently, when deciding whether to visit site L, the consumer does not *know* whether there will be a link to site R, nor, in the case there is a link, what that link will say. For this reason, links observed at the individual level provide sufficient exogenous variation to identify their impact on demand. A more technical treatment of this issue can be found in Appendix D.

4.4.3 Parameter Restrictions and Other Issues

We next summarize parameter normalizations that are necessary for estimation. First, recall the term N appears in Equation (14) and represents an upper limit on values of $K_{i,d,t}$ (the number of new news items seen). Model fit is insensitive to this value, as the λ_i's simply scale up or down at different values of N; we set $N = 30$ during estimation. Second, we set $\tau_v = 1$, as we can only identify the ratio τ_s/τ_v. Finally, we restrict $\sum_j z_j = 0$, $\eta_v = 0$, and $\zeta_v = 1$, as the midpoint of expected match utility is not separately identified from the value of the outside option.

5 Results

We estimate a number of nested versions of our model in order to understand how various aspects of the data and model rationalize consumers' browsing. We first compare the various specifications in terms of their fit with the observed data and show the full model provides the best fit.We then discuss parameter estimates from the full specification and their implications for our main research questions.

5.1 Model Fit and Comparison

We consider the full model presented in the previous sections and two nested specifications: The first nested model differs from the full model by restricting the discount parameter δ to be zero. Because this is the same as modeling consumers as if they ignore the value of future browsing when choosing which sites to visit, we refer to this specification as *myopic*. Comparing the full and myopic models provides a view into how much the forward-looking effect of links—anticipation of benefits from a site's outbound links—matters for consumers choices. In the second nested model, we restrict the parameter for the informativeness of links, τ_s, to be zero. Because this corresponds with a model in which excerpts provide no signaling value whatsoever, we refer to this third specification as *no signals*. Comparisons with this specification show the importance of links in explaining individual browsing choices.

We compare the models using two measures of fit. First is the median absolute percent error (MAPE) of posterior predictive distribution of the total amount of browsing, which provides a robust summary of fit with the sample data. Second is the expected deviance, which provides a measure of predictive accuracy (Gelman et al. 2004). Table 5 shows the full specification performs best for both measures, followed by the myopic model, and that the differences between fit statistics are meaningful. This model comparison collectively suggests that links not only provide useful information that allows consumers to find better matching content, but that consumers also anticipate these benefits and use them in the way suggested by the full model. Because models with more parameters may have lower expected deviance due to overfitting, Table 5 also presents

Table 5: Model Fit Statistics

Model	Parameter Restrictions	MAPE	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

Table 6: MAPE of Posterior Predictive Distribution of Total Traffic by Site

Model	Parameter Restrictions	*celebuzz*	*dlisted*	*egotastic*	*perezhilton*	*thesuperficial*
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

the AIC and BIC, which penalize expected deviance by $2k$ and $\ln(n)k$ respectively (for k equal to the number of unrestricted parameters, and n equal to the number of observations).

Table 6 shows the MAPE of total traffic at each site. All three models shown here fit total traffic at *perezhilton* and *egotastic* better than the other three sites. The full model performs relatively worse at *perezhilton* in terms of prediction error, perhaps because *perezhilton* does not provide or receive as many out- and in-bound links as the other four sites. In terms of overall fit, however, the full model has the lowest prediction error, and we present results from this specification below (see Appendix F for further details regarding other specifications).

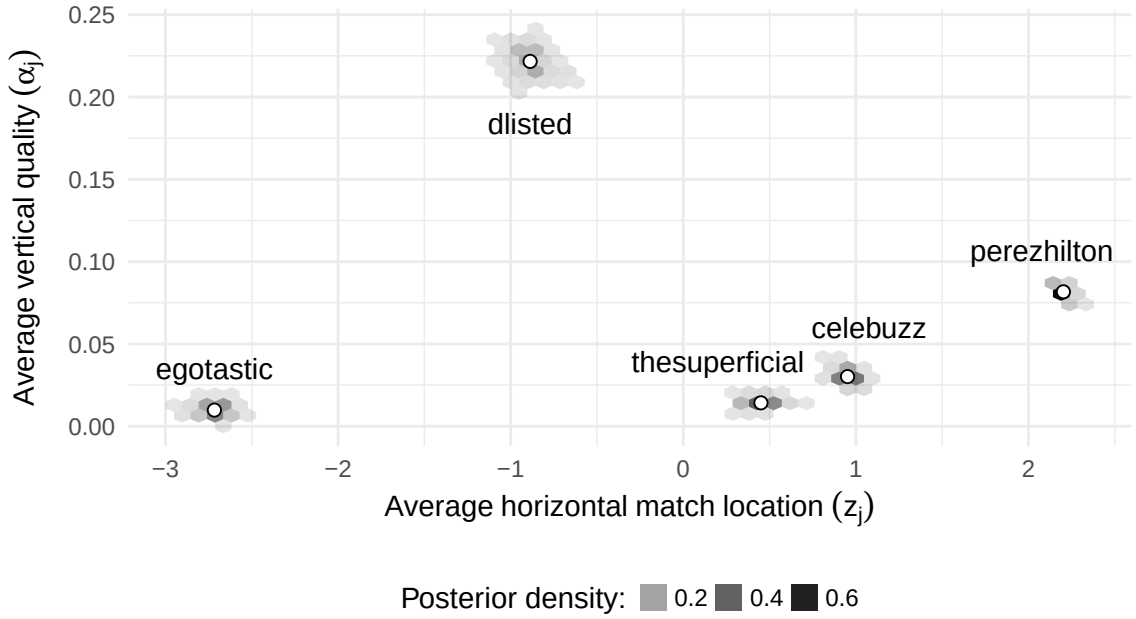
Recall that sites, in our model, provide two dimensions of utility to consumers, a horizontal match component and a vertical quality component. We first present the parameters related to match and the informativeness of links, then present the parameters related to vertical quality and

Table 7: Horizontal (z) and Vertical (α) Quality Parameter Estimates by Site

Site	z_j	α_j
celebuzz	0.95	0.030
	(0.04)	(0.003)
dlisted	-0.89	0.222
	(0.07)	(0.007)
egotastic	-2.72	0.010
	(0.06)	(0.002)
perezhilton	2.20	0.082
	(0.02)	(0.003)
thesuperficial	0.45	0.014
	(0.07)	(0.002)

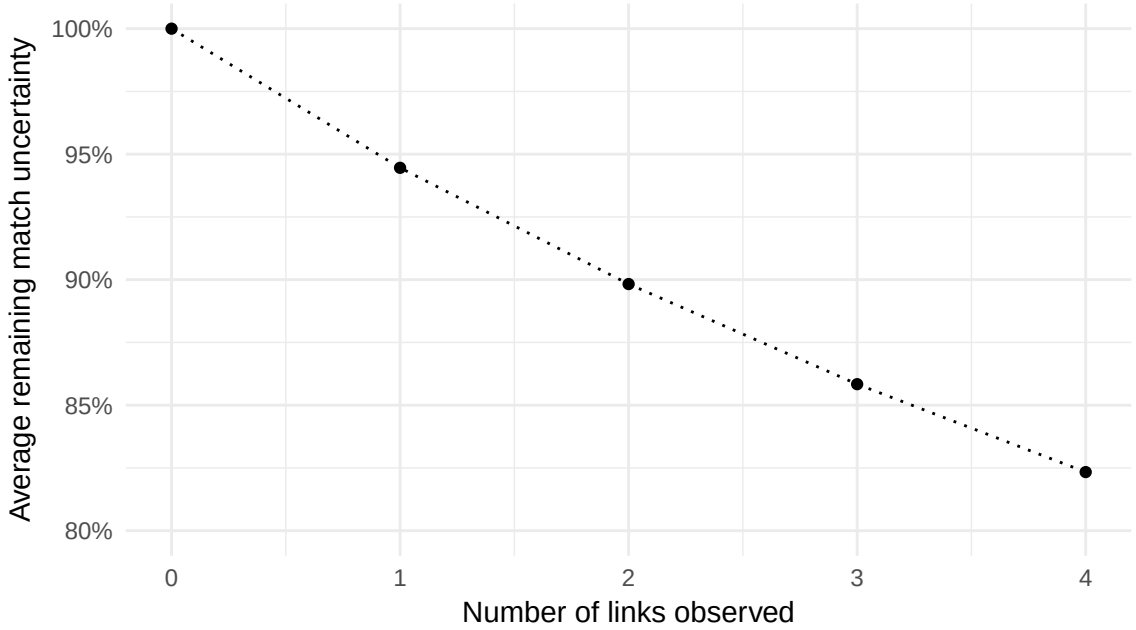
NOTES: Estimates are posterior means with standard deviations in parentheses.

Figure 2: Joint Posterior Distribution of Sites' Average Vertical Quality (α_j) and Horizontal Match Location (z_j)



NOTES. White points indicate locations of posterior means.

