NBER WORKING PAPER SERIES

IDEOLOGY AND ONLINE NEWS

Matthew Gentzkow
Jesse M. Shapiro

Working Paper 19675
http://www.nber.org/papers/w19675

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2013

Chapter in preparation for forthcoming NBER volume The Economics of Digitization: An Agenda. We are grateful to the project organizers, Shane Greenstein, Avi Goldfarb, and Catherine Tucker, and to the participants for helpful comments during the preparation of the chapter. We are especially grateful to Ben Jones for an outstanding discussion. We acknowledge financial support from the Alfred P. Sloan Foundation (through the Sloan Research Fellowship), the National Science Foundation, and the Initiative on Global Markets and the George J. Stigler Center for the Study of the Economy and the State at the University of Chicago. Wharton Research Data Services (WRDS) was used in preparing this chapter. This service and the data available thereon constitute valuable intellectual property and trade secrets of WRDS and/or its third-party suppliers. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Ideology and Online News
Matthew Gentzkow and Jesse M. Shapiro
NBER Working Paper No. 19675
November 2013
JEL No. D83,L86

ABSTRACT

News consumption is moving online. If this move fundamentally changes how news is produced and consumed it will have important ramifications for politics. In this chapter we formulate a model of the supply and demand of news online that is motivated by descriptive features of online news consumption. We estimate the demand model using a combination of microdata and aggregate moments from a panel of Internet users. We evaluate the fit of the model to key features of the data and use it to compute the predictions of the supply model. We discuss how such a model can inform debates about the effects of the Internet on political polarization and other outcomes of interest.

Matthew Gentzkow
University of Chicago
Booth School of Business
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
gentzkow@chicagobooth.edu

Jesse M. Shapiro
University of Chicago
Booth School of Business
5807 S. Woodlawn Avenue
Chicago, IL 60637
and NBER
jmshapir@uchicago.edu

1 Introduction

The news media are a fundamental democratic institution. Access to the news affects political participation (Gentzkow et al. 2011) and the portrayal of the news affects how voters vote (Della Vigna and Kaplan 2007). Digital news is still in its infancy, with digital platforms accounting for only 8 percent of time spent consuming news in the US (Edmonds 2013). Yet it seems inevitable that this share will climb as new technologies develop and diffuse. If this march of technology will transform the Fourth Estate, it may thereby transform democratic politics.

Key to understanding how the rise of digital media will affect politics is understanding how it will affect the breadth and depth of sources from which Americans get their news. These effects are theoretically ambiguous (Mullainathan and Shleifer 2005). On the one hand, the Internet enables inexpensive access to a tremendous range of sources. On the other hand, inexpensive customization may permit highly specialized outlets that serve niche tastes and create echo chambers of self-confirming ideological banter (Sunstein 2001).

In this chapter we formulate an estimable economic model of the production and consumption of online news. We estimate the demand side of the model using a combination of microdata and aggregate moments from a panel of Internet users. We evaluate the fit of the model to key features of the data and use it to explore predictions for the supply of news.

Our model is designed to parsimoniously capture important empirical features of online news consumption. In the model, sites are endowed with two attributes: an ideology and an overall quality. Households are likewise endowed with an overall taste for news and with an ideology. Households choose news sites based on the ideological match between the site and the household. News sites face fixed costs of content that depend on quality and possibly on ideology. News site revenue is from advertising, and advertising revenue depends on audience metrics.

We estimate the demand portion of the model using panel microdata on a sample of Internet users from comScore. For each user we observe total visits to a set of five news sites in 2008. For identification we supplement these data with the overall share conservative on each site, as measured through a separate comScore survey. The demand model fits many aggregate moments well, though it predicts more cross-visiting between news outlets than is present in the data.

We then turn to our supply model. We show that the economics of advertising competition may lead to an important incentive to differentiate ideologically. For a benchmark model of advertising competition, we compute (in the spirit of Gentzkow and Shapiro 2010) the extent to which different news sites are close to their optimal ideological position given the positions of other sites.

The model we present in this chapter complements the descriptive analysis in Gentzkow and Shapiro (2011). In that paper, which we describe in more detail later in the chapter, we use data on the size and ideological composition of online news to construct a measure of ideological segregation for the Internet and to compare the Internet to other media and to non-media domains in which political interaction takes place. We find that the extent of ideological segregation online is low both in absolute terms and in comparison to other domains of interaction.