Figure 3: 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), $\text{var}\left(\mu_{j,d} | n_{j,d} = 0, \ldots, 4\right) / \text{var}\left(\mu_{j,d} | n_{j,d} = 0\right)$, estimated as the posterior mean of $\left(\tau_s n + 1\right)^{-1}$.

Table 8: Consumer Heterogeneity Parameter Estimates

	Horizontal Match Location (υ)	Vertical Quality Preference ($\log \lambda$)	Cost ($\log \gamma$)
Observed factors			
Female	0.92*	0.79*	0.21*
	(0.18)	(0.26)	(0.10)
Age<25	0.03	−0.56*	0.03
	(0.19)	(0.25)	(0.10)
Age>55	0.08	−1.03*	−0.25
	(0.32)	(0.50)	(0.17)
Income	0.37	0.41	0.17
	(0.24)	(0.37)	(0.14)
Children	0.10	0.02	−0.13
	(0.24)	(0.34)	(0.12)
Household Size	0.09	−0.21	0.08
	(0.24)	(0.33)	(0.12)
African American	−1.38*	0.86	0.09
	(0.36)	(0.52)	(0.20)
Intercept (η)	0.00	−1.94*	1.29*
	-	(0.36)	(0.14)
Unobserved factors			
Prior variance (ζ^2)	1.00	1.30	0.51
	-	(0.10)	(0.04)
Posterior variance	2.21	1.34	0.22
Total heterogeneity			
Posterior variance	2.55	1.66	0.24
Explained by observed factors	12.7%	16.0%	6.0%

NOTES: Estimates are posterior means with standard deviations in parentheses. For observed heterogeneity parameters, asterisks indicate estimates with 95% CI's excluding zero.

browsing cost. We conclude with a discussion of how the parameter estimates yield insights for understanding differentiation among news sites.

5.2 Horizontal Match and Link Informativeness

The average match utility consumer *i* receives from site *j* has two components, a site-specific match location, z_j, and a consumer-specific preference for this location, v_i (c.f. Equation (2)). Here we discuss estimates for both sets of parameters, before turning to the informativeness of links, τ_s.

Sites appear to be horizontally differentiated according to whether they emphasize content that is more *sexy* (e.g., pictorials of attractive female entertainers and models) or *gossipy* (e.g., reporting on the breakup of a celebrity couple). Posterior means and standard deviations for average match locations (z_j) are shown in Table 7, and posterior densities are depicted along the x-axis in Figure 2. The posterior distributions of match locations are negative for *egotastic* (-2.72) and *dlisted* (-0.89), and positive for *thesuperficial* (0.45), *celebuzz* (0.95) and *perezhilton* (2.20). This ordering gives rise to our qualitative interpretation, as it is consistent with the relatively high amount of salacious content published by *egotastic*, *dlisted*, and *thesuperficial*; and, to a lesser extent, these sites' greater reliance on humor and sarcasm when reporting on celebrities. Although *celebuzz* and *perezhilton* also publish sexually-oriented content, they do so less frequently and feature male celebrities much of the time. And although reporting at *celebuzz* and *perezhilton* includes humor and sarcasm, posts at these sites align more closely with traditional tabloid celebrity gossip compared to the other three.

Consumers' preferences for sites' match locations (v_i) are highly heterogeneous, as shown in column 1 of Table 8. This heterogeneity is partly explained by two demographic variables. The most important variable is gender: The match preference coefficient for gender is positive with a 95% Bayesian credible interval (CI) that excludes zero, indicating higher preference among males for sites with $z_j < 0$ (i.e. the sexy content of *egotastic* and *dlisted*), and higher preference among females for site with $z_j > 0$ (the gossipy content of *celebuzz* and *perezhilton*). 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 5 panelists. Altogether, demographic variables account for 12.7% of the total heterogeneity in consumers' preferences for match location.

The true location of each site, and hence the actual amount of match utility received, deviates each day, and links provide consumers with signals about these daily deviations (c.f. Equation (4)). The informativeness of links and excerpts is reflected in the parameter τ_s, which formally represents the precision of link signals around sites' true match locations (relative to the daily variation in match location). The marginal posterior distribution of τ_s has a 95% CI of $(0.01, 0.22)$ with a mean of 0.06. The inverse root of this parameter, $\tau_s^{-1/2}$, is easier to interpret, as it represents

the ratio of the standard deviations of signals and daily match deviations; the posterior mean is 5.2 with a 95% CI of (2.1, 10.3).

Accordingly, excerpts provide informative, but noisy signals about site content. After seeing an excerpt to another site, consumers have a better idea of the coverage at the linked site, but they are still far from certain. We show in Section 6 that the collective information provided by linking across all sites has a meaningful impact on browsing. Figure 3 further illustrates the informativeness of links by showing the reduction in uncertainty about a site's match utility after observing increasingly more links. Observing one link reduces uncertainty about match utility by about 6%; seeing a second link reduces uncertainty by another 4%. Overall, we find compelling evidence that links provide informative signals about match utility at other sites.

5.3 Vertical Quality

Just as with horizontal match utility, we also find site differentiation and heterogeneous preferences for the vertical component of utility related to news volume. Posterior means and standard deviations for sites' vertical qualities (α_j) are also shown in Table 7; posterior densities are depicted along the *y*-axis in Figure 2. *dlisted* is estimated to provide on average the highest level of vertical quality, and *egotastic* and *thesuperficial* the lowest. These estimates reflect both consumers' browsing habits, as well as differences in the amount of content (number of words) published at each site.

Consumers are heterogeneous in their preference for vertical quality (λ_i), as shown in column 2 of Table 8. Demographic variables explain 16% of this heterogeneity, with female consumers and those aged 25–55 receiving the most utility from vertical quality.

5.4 Browsing Cost

We turn next to the parameters for browsing costs. Column 3 of Table 8 shows female consumers have higher browsing costs (γ_i) than males. Collectively, the demographic variables explain just 6% of the variation in $\log \gamma_i$. Consumers with the highest browsing costs visited the fewest number of sites, and were more likely to choose *egotastic* and *perezhilton* (both sources of high match utility) at the start of their sessions. The estimate for γ^w indicates browsing costs are about 7.2% (SD 0.02%) higher on weekends. Browsing costs are measured relative to the value of the outside option, hence this result is also consistent with an outside option (not browsing) that is more valuable on weekends (Ahn et al. 2015).

5.5 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 0.256 (0.253) and a 95% CI of

$(0.001, 0.645)$. 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). This is potentially due to the model imposing a single discount rate for all consumers, when in fact some individuals behave myopically and others in a forward-looking manner.

5.6 Link Frequencies and Site Differentiation

Here 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), vertical qualities (α_j), and linking frequencies (ω_j). Figure 4 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.

From Figure 4 we can see that sites tend to link to competitors with similar values of z_j (i.e., their closest neighbors along the x-axis).[11] 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, excerpts tended to come from sites with very different match locations (e.g., if *egotastic* were to link to *perezhilton*), then consumers would find excerpts to be far less useful, even though the excerpt might be highly informative. As we demonstrate next via counterfactual simulation, a significant portion of some sites' value to consumers stems from their tendency to excerpt from other sites.

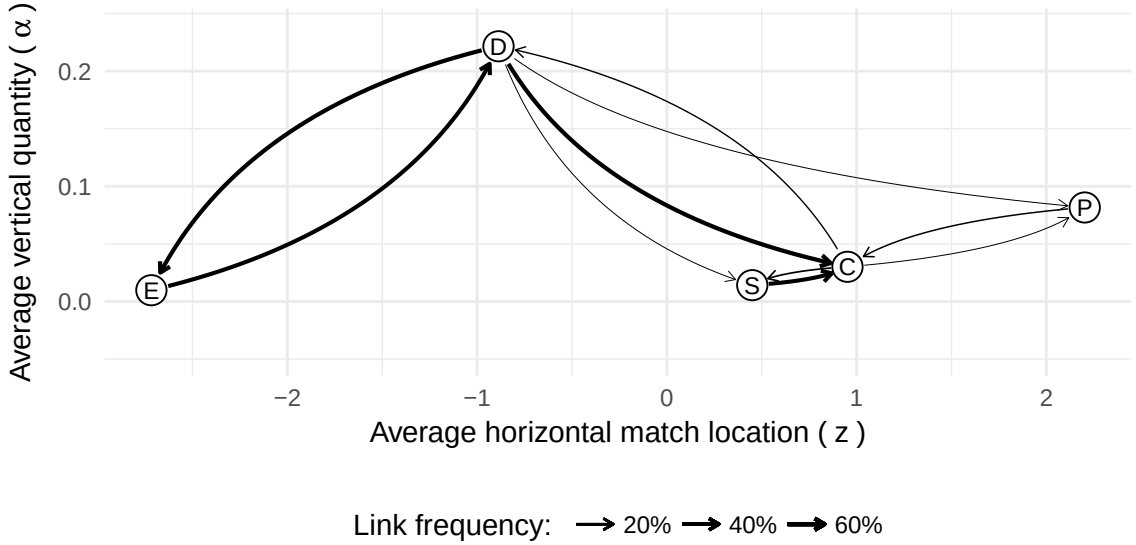
6 Counterfactual Analysis

What are the demand implications of links and excerpts for news consumption? Although the theoretical model shows how excerpting can be either beneficial or detrimental to news sites, it also indicates the importance of contextual factors in determining the total effect of linking. That is, the theoretical results provide an ambiguous answer to the question of linking's effects on traffic.