The value of the model is that it permits evaluation of counterfactual changes in tastes or technology that by definition cannot be envisioned by descriptive statistics alone. Although we do not undertake such calculations here, the model could in principle be used to calculate how the configuration of the market and the consumption of news will change as fixed costs fall, or as news domains sub-divide into more specialized or customized outlets. Because the model incorporates the advertising market, it can also confront changes in the online advertising market, and predict how these will change the mix of products on offer.

The model may also provide a window into the underlying motivations of online news consumers. Where our model fails to fit the facts, there is room for additional modeling to more accurately capture the structure of consumer preferences.

The remainder of the chapter is as follows. Section 2 provides background on broader issues surrounding digitization and the consumption of political news. Section 3 summarizes our data and the descriptive evidence in Gentzkow and Shapiro (2011). Section 4 presents our model. Section 5 discusses our estimation strategy and presents our results. Section 6 concludes with an agenda for future work.

2 Digitization and Political News

There is good evidence of rising elite polarization in the US. Roll-call voting records in the US Congress show a widening gap between the parties since the 1970s (McCarty et al. 2006). Though the evidence for a rise in polarization among non-elites is weaker (Fiorina and Abrams 2008), there are important patterns in the data that suggest strengthening party identification among at least some groups of voters (Prior 2013).

A possible explanation for these patterns comes from widening media choice. The rise of cable television, and the subsequent rise of the Internet, proliferate options that may change how citizens obtain the news. Prior (2005) shows that expanding media choice reduces political engagement among those seeking entertainment but increases it among those seeking information. Prior (2013) reviews evidence on other channels by which the media may influence political polarization.

A central theme in the literature on media and polarization is selective exposure. With many choices, it is easier for an individual with a strong ideological predisposition to consume like-minded news. This can reduce the moderating influence of mainstream media and can result in an ideologically pigeonholed society (Sunstein 2001).

The logic for this type of effect is as follows. Imagine the news is differentiated only horizontally and that news outlets are arrayed on a unit line segment from left to right, along which citizens are uniformly distributed. Suppose that there are J news outlets, equally spaced along the line, and each citizen consumes news from the outlet closest to her. In a world with $J = 1$ news outlet, everyone sees the same news, and the news outlet optimally caters to a broad audience. In a world with $J = 2$, those on the extreme right share an outlet with those on the moderate right, and similarly for the left. So, right-wingers see right-wing news, but extreme right-wingers may not get extreme right-wing news. In a world with $J = 3$, those close to the center (right or left) share an outlet and those on the wings get dedicated outlets, though perhaps not yet fringe outlets. And so on. As J rises, news outlets serve narrower audiences, and so presumably serve them with narrower content. Mullainathan and Shleifer (2005) formalize this type of logic in a model with much richer economic forces.

The logic of this prediction is strongest in a model with purely horizontal differentiation and each citizen consuming news from a single outlet. The addition of vertical attributes and the option to visit multiple outlets both complicate the picture. To see why, step away from the news domain and consider another: the market for DVDs. When DVDs were rented via brick-and-mortar shops, catalogs were often limited to the top films of the day. The advent of rental services like Netflix meant that choice expanded tremendously because inventory costs fell by orders of magnitude (Anderson 2006). Obscure films were now widely available.

But expanding choice did not polarize the movie rental market. Data from Quickflix (an Australian DVD-by-mail service) show that those renting movies from the bottom decile by popularity devote only 8 percent of their rentals to movies in that group, and over a third to movies in the top decile by popularity. Subscribers who rent at least one movie from the least popular decile rent more than twice as many movies total as those who rent at least one from the most popular decile (Elberse 2008).

Put differently, those with niche tastes are still highly engaged with mainstream content, a finding that resonates with evidence from other domains such as cable television (Webster and Ksiazek 2012). Some watch ESPN and some watch the Food Network. Both groups meet at CBS.

In Gentzkow and Shapiro (2011) we show that something similar is at work in online news. The Internet makes sites with extreme content available. But the visitors to these sites get the majority of their news elsewhere, and as a result, patterns of viewership are not well approximated by the simple horizontal model that we sketch above.

There are two reasons. First, in the purely horizontal model, an extreme liberal consumes news from the most liberal news outlet and no other outlet. In practice, she might combine reading from a progressive blog with reading of a middle-of-the-road website like cnn.com. Second, in the purely horizontal model, all outlets are equally good. In practice, they are not, and since quality is primarily a fixed cost, quality is highest where the market is largest, which is in the middle of the road. There are websites that spin the news from a Neo-Nazi perspective, but even accounting for the perspective the overall quality and timeliness of their coverage is poor.