From the perspective of regulators or firms, ambiguous theoretical outcomes are insufficient for setting policy. For example, in cases where publishers have sued *Google News*, or where legislators have severely curtailed aggregators' ability to link to local sources, there has been a presumption of harm, or at a minimum, that excerpting sites benefit disproportionately from linking (Concha et al. 2015; Chiou and Tucker 2015; Athey et al. 2017). Thus, to understand how links affect behavior in this empirical setting, we conduct counterfactual simulations in which we exogenously manipulate linking and simulate the impact of these changes on browsing. In the remainder of

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

Figure 4: 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 and α. C = *celebuzz*; D = *dlisted*; E = *egotastic*; P = *perezhilton*; S = *thesuperficial*.

this section, we describe this approach, our main insights, and discuss the value of excerpts in this setting.

6.1 Procedure

The objective of this analysis is to understand how excerpting affects demand metrics including the volume of consumer browsing, the flow of traffic between sites, and the number of visitors to each site. The empirical distribution of links between sites, listed in Table 2 and depicted in Figure 4, provides the baseline for these comparisons. Our counterfactual simulation entails removing these links, updating consumers' expectations about (the absence of) excerpting, and simulating browsing.

We simulate the full 92 day sequence of browsing S times for every consumer under the baseline and counterfactual scenarios, with each of the S simulations corresponding to a sample drawn from the data-augmented posterior distribution of the model parameters. For each simulation, we calculate a quantity of interest (e.g. the expected increase in visitors to a particular site) and then average these quantities over these S simulations (i.e., we integrate over the posterior distribution of the parameters).

The counterfactual scenario entails setting the ω's (long-run beliefs for sites' average link probabilities) and n's (number of links observed) to zero. Because consumers' expectations about the links they will encounter at certain sites depend on the ω's, we re-estimate the value function for each of the S parameter draws. Because this is computationally expensive, we set $S = 500$. To

Table 9: Total Effects of Linking on Consumers and Sites

	Percent Change	
	Median Consumer	All Consumers
Number of browsing sessions	0.59%∗	0.11%∗
Sites visited per browsing session	0.14%∗	−0.05%
Total sites visited	0.54%∗	0.06%
Visits to:		
celebuzz		0.10%
dlisted		0.18%
egotastic		0.14%
perezhilton		0.09%
thesuperficial		0.01%

NOTES: Percent changes and differences are expressed relative to the counterfactual with no linking. * indicates 95% bootstrap CI around the estimate excludes 0.

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, the results in this section are framed as changes from the counterfactual with no linking to the baseline with linking. Hence when speaking of a quantity y as the expected percent change from links, we mean $\mathbb{E}_\theta\left[\left(y^{baseline} - y^{counter}\right)/y^{counter}\right]$.

6.2 Results

We present the results in two stages. First, we discuss the total effect of links on consumers and sites at the aggregate level. We then decompose this total effect into two theoretically distinct effects of linking on choice: 1) the effect of forward-looking expectations about links on the propensity to initiate a browsing session (i.e., prior to observing specific links), and 2) the impact of observing a particular link on subsequent choices.

6.2.1 Total effect of linking

Linking has a positive impact on browsing for the median consumer, as shown in Table 9. When we compare the baseline with linking to a counterfactual without links, the number of days on which the median consumer initiates a browsing session increases by .59%, the number of sites visited per session by .14% and the total number of site visits by .54%.

Table 9 also shows the average effect of linking on the total number of sessions and site visits (i.e., total demand across all consumers). The expected increases at the aggregate level are smaller (or even negative) compared to those for the median consumer, as the increases 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.

Table 10: Decomposition of Linking Effects by Step in Browsing Session and Prior Exposure to Link

	Percent Change in Total Visitors		Difference in Choice Probability at Steps $t > 1$	
	Visitors Arriving at Step $t = 1$	Visitors Arriving at Steps $t > 1$	Exposure to Actual or Removed Links	No Prior Exposure to Links
celebuzz	0.21%	0.45%	0.04%	0.01%
dlisted	0.21%	0.42%	0.26%∗	−0.02%
egotastic	0.03%	1.32%∗	0.08%∗	0.00%
perezhilton	0.21%∗	−0.55%	0.47%	−0.14%∗
thesuperficial	0.76%	−0.14%	0.27%∗	−0.03%
All sites	0.11%∗	−0.07%	0.14%∗	−0.02%∗

NOTES: Percent changes and differences are expressed relative to the counterfactual with no linking. * indicates 95% bootstrap CI around the estimate excludes 0.

Turning to the site-specific browsing results, Table 9 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 9 are averaged across conditions in which links affect choices to varying degrees and via different mechanisms. At the start of a browsing session, for instance, only forward-looking consumers' expectations about links (i.e. ω) influence choice. In contrast, subsequent browsing is also affected by the observation of links (i.e. n), which change beliefs about horizontal match at other sites. Because sites do not always link and consumers do not visit every site, we interpret the results in Table 9 as the total effects of a broader policy of excerpting, with the understanding that some consumers may see few, or perhaps none of the linked content. To understand how exposure to any particular link affects consumers' choices, we must further decompose the total effect into its constituent parts.

6.2.2 Decomposition of the total effect of linking

The effect of changes in initial beliefs about linking probabilities. Our theory suggests links should have a different effect on decisions at step $t = 1$ compared to later steps: At step $t = 1$ of a browsing session, choices are affected by forward-looking consumers' expectations about links they may encounter, but not by updated beliefs about daily match (as no links have been seen yet). The first two columns of Table 10 show how linking changes site traffic differently at the first step of consumers' browsing sessions (when only expectations of links contribute to choice) compared to later steps (when both expectations and prior exposure to links matter).

The differences for *egotastic*—which gains little traffic at step $t = 1$ (suggesting its outbound likes are of relatively little value), but a substantial amount at later steps—provide a useful illustration of how the effects of links depend on how sites are differentiated and which other sites they link to. Recall that *egotastic* gives and receives a large number of links, but only in exchange with

dlisted, whereas *dlisted* 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 in expectation than *dlisted*'s. Accordingly, at step $t = 1$, *egotastic* actually loses a portion of its audience to *dlisted*, offsetting any gains from providing its own outbound links, and leading to a net increase of just .03%. Nevertheless, because *dlisted* links to egotastic rather frequently, the number of visitors to *dlisted* at later steps increases by 1.32%.

Isolating the effect of links on the exposed. Although the increase in traffic at later stages is relatively large for *egotastic*, it still represents an average over cases in which some consumers see a link and others do not. Thus, we would like to compare consumers' choices after they have been exposed to a link with those same choices under the counterfactual without linking. The challenge with making this comparison is that, as Figure 4 shows, sites often link to their closest neighbors in terms of match location, and 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 it under the counterfactual of no linking.[12] By comparing baseline choices in which a link was observed with counterfactual choices in which a removed link *would have been observed,* had it not been deleted, we difference out cases when individuals see a link to a site they would visit even if there was no link. That is, the effect we measure is almost entirely due to the exogenous presence or absence of the link itself. These differences in consumers' propensity to visit a linked site are somewhat analogous to click-through rates for (untargeted) Internet ads, in the sense that they are calculated predicated on exposure to a particular link (or ad).

The third and fourth columns of Table 10 summarize the results of this analysis. The third column of Table 10 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 an excerpt is .26% higher than it would be without the link. This represents a 3.8% increase in the amount of *dlisted*'s traffic coming from linking sites.

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. The magnitude of this estimate is higher than paid forms of links, such as display advertising, which typically have click-through rates less than .05% (Lambrecht and Tucker 2013; Lewis et al. 2011; Chaffey 2017).

[12]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 missing link. That is, these consumers would have seen the link to site R, if not for us removing it under the counterfactual.

7 Conclusion

Understanding how excerpting among news sites affects consumers is relevant to 1) content producers, who need to know whether excerpting will be more beneficial to their own sites or the competition's; 2) policy makers, who need to understand whether excerpting generates value for consumers or creates an unfair competitive advantage for content aggregators; and 3) advertisers, who need to know how changes in linking affect the reach and frequency of ads running on multiple sites. In this paper, we present a theory that distinguishes the effects of excerpting on the linking and linked site, and thus generates new insights into the question of why excerpts can be beneficial to the excerpted site in some circumstances and detrimental in others. Moreover, we quantify the magnitude of these effects in an empirical setting in order to assess how excerpting influences consumers browsing for Internet news. These efforts advance our understanding of excerpting, and more generally, the consumption of Internet news.

A novel aspect of this research is that excerpts are modeled as signals of consumers' heterogeneous match with content at the excerpted site. This signaling mechanism allows excerpts to either increase or decrease the likelihood of subsequently visiting the linked site, depending on the valence of the signal. Yet even though our model allows any given link to have a negative effect at the individual level, it also provides a theoretical rationale for why the practice of excerpting may still be generally positive for both the linking and linked sites at the aggregate level.

Specifically, our theoretical results indicate that when the prior probability of visiting an excerpted site is already low, any decreases in the probability of visiting the excerpted site due to lower expected match will be smaller in magnitude than any increases due to higher expected match. For this reason, excerpting should generally have a positive direct effect on the linked site. However, we find there is another factor at work. Because consumers value excerpts, sites that offer them become more popular. If this increase in popularity is large enough, excerpting can also have a negative indirect effect on the excerpted sites.

Our empirical results reinforce these insights. Excerpting benefits the excerpted site by increasing the share of its traffic originating at the linking site, and benefits the linking site by making it more popular at the start of consumers' browsing sessions. Although the average overall impact is positive for both sites, the effects are heterogeneous, depending on the characteristics of the linking and excerpted sites, and by extension, their relative popularity at different stages of consumers browsing sessions. Our results indicate that excerpting typically increases traffic at both the excerpting and linked sites. Compared to a counterfactual in which all links have been removed, exposure to a link increases the probability of visiting the linked site on average by .14%, a 2.3% increase. That is, excerpting from another site increases the excerpted site's traffic at a level greater

than that for typical display ads (which have click-through rates in the range of .05%). When aggregated over time and across sites, the impact of linking among news sites translates into meaningful increases in news consumption. Owing to the nature of digital advertising, such gains in traffic translate directly into higher ad revenue and profit.

In addition to generating new theoretical and substantive insights about the consumption of news on the Internet, this study also provides a number of methodological advances. First, our model of excerpts as match signals can be easily applied to other settings where consuming one product leads to learning about another, or where a firm's advertising contains information about its competitors. Second, we formulate a model of news consumption with learning that can be directly applied to the study of other (non-Internet) news media. And third, our estimation procedure provides a template for more efficient Bayesian estimation of single-agent dynamic discrete choice models.

There are a number of limitations to this study that may provide the basis for future extensions. First, we do not model the strategic decision of whether to link to another site. The decision of which sites to link to may depend, for example, on how similar sites are, or on their relative market power, as well as the distribution of consumer preferences. An empirical study that accounts for these factors might provide new insights into the popularity of excerpting among blogs and news sites. Another limitation of this study is that the match locations and link signals are unobserved. An interesting extension would be to model the site's match location as a function of its content, and the match signal as a function of the text immediately surrounding the link. Such insights would guide the design and content of links. Finally, this study has limited its focus to the practice of excerpting among Internet news sites. But excerpting is far more widespread than the specific context of news sites. Thus, it would be valuable to understand how the effects of excerpting differ in other contexts, such as Twitter, Internet discussion boards, and other social media.

References

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

Ahn, D.-Y., J. A. Duan, and C. F. Mela. 2015. "Managing user-generated content: A dynamic rational expectations equilibrium approach." *Marketing Science* 35 (2): 284–303.

Allen, B. 1990. "Information as an economic commodity." In *Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association*, ed. by R. L. Oaxaca and W. St. John, 268–273. Pittsburgh: American Economic Association.

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

— . 1986. "The demand for (differentiated) information." *The Review of Economic Studies* 53 (3): 311.

Athey, S., and M. Mobius. 2012. “The impact of news aggregators on Internet news consumption: The case of localization.” Working paper.

Athey, S., M. M. Mobius, and J. Pál. 2017. “The Impact of Aggregators on Internet News Consumption.” Stanford University Graduate School of Business Research Paper No. 17-8. Available at SSRN: https://ssrn.com/abstract=2897960.

Bell, B. M. 2007. *CppAD: A Package for Differentiation of C++ Algorithms*. `http://www.coin-or.org/CppAD`.

Chaffey, D. 2017. “Display advertising clickthrough rates.” Visited on 11/15/2017. `https://web.archive.org/web/20171115183841/https://www.smartinsights.com/internet-advertising/internet-advertising-analytics/display-advertising-clickthrough-rates/`.

Ching, A. T., et al. 2012. “A practitioner’s guide to Bayesian estimation of discrete choice dynamic programming models.” *Quantitative Marketing and Economics* 10 (2): 151–196.

Chiou, L., and C. Tucker. 2015. “Content aggregation by platforms: The case of the news media.” NBER Working Paper No. 21404.

Concha, P. P. de la, A. G. García, and H. H. Cobos. 2015. *Impacto del nuevo Artículo 32.2 de la Ley de Propiedad Intelectual*. Tech. rep. NERA Economic Consulting. Summarized in `https://web.archive.org/web/20150814111804/http://www.aeepp.com/noticia/2272/actividades/informe-economico-del-impacto-del-nuevo-articulo-32.2-de-la-lpi-nera-para-la-aeepp.html` as accessed via Google Translate.

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

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

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

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

Gelman, A., et al. 2004. *Bayesian Data Analysis*. 2nd. Chapman & Hall/CRC.

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

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

George, L. M., and C. Hogendorn. 2013. “Local news online: Aggregators, geo-targeting and the market for local news.” Working paper.

Girolami, M., and B. 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, A. 2002. “Analyzing website choice using clickstream data.” *Advances in Applied Microeconomics* 11:209–230.

Griewank, A., D. Juedes, and J. 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.

Imai, S., N. Jain, and A. Ching. 2009. “Bayesian estimation of dynamic discrete choice models.” *Econometrica* 77 (6): 1865–1899.

Johnson, E. J., et al. 2004. "On the depth and dynamics of online search behavior." *Management Science* 50 (3): 299–308.

Kim, J. B., P. Albuquerque, and B. J. Bronnenberg. 2010. "Online demand under limited consumer search." *Marketing Science* 29 (6): 1001–1023.

Lambrecht, A., and C. Tucker. 2013. "When does retargeting work? Information specificity in online advertising." *Journal of Marketing Research* 50 (5): 561–576.

Lee, S., F. Zufryden, and X. Drèze. 2003. "A study of consumer switching behavior across Internet portal web sites." *International Journal of Electronic Commerce* 7 (3): 39–63.

Leskovec, J., L. Backstrom, and J. Kleinberg. 2009. "Meme-tracking and the dynamics of the news cycle." In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, 497–506. New York: ACM.

Lewis, R. A., J. M. Rao, and D. H. Reiley. 2011. "Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising." In *Proceedings of the 20th international conference on World Wide Web*, 157–166. ACM.

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

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

Park, Y.-H., and P. S. Fader. 2004. "Modeling browsing behavior at multiple websites." *Marketing Science*: 280–303.

Pew Research Center. 2016. "The modern news consumer: News attitudes and practices in the digital era." `http://web.archive.org/web/20160713044619/http://www.journalism.org/files/2016/07/PJ_2016.07.07_Modern-News-Consumer_FINAL.pdf`.

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

Rossi, P. E., G. M. Allenby, and R. McCulloch. 2005. *Bayesian statistics and marketing*. John Wiley & Sons, Ltd.

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

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

West, M., and J. Harrison. 1999. *Bayesian forecasting and dynamic models*. 2nd. Springer-Verlag.

A Microfoundations for the Model of Information Availability

As mentioned in the main text, there are at most N unique news bits available to consumer i on each day d (to simplify notation, we suppress the i and d subscripts in this section). These bits are distributed heterogeneously across sites, and a bit can appear at more than one site, or perhaps none at all. When a consumer encounters a news bit for the first time, it provides an amount of utility, and thereafter enters the consumer's state of knowledge; hence further encounters with that bit (at other sites) will provide no further utility.

The probability of any bit b appearing at site j is composed of two factors. The first pertains to the availability of the bit in the environment, the second to its availability at site j. The first

factor is the random variable $\pi_b \in (0,1)$ and is common across all sites; it represents the baseline Bernoulli probability that bit b would appear at a (hypothetical) site capable of covering all of the day's news. The second factor is the site-specific parameter $\alpha_j \in (0,1)$ and is particular to site j, but common across all bits. The two factors jointly define the probability that any bit b appears at site j, which as noted in the main text, is $1-\left(1-\pi_b\right)^{\alpha_j}$. This probability is such that if site j publishes more information on average, i.e., $\alpha_j \to 1$, the probability news bit b is at site j is π_b—and if j publishes less information on average ($\alpha_j \to 0$), the probability b is at site j goes to zero. The parameter α_j thus attenuates the probability of finding information at site j relative to the overall news environment.