In this chapter we will review the evidence in Gentzkow and Shapiro (2011) and complement it with a model that can rationalize patterns of online news consumption. Though we focus on the news, the model we present may also be useful in understanding consumption in other media domains that have undergone transformative increases in product variety.

3 Data and Descriptive Evidence

In this section we describe our data sources and we summarize the descriptive evidence in Gentzkow and Shapiro (2011) regarding the ideological segregation of online news. Portions of this section are excerpted from Gentzkow and Shapiro (2011).

3.1 Data Sources

Our data on Internet news consumption come from comScore. We construct a universe of 119 national political news and opinion websites for which it is possible to measure both the size and ideology of the audience (Gentzkow and Shapiro 2011).

We measure site size using the average daily unique visitors to each site over the twelve months in 2009 from comScore Media Metrix. Media Metrix data come from comScore's panel of over one million US-resident Internet users. Panelists install software on their computers to permit monitoring of their browsing behavior, and comScore uses a passive method to distinguish multiple users of the same machine.

We measure site ideology using data from comScore Plan Metrix. Plan Metrix data come from a survey distributed electronically to approximately 12,000 comScore panelists. The survey asks panelists the question "In terms of your political outlook, do you think of yourself as...? [very conservative / somewhat conservative / middle of the road / somewhat liberal / very liberal]". The average number of daily unique visitors in each category is reported by comScore for each site for each month. We average these figures over the twelve months in 2009.

We also use comScore microdata on the browsing behavior of a subset of panelists obtained from Wharton Research Data Services (WRDS). The data include 50,000-100,000 machines per year and contain the domain name of each site visited.

Relative to the site-level aggregates, the microdata have two important limitations. First, because the comScore microdata are defined at the domain level (e.g., yahoo.com), we cannot distinguish news content on sub-pages of large sites such as aol.com and yahoo.com. Sites such as Yahoo! News and AOL News are therefore excluded from the microdata sample. Second, the microdata do not distinguish between multiple users of the same machine.

In this chapter, we use a subset of the data for structural estimation. We focus on five sites: foxnews.com, nytimes.com, huffingtonpost.com, drudgereport.com, and cnn.com. We use the 2008 comScore microdata panel and we limit to machines that visit the universe of news sites in Gentzkow and Shapiro (2011) no more than 100 times total throughout the year.

3.2 Descriptive Features of Online News Consumption

In Gentzkow and Shapiro (2011) we use data on the news consumption habits of a panel of Internet users to evaluate whether news online constitutes an "echo chamber" in which people hear only their own views. To do this, we measure the ideological segregation of online news using an approach borrowed from the literature on racial segregation.

For each news outlet, we define the *share conservative*: the share of users who report their political outlook as "conservative" among those who report being either "conservative" or "liberal." We then define each individual's *conservative exposure* to be the average share conservative on the outlets she visits. For example, if the only outlet an individual visits is nytimes.com, her exposure is defined as the share conservative on nytimes.com. If she visits both nytimes.com and foxnews.com, her exposure is the average of the conservative shares on these two sites. Next, we define the *isolation index* (White 1986; Cutler et al. 1999) as the difference in the average conservative exposure of conservatives minus the average conservative exposure of liberals. If conservatives only visit foxnews.com and liberals only visit nytimes.com, the isolation index will be equal to 100 percentage points. If both conservatives and liberals get all their news from cnn.com, the two groups will have the same conservative exposure, and the isolation index will be equal to zero.

We find that news consumption online is far from perfectly segregated. The average Internet news consumer's exposure to conservatives is 57 percent. (Excluding self-described moderates, about two-thirds of the US population self-describes as conservative.) The average conservative's exposure is 60.6 percent, similar to a person who gets all her news from usatoday.com. The average liberal's exposure is 53.1 percent, similar to a person who gets all her news from cnn.com. The isolation index for the Internet is 7.5 percentage points, the difference between the average conservative's exposure and the average liberal's exposure.

News consumers with extremely high or low exposure are rare. A consumer who got news exclusively from nytimes.com would have a more liberal news diet than 95 percent of Internet news users, and a consumer who got news exclusively from foxnews.com would have a more conservative news diet than 99 percent of Internet news users.

The isolation index we estimate for the Internet is higher than that of broadcast television news (1.8), cable television news (3.3), magazines (4.7), and local newspapers (4.8), and lower than that of national newspapers (10.4). We estimate that eliminating the Internet would reduce the ideological segregation of news and opinion consumption across all media from 5.1 to 4.1.