From the consumer's perspective, it is not necessary to predict the entire set of information at each site, but instead just the number of bits that were not already seen. However, to simplify the exposition, consider the case when there is just a single bit available in the environment and the consumer is keeping track of it. The baseline bit probability π_b is an i.i.d. uniform random variable: $\pi_b \sim U(0,1)$. We assume the consumer's prior belief about the availability of news content is consistent with this.

Denoting by α_t the value of α_j for the site visited at step t, we can write the likelihood of *not* having seen bit b at any of the previous $t-1$ sites as $\left(1-\pi_b\right)^{\alpha_1}\cdots\left(1-\pi_b\right)^{\alpha_{t-1}} = \left(1-\pi_b\right)^{A_t}$, where $A_t \equiv A\left(h_t\right)$ denotes the sum of α_j's for the sites already visited, as defined in Equation (13). Combining this likelihood with the uniform prior distribution for π_b leads to a beta posterior distribution for π_b:

$$p\left(\pi_b|A_t\right) = \frac{\left(1-\pi_b\right)^{A_t}}{\int_0^1\left(1-\pi_b\right)^{A_t}d\pi_b} = \left(1-\pi_b\right)^{A_t}\left(1+A_t\right) = Beta\left(\pi_b\middle|1,1+A_t\right) \tag{A.1}$$

The (step-ahead forecast) probability of finding bit b at the next site j is then derived by integrating over the posterior distribution of π_b and the probability bit b has been published by site j.

$$\Pr\left[b \text{ is at } j \,\middle|\, b \text{ not seen previously}\right] = \int_0^1\left(1-\left(1-\pi_b\right)^{\alpha_j}\right)\left(1-\pi_b\right)^{A_t}\left(1+A_t\right)d\pi_b = \frac{\alpha_j}{1+\alpha_j+A_t} \tag{A.2}$$

Because unseen bits represent a priori unknown news information, these bit probabilities are exchangeable for any bits that haven't already been encountered. Hence, the number of new bits to be received by visiting the next site j (Equation (12) in the main text) is binomial, comprising the sum of $N-K_t$ Bernoulli draws each with success probability $\alpha_j/\left(1+\alpha_j+A_t\right)$.

B State Variables and Transition Probabilities

Here we provide a full specification of the state variables and their transition probabilities. The consumer's information state is the set $I_t \equiv \left\{n_t, \bar{s}_t, K_t, h_t\right\}$, and we note that $A_t \equiv \sum_j h_{t,j}\alpha_j$. The

probability of the next I' conditional on the previous I_t and the decision to visit site j is denoted

$$f\left(I'|I_t,j\right)=f\left(n',\overline{s}',K',h'|n_t,\overline{s}_t,K_t,h_t,j\right) \tag{B.1}$$

We decompose this distribution in the following way:

$$f\left(I'|I_t,j\right)=p\left(\overline{s}'|n',n_t,\overline{s}_t\right)p\left(n'|n_t,j\right)p\left(K'|K_t,h_t,j\right)p\left(h'|h_t,j\right) \tag{B.2}$$

The distribution of $p\left(K'|K_t,h_t,j\right)$ is specified in (12). The distribution of h' is deterministic conditional on the choice j. Using $\dot{\delta}_x$ to indicate the degenerate (Dirac) distribution with a point mass at x, we write this distribution as

$$p\left(h'|h_t,j\right)=\dot{\delta}_1\left(h_{t,j}\right),$$

which indicates the next h' is equal to the previous h_t after changing h'_j to 1. The evolution of n and $\overline{s}$ are specified as follows. First, the distribution of n' is discrete: for any site $j'\neq j$ that has not yet been visited, $n'_{j'}$ will equal $n_{t,j'}+1$ with probability $\omega_{j,j'}$, and $n_{t,j'}$ with probability $1-\omega_{j,j'}$. Formally,

$$p\left(n'_{j'}|n_t,j\right)=\omega_{j,j'}\dot{\delta}_{n_{t,j'}+1}\left(n'_{j'}\right)+\left(1-\omega_{j,j'}\right)\dot{\delta}_{n_{t,j'}}\left(n'_{j'}\right) \tag{B.3}$$

When a new link to site j' is observed, the value of $\overline{s}'_{j'}$ evolves according to the rules for Bayesian updating of standard Normal conjugate distributions, and when no new link is observed, then $\overline{s}'_{j'}=\overline{s}_{t,j'}$. Formally,

$$p\left(\overline{s}'_j|n'_j,n_{t,j},\overline{s}_{t,j}\right)=\begin{cases}N\left(z_j+\frac{\tau_s n_{t,j}\left[\overline{s}_{t,j}-z_j\right]}{\tau_s n_{t,j}+\tau_v},\tau_s^{-1}+\left[n_{t,j}\tau_s+\tau_v\right]^{-1}\right), & n'_j=n_{t,j}+1\\ \dot{\delta}_{\overline{s}_{t,j}}\left(\overline{s}'_j\right), & n'_j=n_{t,j}\end{cases} \tag{B.4}$$

C Word Counts and Information Quantity State Variables

In this section we provide technical details about the relationship between word counts, $w_{j,d}$, and consumers' state variables for news quantity, $K_{i,d,t}$. As shown in Equation (12), the state variable $K_{i,d,1}$ follows a binomial distribution with expectation $\mathbb{E}\left[K_{i,d,1}\right]=N\alpha_j/(1+\alpha_j)$. If we lacked word count data ($w_{j,d}$), we would draw data-augmented values of $K_{i,d,1}$ from this distribution during estimation. The word count data, however, function as noisy measures of the total amount of news content published at each site each day, allowing us instead to draw data-augmented values of $K_{i,d,1}$ during estimation from a binomial distribution with expected value

$$\mathbb{E}\left[K_{i,d,1}\middle|w_{j,d}\right]=N\,q\left(w_{j,d}\right) \tag{C.1}$$

The function $q\left(w_{j,d}\right)$ translates the number of words published at site j on day d to the appropriate scale, and is described below. First, note that we only draw values of $K_{i,d,1}$ using Equation (C.1), but not subsequent values of $K_{i,d,t}$ for steps $t>1$. Moreover the consumer's beliefs are always represented by Equation (12).

In selecting a function $q\left(w_{j,d}\right)$, we face the following constraint: The function $q\left(w_{j,d}\right)$ must

map $w_{j,d}$ to the interval $\left(0, \frac{1}{2}\right)$ because the parameters α_j lie within the interval $(0, 1)$, and thus $\alpha_j / \left(1 + \alpha_j\right) \in \left(0, \frac{1}{2}\right)$. The following half-logit function satisfies this restriction.

$$q\left(w_{j,d}\right) = \frac{2}{1+\exp\left(-w_{j,d}c\right)} - 1 \qquad c \equiv \frac{\log 3}{\max\left\{w_{j,d}\right\}} \tag{C.2}$$

Equation (C.2) is such that if a site publishes zero words on day d, consumer i would see a quantity of news with expected value $K_{i,d,1} = 0$; if the site publishes $\max\left\{w_{j,d}\right\}$ words, then consumer i would see a quantity of news with expected value $K_{i,d,1} = N/2$.

D Identification Details

Here we provide further technical details regarding model identification. As in the main text, we first discuss the data moments identifying the model parameters before considering the independence requirements for error terms related to unobserved utility, and parameter normalizations.

D.1 Identification of Structural Parameters

Table 11 summarizes which data moments and model assumptions identify the structural parameters. The first row of Table 11 indicates requirements for identifying consumers' expected match at step 1 of the session (denoted $\mathbb{E}\left(\mu_{i,1}\right)$), and their browsing costs (denoted γ_i). Identification is due to differences in choice shares at the start of the browsing session (denoted $a_{i,1}$), conditional on the observed link frequencies from the first site visited to all other sites (denoted ω_1). Separate identification of $\mathbb{E}\left(\mu_{i,1}\right)$ from γ_i depends on a restriction that ensures the $\mathbb{E}\left(\mu_{i,1}\right)$'s sum to 0, additivity in the utility function, and independence of the $\epsilon_{i,j,d,1}$'s (idiosyncratic private shocks to utility). Separate identification of the structural parameters for site's horizontal locations (z_j) and consumers horizontal tastes (v_i) is due to a factorization of expected match utility into site-specific and consumer-specific components, as shown in the second row of Table 11.

The third row of Table 11 indicates that identification of the discount parameter (δ) comes from the joint distribution of the link probabilities and consumers' choice shares across all steps of the browsing session (denoted ω), as well as the exclusion restriction that link frequencies only affect choices through consideration of future utility (or stated equivalently—that observing links does not directly generate utility). Note that conditioning on ω is not necessary to identify any of the structural parameters in models which assume consumers are myopic (i.e., when $\delta = 0$).

Identification of the remaining structural parameters depends on choice shares at the second step of the browsing session ($t = 2$).[13] As shown in row 4 of Table 11, conditional on the link frequencies from the second site visited to all others (denoted ω_2) and the number of words seen at the first site visited (denoted $w_{i,1}$), differences in the joint distribution of choice shares at the

[13] Subsequent choices ($t > 2$) provide additional information used during estimation, but are not necessary to establish identification.

second step of the browsing session (denoted $a_{i,2}$) and the number of excerpts that were seen at the first site (denoted $n_{i,1}$) identify the informativeness of links (the ratio τ_s/τ_v). Identification of this ratio further depends on the assumption of Bayesian updating, the distributional assumptions for daily match deviations ($\nu_{j,d}$) and signals ($s_{j,\ell,d}$), and the assumption that the $\nu_{j,d}$'s and $s_{j,\ell,d}$'s are independent of the $\epsilon_{i,j,d,t}$'s. Intuitively, after controlling for everything else the consumer knows about a site, if her choice shares are correlated with the number of links she has seen, then links are informative, whereas if choices are uncorrelated with the number of links she has seen, then links are uninformative.

As rows 5 and 6 of Table 11 indicate, similar arguments establish identification for expected utility from vertical quality. Conditional on everything else the consumer knows about a site, variation in choice shares and the number of words already seen (plus the assumptions surrounding Bayesian updating) identifies expected vertical utility. Hence, after controlling for everything else, if the consumer is more (less) likely to end a particular session after visiting a site with a large (small) amount of content, then expected vertical utility from the second must have been low (high) that day. Separate identification of the structural parameters for vertical quality (α_j) and vertical preference (λ_i) comes from a factorization of expected vertical utility into a site-specific and consumer-specific components.

The remaining rows of Table 11 summarize the estimation parameters related to υ_i, λ_i, and γ_i. Identification of these depends on differences in the distribution of the structural parameters that correlate with the observed demographic variables (D_i). Finally, the cost shifter for weekends and holidays (γ_w) is identified from differences in the amount of browsing on these days compared to weekdays.

D.2 Independence of Unobserved Contributions to Utility

To obtain unbiased estimates of model parameters, we make two assumptions about unobserved components of the utility from visiting a site. First, conditional on the number of links to site j that were previously seen as of step t ($n_{i,j,t}$), the idiosyncratic private shock to utility for site j ($\epsilon_{i,j,d,t}$) is assumed to be independent of site $j's$ daily deviation in match utility ($\nu_{j,d}$): $\epsilon_{i,j,d,t} \perp \nu_{j,d} | n_{i,j,t}$. Second, conditional on $n_{i,j,t}$, $\epsilon_{i,j,d,t}$ is assumed to be independent of the signal value of links from site j to some other (as-of-yet unvisited) site k ($s_{k,j,d}$): $\epsilon_{i,j,d,t} \perp s_{k,j,d} | n_{i,j,t}$. Owing to the nature of information consumption, these assumptions are compatible with our setting as we explain next.

First, a violation of the assumption $\epsilon_{i,j,d,t} \perp \nu_{j,d} | n_{i,j,t}$ can occur if information about $\nu_{j,d}$ (from sources other than inbound links) is obtained prior to visiting site j. Such sources include prior visits to site j, links to site j from previously visited sites that are outside the estimation sample, or word of mouth from other readers (e.g., via email). The data rule out the first of these. Although

Table 11: Summary of Identifying Conditions for Model Parameters

Parameters	Identifying Variation	Key Assumptions
Structural		
$\mathbb{E}(\mu_{i,1})$, γ_i	$a_{i,1}\vert\omega_1$	Additivity in utility function
z_j, υ_i	$\mathbb{E}(\mu_{i,1})$	Functional form for utility from horizontal match
δ	$(a_{i,1}, a_{i,2}, \dots, \omega_1, \omega_2, \dots)$	Exclusion of ω from utility
τ_s/τ_ν ratio	$(a_{i,2}, n_{i,1})\vert\omega_2, w_{i,1}$	Bayesian updating, distributions of ν and s
$\mathbb{E}(\beta_{i,2})$	$(a_{i,2}, w_{i,1})\vert\omega_2, n_{i,1}$	Bayesian updating, distribution of K
α_j, λ_i	$\mathbb{E}(\beta_{i,2})$	Functional form for utility from vertical quality
Estimation		
ϕ_υ	$\upsilon_i\vert D_i$	Distribution of υ_i
ϕ_λ, η_λ, ζ_λ^2	$\lambda_i\vert D_i$	Distribution of λ_i
ϕ_γ, η_γ, ζ_γ^2	$\gamma_i\vert D_i$	Distribution of γ_i
γ_w	$(a_{i,1}, weekend)$	

NOTES: $a_{i,1}$ and $a_{i,2}$ denote which sites are visited first and second within a session; ω_1 and ω_2 denote the average linking frequency from the first and second sites visited to all other sites; $w_{i,1}$ denotes the number of words encountered at the first site visited within the session; $n_{i,1}$ denotes the number of links encountered at the first site visited; $\mathbb{E}(\mu_{i,1})$ is equal to the expected horizontal match utility from the first site visited in a session; and $\mathbb{E}(\beta_{i,2})$ is equal to the expected vertical utility from the second site visited in the session (when visited at step $t = 2$).

the latter two cannot be entirely ruled out, their potential to introduce bias is minimal because we model excerpts as noisy, rather than perfect signals of sites' daily match locations.

Second, for a violation of the assumption $\epsilon_{i,j,d,t} \perp s_{k,j,d}\vert n_{i,j,t}$ to occur, the consumer needs to know something about the excerpt from site j to another site k prior to visiting site j. Until they visit site j and see its content, however, the consumer does not know whether site j links to site k on any particular date (rather, they only know the long-run linking frequencies), nor, if there is a link, do they know what its signal contains (as it pertains to news at site k, which the consumer hasn't seen).

D.3 Parameter Normalizations, Transformations, and Prior Distributions

Here we present the parameter normalizations listed in Section 4.4, transformations of the data-augmented state variables, and the prior distributions of the remaining parameters.

The parameter N (related to news quantity) cannot be separately identified from the individual vertical quality preference parameters, λ_i. For example, doubling the number of bits N and dividing λ_i by two would yield the same choice probabilities. We normalize $N = 30$, reflecting an upper limit of 30 celebrity news items each day.

Daily deviations in match position, $\nu_{j,d}$, and match signals from excerpts, $s_{j,k,d}$, are both latent constructs, and we cannot separately identify their scales. Instead, we set $\tau_\nu = 1$ and interpret τ_s as the ratio of their precisions. Average match locations, z_j, are also latent constructs, and we normalize them with respect to consumer's match preferences, υ_i, by setting the mean of the z_j's to be zero. To avoid a degenerate posterior density for the υ_i's of the type described in Roos and

Shachar (2014), we set the prior intercept and scale of the υ_i's to $\eta_\upsilon = 0$ and $\zeta_\upsilon = 1$, respectively.

Because the state variables for news quantity, K, and average signal value for each site, $\bar{s}$ are unobserved, we use data augmentation (Tanner and Wong 1987; Rossi et al. 2005) to sample these state variables along with the model primitives and then integrate over them numerically. To improve the efficiency of our sampling procedure, we transform the $\bar{s}$'s. First, we define $s^*_{j,\ell,d} \equiv \left(s_{j,\ell,d} - z_j - \nu_{j,d}\right)\tau_s^{-1/2}$, so that $s^*_{j,\ell,d}$ follows a standard normal distribution independent of z_j and $\nu_{j,d}$. Second, we enforce the identifying restrictions $\mathbb{E}\left(s^*_{j,\ell,d}\right) = 0$ and $\mathbb{V}\left(s^*_{j,\ell,d}\right) = 1$ via pairwise sampling of the s^*'s using the method of Musalem et al. (2009). A parallel strategy is used to sample the data augmented $\nu_{j,d}$'s.

The prior distributions for the remaining parameters are:

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

E MCMC Sampling Procedure

Here we provide an overview of our estimation approach (further details can be found in the Online Appendix). We use the method of Imai, Jain, and Ching (2009, hereafter IJC) to sample from the data-augmented posterior distribution of the model parameters. The IJC procedure is based on a standard Metropolis-Hastings (M-H) sampler augmented with a method for calculating the emax function (Equation (18)). Compared to the standard nested fixed point algorithm for approximating the emax function (Aguirregabiria and Mira 2010), IJC's method requires significantly fewer computational resources (see Imai et al. 2009 and Ching et al. 2012 for further discussion of IJC's advantages).

But even though the computational gains from IJC are great, they come at a cost: The procedure can produce sample chains that are highly autocorrelated (compared to the same model without forward-looking consumers). To alleviate this autocorrelation, we use Girolami and Calderhead's (2011) MMALA procedure to construct high-quality proposal distributions for the M-H accept/reject steps in IJC. These proposal distributions have two important qualities: First, the deterministic component of the proposal distribution usually lies in the direction of higher density regions of the parameter space (relative to the current parameter vector). Second, the covariance of the random component is adjusted at each step to approximate the curvature of the posterior distribution. Together, these features greatly improve the rate of convergence and reduce autocorrelation.