Online segregation is somewhat higher than that of a social network where individuals matched randomly within counties (5.9), and lower than that of a network where individuals matched randomly within zipcodes (9.4). It is significantly lower than the segregation of actual networks formed through voluntary associations (14.5), work (16.8), neighborhoods (18.7), or family (24.3). The Internet is also far less segregated than networks of trusted friends (30.3) and political discussants (39.4).

Figure 1 shows the relative segregation of different domains graphically.

4 Model

The facts we describe above suggest the elements of a satisfactory model of the production and consumption of news online.

News outlets differ in two dimensions: a vertical or quality dimension, and a horizontal or ideology dimension. Accounting for quality variation is critical because most online news consumption is concentrated among a very small number of outlets. In Gentzkow and Shapiro (2011) we report that the top 20 sites account for nearly 80 percent of the daily visits to news outlets online. Accounting for variation in ideology is critical because it is an important driver of demand. For example, 78 percent of visitors to drudgereport.com are conservative as against 22 percent for huffingtonpost.com. Accounting for ideology is also important because many of the concerns about the effects of the Internet relate to its effects on the ideological composition of the news diet.

News consumers differ in two ways as well: their overall taste for consuming news online, and their ideology. We have already stressed the importance of modeling ideology. As we document below, there is enormous heterogeneity across households in the amount of online news consumed, suggesting significant heterogeneity in the overall taste for news (or equivalently in the value of the outside option).

News outlets, especially those with no offline presence, primarily compete for advertising revenue. The growing literature on platform competition with multi-homing (Armstrong 2002; Ambrus and Reisinger 2006; Anderson et al. 2010; Athey et al. 2013) shows that two outlets compete in the market for advertising to the extent that they have audience in common. Therefore, a news outlet's revenue will increase to the extent that the outlet garners a greater audience, and especially to the extent that its audience does not overlap with the audience of other outlets. The importance of audience overlap in determining advertising revenue also suggests that a good model should allow for significant multi-homing by consumers.

News outlets face costs of news production. Improving along the vertical or quality dimension requires payment of fixed costs that do not depend on the size of the audience (Berry and Waldfogel 2010). The costs of varying along the horizontal or ideology dimension are less clear; we will think of these costs as negligible for the purposes of discussion.

In this model, only a small number of outlets will want to make large investments in quality (Shaked and Sutton 1987), and those that do will want to appeal to the widest possible audience. This helps explain the dominance of a small number of relatively centrist sites. Moreover, the incentive to operate outlets on the ideological fringe depends greatly on whether fringe outlets attract unique audience. To the extent that their audience mostly overlaps with that of the major sites, they will obtain vanishingly small revenues in a model in which only unique audience pays.

4.1 Model of Demand

4.1.1 Setup and Notation

There is a set of news outlets $\{1,...,J\}$ indexed by j and a set of consumers $\{1,...,I\}$ indexed by i. Each consumer has T_i occasions to consume news online. At each occasion $t \in \{1,...,T_i\}$ each consumer must choose one news outlet. We can think of an occasion as a unit of time—a minute, say—that is small enough so that it is impractical to visit multiple outlets on the same occasion. Let $y_{it} \in \{1,...,J\}$ denote consumer i's choice at occasion t.

Each consumer i has a time-constant ideology τ_i, and ideologies are distributed i.i.d. across consumers with known pdf $\phi\,()$.[1] Each consumer i has a taste for news μ_i, with μ_i distributed i.i.d. $Gamma\,(\theta,\theta)$ conditional on τ_i.

Conditional on τ_i and μ_i, the number of occasions to consume news T_i is distributed across

[1]In estimation we will assume that τ is standard normal. The assumption that the pdf $\phi\,()$ is known is necessary in order to pin down the scale of outlet ideology. In the normal case, an equivalent alternative would be to normalize the scale of outlet ideology and allow the standard deviation of τ to be a model parameter.

consumers as $Pois(\lambda_i)$, where

$$\log(\lambda_i) = \log f(\tau_i) + \log(\mu_i) \tag{1}$$

Conditional on τ_i, this defines a negative binomial count model (Greene 2012, Ch. 18.4.4).[2] We include $f(\tau_i)$ in the arrival probability to capture the possibility that taste for news is correlated with ideology.