To construct the MMALA proposal distribution, one must know the values of the first, sec-

ond, and third partial derivatives of the target log-density function. For single-agent DDC's, these derivatives are not available in convenient closed forms, so we obtain their values through a technique known as automatic differentiation (also referred to as AD; Griewank et al. 1996; Su and Judd 2012). AD is a procedure for automatically augmenting computer code such that while evaluating the value of an arbitrary function $f(x)$, the augmented program also evaluates $f'(x)$, $f''(x)$, etc. by algorithmically applying the chain rule corresponding with the basic operations (addition, multiplication, etc.) comprising the original function. The M-H proposal distributions we construct are based on derivatives of the model posterior distribution while ignoring IJC's numerical approximation to the emax function (in our case, the increased numerical efficiency from performing AD on IJC's approximation to the emax function does not offset the higher computational expense).[14]

Estimation code is written in MATLAB and C++ using the CppAD library for automatic differentiation (Bell 2007), and tested using the method of posterior quantiles (Cook et al. 2006). The testing procedure ensures that the estimation code can recover parameter values based on simulated data.

F Alternative Model Specifications

Posterior estimates for the site parameters for horizontal (z_j) and vertical (α_j) quality are shown in Table 12 for four model specifications. The first three (*no signals*, *myopic*, and *full*), which are presented in the main text, differ in the number of restricted parameters that are estimated. The no signals model assumes consumers are myopic, and that excerpts do not provide information about future horizontal match utility. The myopic model differs from the no signals model by estimating how much links affect consumers' choices of where to browse next. The improvement in fit is significant, as shown in Table 5. The full model differs from the myopic by estimating the discount parameter, thus allowing consumers to anticipate the value of links in their browsing choices.

The fourth specification shown in Table 12 (*larger sample*) is based on the full model, but estimated using a sample for which the inclusion criteria are less restrictive than those described in Section 3.1.2. Specifically, the minimums for the criteria "number of sessions in each month visiting any of our 5 sites" are lowered from 5 to 4; for "total number of sessions visiting any of our 5 sites" from 16 to 12; and for "average number of sessions per month (visiting any site)" from 4 to 3. Using these less restrictive criteria increases the number of panelists from 127 to 155. Although the number of panelists increases by 22%, the number of site visits (i.e., non-zero observations)

[14] Markov chains sampled from the model with myopic consumers using: 1) MMALA proposals, and 2) random walk proposals (with the same target M-H acceptance rate) indicate that the MMALA chain has lag-1, -5, and -50 autocorrelations that are 19%, 36%, and 56% lower, and effective sample sizes that are 13 times higher. In other words, 1/13 of the draws are needed to obtain the same efficiency. Additional detail about the sampling algorithm is provided in Section H of the Online Appendix.

Table 12: Site Parameter Estimates for Alternative Models

	Parameter	Model			
		No signals	Myopic	Full	Larger sample
Model feature					
Links and excerpts			x	x	x
Forward-looking consumers				x	x
Less restrictive inclusion criteria					x
Site					
celebuzz	z_j	0.38	0.57	0.95	0.94
		(0.28, 0.49)	(0.49, 0.64)	(0.85, 1.04)	(0.85, 1.02)
	α_j	0.034	0.032	0.03	0.031
		(0.026, 0.042)	(0.028, 0.036)	(0.025, 0.036)	(0.026, 0.038)
dlisted	z_j	-0.65	-0.63	-0.89	-0.8
		(-0.71, -0.58)	(-0.72, -0.56)	(-1.01, -0.69)	(-0.93, -0.70)
	α_j	0.22	0.2	0.22	0.23
		(0.21, 0.23)	(0.20, 0.21)	(0.21, 0.23)	(0.21, 0.26)
egotastic	z_j	-1.83	-1.95	-2.72	-2.69
		(-1.91, -1.77)	(-2.02, -1.88)	(-2.85, -2.60)	(-2.76, -2.62)
	α_j	0.0085	0.0099	0.0098	0.01
		(0.0029, 0.014)	(0.007, 0.013)	(0.0058, 0.015)	(0.0048, 0.017)
perezhilton	z_j	2.31	1.87	2.2	2.13
		(2.26, 2.37)	(1.84, 1.91)	(2.15, 2.25)	(2.09, 2.17)
	α_j	0.077	0.077	0.082	0.084
		(0.07, 0.084)	(0.072, 0.082)	(0.076, 0.088)	(0.078, 0.09)
thesuperficial	z_j	-0.22	0.14	0.45	0.43
		(-0.35, -0.10)	(0.045, 0.23)	(0.31, 0.59)	(0.29, 0.56)
	α_j	0.014	0.014	0.014	0.014
		(0.0082, 0.02)	(0.012, 0.017)	(0.01, 0.018)	(0.0098, 0.02)

increases by only 9%.

Table 12 shows that the site parameters related to vertical quality due to the volume of news published (α_j) are highly consistent across specifications, whereas the horizontal match locations (z_j) change more, while retaining the same general spatial configuration.

Online Appendix

G Simulation Procedure

Here we document the simulation procedure on which the analysis reported in Section 2.6 is based, and provide further detail not reported in the main text. We simulate browsing for two types of consumers—1) myopic, and 2) forward-looking (with a discount rate of $\delta = 1$)—under three types of excerpting—1) no links, 2) links are noisy signals ($\tau_s = .2$), and 3) links are informative signals ($\tau_s = 2$). We simulate 30,000 browsing sessions under each of the six conditions.

We set the consumer's cost to $\gamma = 2$, her match preference to $\upsilon = 2$, and locate both sites at $z = 0$ (thus ensuring equal match utility on average). Site L always links to site R, but not the reverse; hence $\omega_{L,R} = 1$ and $\omega_{R,L} = 0$.

Initiating a browsing session. Figure 5 shows the probability of initiating a browsing session (i.e., visit at least one site on any given day) under the six conditions. Forward-looking consumers are increasingly likely to initiate browsing sessions as links become more informative. When links are especially informative, the anticipated future benefits are even higher because consumers can choose to visit the linked site only when it provides very high match. Myopic consumers, on the other hand, are insensitive to the precision of link signals, since they cannot anticipate the future benefits from seeing excerpts.

Share of sessions starting at the linking site. Because the two sites offer identical match utility in expectation, myopic consumers are equally likely to start their sessions at both sites, as seen in Figure 6. Forward-looking consumers behave the same when there are no links, but as links become more informative, they are increasingly likely to start their sessions at the linking site (L)

Figure 5: Probability of Initiating a Browsing Session

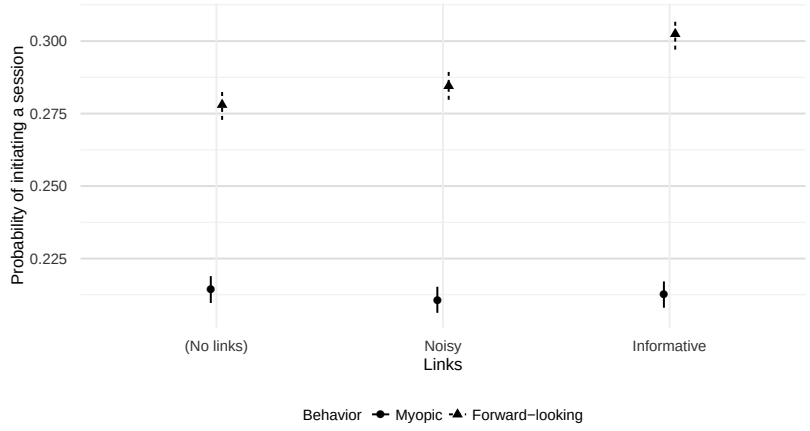


Figure 6: Share of Sessions Starting at Linking Site (L)