A site j is characterized by a quality α_j and an ideology γ_j, where higher values represent higher quality or more right-wing ideology. The utility to consumer i from visiting site j on occasion t is

$$u_{ijt} = \alpha_j - \left(\tau_i - \gamma_j\right)^2 + \varepsilon_{ijt} \tag{2}$$

where ε_{ijt} is a type-I extreme value error, drawn independently across consumers, outlets, and occasions, and independently of μ_i and τ_i. On each occasion, a consumer chooses the site that maximizes her utility:

$$y_{it} = j \iff u_{ijt} \geq u_{ij't} \forall j' \neq j. \tag{3}$$

4.1.2 Choice Probabilities

Let $\pi_j(\tau) \equiv \Pr(y_{it} = j | \tau_i = \tau)$ denote the probability that a household with ideology τ chooses to visit site j on a given occasion, conditional on choosing to consume news. Then:

$$\pi_j(\tau) = \frac{\exp\left(\alpha_j - \left(\tau - \gamma_j\right)^2\right)}{\sum_{j'=1}^{J} \exp\left(\alpha_{j'} - \left(\tau - \gamma_{j'}\right)^2\right)} \tag{4}$$

Let $\pi(\tau) = (\pi_1(\tau), ..., \pi_J(\tau))$ denote the vector of πs.

4.1.3 Likelihood

An econometrician observes the sequence $\{y_{it}\}_{t=1}^{T_i}$ for each consumer i. Let $K_{ij} = \sum_{t=1}^{T_i} 1_{y_{it}=j}$ denote the number of visits to site j made by consumer i. Let $\mathbf{K_i} = \left\{K_{ij}\right\}_{j=1}^{J}$ denote the vector of visit counts for consumer i.

[2] Formally, $T_i | \tau_i \sim NB\left(\theta, \frac{f(\tau_i)}{f(\tau_i)+\theta}\right)$.

Let

$$B(\tau_i, T_i) = \frac{\Gamma(\theta + T_i)}{\Gamma(T_i + 1)\Gamma(\theta)} \left(\frac{f(\tau_i)}{f(\tau_i) + \theta}\right)^{T_i} \left(\frac{\theta}{f(\tau_i) + \theta}\right)^{\theta}$$

denote the negative binomial probability that a household with ideology τ_i has T_i occasions to consume news.

Let $Multinomial(\mathbf{K_i}, T_i, \pi(\tau_i))$ denote the probability of visit counts $\mathbf{K_i}$ given T_i occasions and ideology τ_i.

The conditional likelihood for household i given ideology τ_i is then

$$L(T_i, \mathbf{K_i}|\tau_i) = B(\tau_i, T_i)\, Multinomial(\mathbf{K_i}, T_i, \pi(\tau_i))$$

The unconditional likelihood for household i is

$$L(T_i, \mathbf{K_i}) = \int_{-\infty}^{\infty} L(T_i, \mathbf{K_i}|\tau_i)\, \phi(\tau_i)\, d\tau_i. \tag{5}$$

The unconditional log likelihood of the data is

$$\ln(L) = \sum_{i=1}^{I} \ln L(T_i, \mathbf{K_i}). \tag{6}$$

Here we make explicit the dependence on T_i just for emphasis; T_i is just the sum of the elements of the vector $\mathbf{K_i}$.

The parameters of the likelihood, which we have suppressed in the notation above, are θ, $\{\alpha_j, \gamma_j\}_{j=1}^{J}$, and any parameters of the function $f()$.

4.1.4 Constraints

Let $c_i = 1_{\tau_i > \tau_0}$ be an indicator for whether a household reports being conservative, where τ_0 is a cutoff.

With some abuse of notation, let

$$c_j = \frac{\sum_{i=1}^{I} c_i K_{ij}}{\sum_{i=1}^{I} K_{ij}} \tag{7}$$

denote the share of visitors to site j who are conservative.

The econometrician observes $\{c_j\}_{j=1}^{J}$. The econometrician can therefore impose the following

J constraints:

$$c_j = \frac{\int_{\tau_0}^{\infty} \pi_j(\tau) f(\tau) \phi(\tau) d\tau}{\int_{-\infty}^{\infty} \pi_j(\tau) f(\tau) \phi(\tau) d\tau}. \tag{8}$$

These constraints are necessary to identify τ_0 and the γ_j's in a sample of households whose ideology is unknown.

4.2 Model of Supply of Online News

4.2.1 Setup and Notation

We define several summaries of the number of visits to site j . Let V_j denote the total number of visitors to site j. Let S_j denote the fraction of consumers who visit site j at least once. Let X_j denote the fraction of consumers who visit site j and no other site.