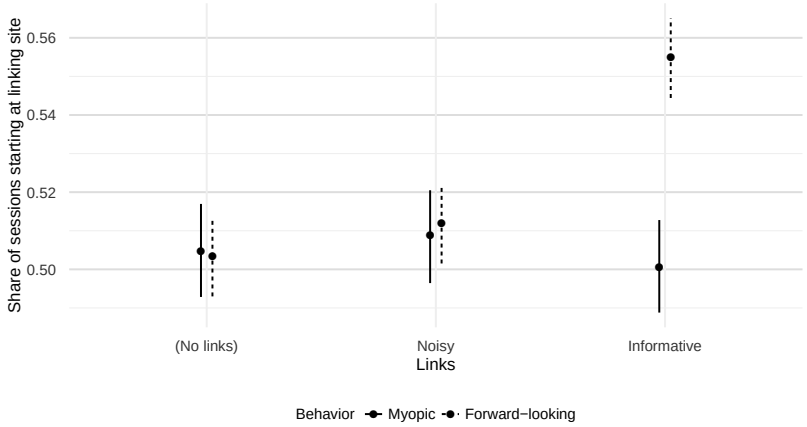
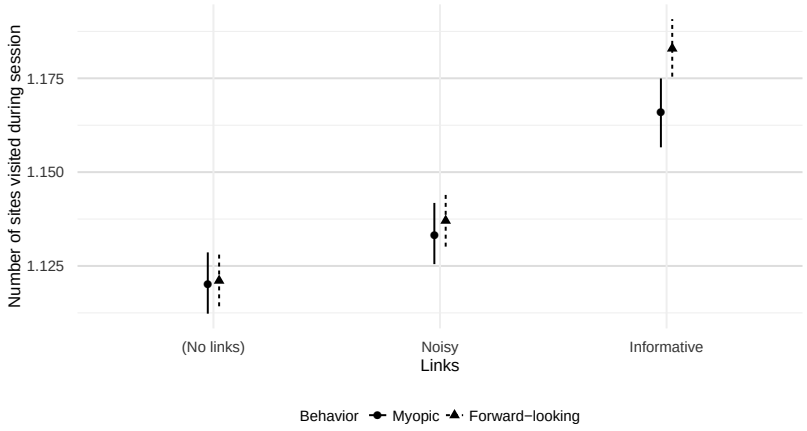


Figure 7: Number of Sites Visited Conditional on Browsing



given the anticipated future benefits from seeing excerpts from site R.

Number of sites visited per session. Figure 7 shows that as links convey more information, both myopic and forward-looking consumers visit more sites (conditional on having initiated a session—i.e., the denominator in this average is the number of sessions in each condition). The increase in session length is due to the consumer being more likely to visit site R after seeing an excerpt at site L. The even greater increase among forward-looking consumers is due to their greater likelihood of initiating their session at site L when links are informative.

Share of sessions visiting the linked site. Figure 8 shows that when links are informative, total traffic at the linked site is higher. The increase in traffic going to site R is highest if consumers are myopic, however, because forward-looking consumers *delay* their visits to the linked site, and sometimes choose to end their session before visiting R.

Figure 8: Share of Sessions Visiting the Linked Site (R)

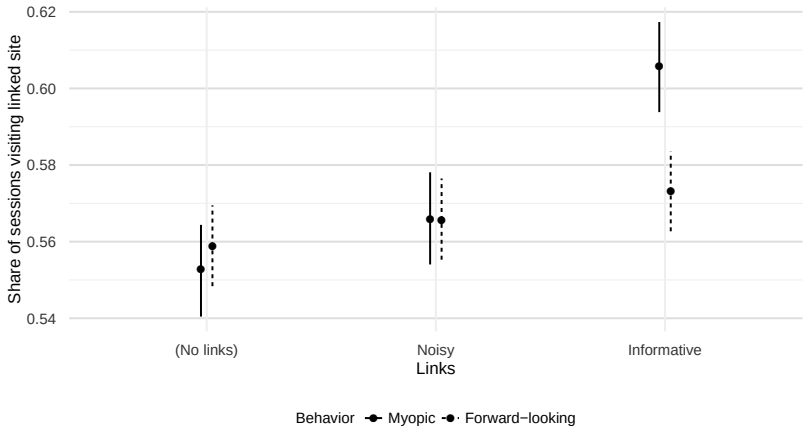
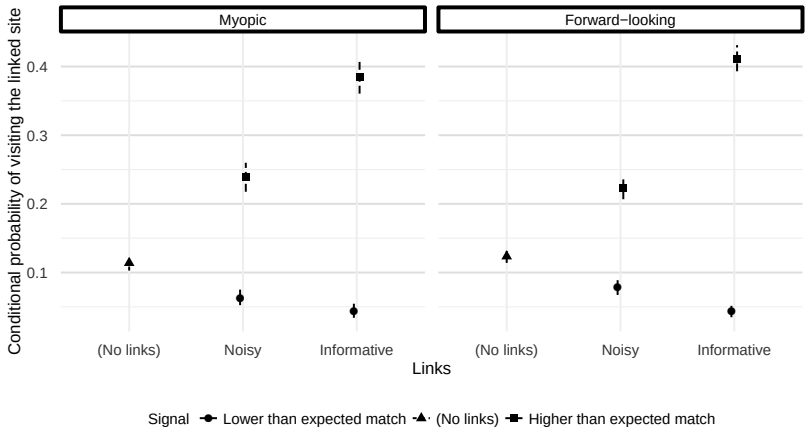


Figure 9: Effect of Signal Valence on Probability of Visiting Linked Site



NOTES. Probabilities are calculated conditional on having chosen to visit the linking site (L) first in the session.

Effect of signal valence. Figure 9 shows the asymmetric effect of match signals on visit probabilities by considering only sessions that begin at site L. When the excerpt at site L signals higher than average match, then the probability of subsequently visiting site R increases. (The amount of the increase is the same for forward-looking and myopic consumers.) Similarly, when the excerpt at L signals lower than average match, then the probability of subsequently visiting site R decreases. The magnitude of the decrease, however, is smaller than the magnitude of the increase because the probability of subsequently visiting R is already low to start with. That is, there is a floor effect limiting the damage that low match signals can inflict on the excerpted site.

H Sampling Algorithm

The general sampling procedure is outlined in Algorithm 1. This algorithm is an application of IJC (Imai et al. 2009), with a few differences. At the lines marked [1] in Algorithm 1, a single procedure calculates both $p(\theta|\mathscr{W})$ and $\mathscr{D}_\theta$ using automatic differentiation. In the MMALA procedure (Girolami and Calderhead 2011), the value of $\mathscr{D}_\theta$ would typically contain derivatives of the log posterior density function. In our setting, however, $\mathscr{D}_\theta$ contains the derivatives of the log posterior function while ignoring the contributions to these derivatives from the IJC emax approximation subroutine. The loss in precision in calculating $\mathscr{D}_\theta$ is compensated for by lower computational burden.

At the lines marked [2] and [3] in Algorithm 1, the function $f(\cdot,\cdot)$ indicates the MMALA proposal distribution described in Girolami and Calderhead (2011). At line [2], the proposal distribution is created conditional on the current parameter vector θ and the derivatives of the log posterior density function evaluated at the point θ, $\mathscr{D}_\theta$. At line [3], the proposal distribution is created conditional on the proposed parameter vector θ^c and the derivatives of the log posterior density function evaluated at the point θ^c, $\mathscr{D}_{\theta^c}$. The proposal distributions are not symmetric, and therefore do not cancel out of the Metropolis-Hastings accept/reject ratio.

Finally, the line marked [4] indicates calculation of a new value function iteration for step t of the IJC sample, as described in (Imai et al. 2009). IJC recommend increasing the efficiency of the sampler by using θ^c to calculate the next approximation of the emax function. Because θ^c has greater distance from θ compared to a random walk sampler owing to the MMALA proposal distribution, we θ to be more efficient when calculate our estimate of the emax function.

References

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

Imai, S., N. Jain, and A. Ching. 2009. "Bayesian estimation of dynamic discrete choice models." *Econometrica* 77 (6): 1865–1899.

Algorithm 1: MCMC sampling procedure. At each iteration, parameters are sampled in blocks. The value function is then iterated and the result either replaces the oldest saved iteration or is appended to the set of saved iterations.

initialize saved MCMC samples: Θ
saved value function iterations: $\mathscr{W}$
foreach *MCMC iteration* t **do**
  foreach *parameter block* $\theta \equiv \theta_b^{(t-1)}$ **do**
    Propose new θ using mMALA proposal distribution:
      calculate marginal posterior probability: $p(\theta|\mathscr{W})$
        derivatives of log posterior probability: $\mathscr{D}_\theta$ // [1]
      set $(m, S) \leftarrow f(\theta, \mathscr{D}_\theta)$ // [2]
      draw $\theta^c \leftarrow N(m, S)$
    Maintain detailed balance:
      calculate $p(\theta^c|\mathscr{W})$ and $\mathscr{D}_{\theta^c}$ // [1]
      set $(m^\circ, S^\circ) \leftarrow f(\theta^c, \mathscr{D}_{\theta^c})$ // [3]
    Accept or reject proposal:
      set $\alpha \leftarrow \frac{p(\theta^c)N(\theta|m^\circ,S^\circ)}{p(\theta)N(\theta^c|m,S)}$
      draw $u \leftarrow U(0,1)$
        if $u < \alpha$ **then set** $\theta_b^{(t)} \leftarrow \theta^c$
      else set $\theta_b^{(t)} \leftarrow \theta$
  Iterate value function using IJC:
    draw $I \leftarrow p(I|\theta^{(t)})$
    calculate $\widehat{W} \leftarrow f(I, \theta^{(t)})$ using IJC // [4]
  Save parameters and value function:
    append $\mathscr{W} \leftarrow \{\widehat{W}, I, \theta^{(t)}\}$
    append $\Theta \leftarrow \theta^{(t)}$