Write the operating profits of outlet j as

$$\Pi_j = a\left(V_j, S_j, X_j\right) - g\left(\alpha_j, \gamma_j\right)$$

where $a\left(V_j, S_j, X_j\right)$ is annual advertising revenue and $g\left(\alpha_j, \gamma_j\right)$ is the annual cost of content production.

The function $a()$ allows for several possible advertising technologies. The case where $a\left(V_j, S_j, X_J\right) = \tilde{a}V_j$ for some constant $\tilde{a}$ corresponds to a constant per-viewer advertising rate. The case where $a\left(V_j, S_j, X_J\right) = \tilde{a}S_j$ exhibits strong diminishing returns to additional impressions to the same viewer on the same site. The case where $a\left(V_j, S_j, X_j\right) = \tilde{a}X_j$ exhibits strong diminishing returns to additional impressions both across and between sites. This last form of diminishing returns is especially interesting in light of the theoretical literature on multi-homing (Armstrong 2002; Ambrus and Reisinger 2006; Anderson et al. 2010; Athey et al. 2013).

The function $g\left(\alpha_j, \gamma_j\right)$ is similarly abstract. A convenient starting point is that $g\left(\alpha_j, \gamma_j\right) = g\left(\alpha_j\right)$ strictly increasing in α_j. Such an assumption implies that it is costly to produce quality but free to locate anywhere on the ideological spectrum for a given quality.

4.2.2 Audience Metrics

Using our demand model it is possible to derive simple expressions for the various audience metrics that we define above.

The number of visits to site j by the average consumer is given by

$$\begin{aligned} V_j &= \int_{-\infty}^{\infty} \sum_{T=0}^{\infty} \pi_j(\tau) T \Pr(T|\tau) \phi(\tau) d\tau \\ &= \int_{-\infty}^{\infty} \pi_j(\tau) f(\tau) \phi(\tau) d\tau. \end{aligned} \tag{9}$$

The derivation uses the fact that $\mathrm{E}(T|\tau) = f(\tau)$.

The share of consumers who ever visit site j is given by

$$\begin{aligned} S_j &= \int_{-\infty}^{\infty} \sum_{T=0}^{\infty} \left(1 - (1-\pi_j(\tau))^T\right) \Pr(T|\tau) \phi(\tau) d\tau \\ &= 1 - \int_{-\infty}^{\infty} \left(\frac{\theta}{f(\tau)\pi_j(\tau)+\theta}\right)^{\theta} \phi(\tau) d\tau. \end{aligned} \tag{10}$$

To derive the second expression from the first, observe that

$$\sum_{T=0}^{\infty} (1-\pi_j(\tau))^T \Pr(T|\tau) = \mathrm{E}_{T|\tau}\left((1-\pi_j(\tau))^T\right) = \mathrm{E}_{T|\tau}\left(\exp\left(T\ln\left(1-\pi_j(\tau)\right)\right)\right) = \left(\frac{\theta}{f(\tau)\pi_j(\tau)+\theta}\right)^{\theta},$$

where the last step follows from the moment generating function of the negative binomial.

The share of consumers who visit site j and no other site is given by

$$\begin{aligned} X_j &= \int_{-\infty}^{\infty} \sum_{T=1}^{\infty} (\pi_j(\tau))^T \Pr(T|\tau) \phi(\tau) d\tau \\ &= \int_{-\infty}^{\infty} \left(\left(\frac{\theta}{f(\tau)(1-\pi_j(\tau))+\theta}\right)^{\theta} - \left(\frac{\theta}{f(\tau)+\theta}\right)^{\theta}\right) \phi(\tau) d\tau \end{aligned} \tag{11}$$

The derivation here is analogous to that for S_j, but begins by noting that $\sum_{T=1}^{\infty} (\pi_j(\tau))^T \Pr(T|\tau) = \mathrm{E}_{T|\tau}\left((\pi_j(\tau))^T\right) - \Pr(T=0|\tau)$.

4.2.3 Equilibrium Choice of Attributes

Given the set of outlets, we suppose that attributes $\{\alpha_j, \gamma_j\}_{j=1}^{J}$ are a Nash equilibrium of a game in which all outlets simultaneously choose attributes. The first order conditions are that

$$\frac{\partial \Pi_j}{\partial \alpha_j} = \frac{\partial \Pi_j}{\partial \gamma_j} = 0 \forall j. \tag{12}$$

The first order conditions are a useful starting point for empirical work, because the game we have specified will in general have many equilibria. (For example, any set of attributes that constitutes an equilibrium is also an equilibrium under a relabeling of the outlets.)

Coupled with an estimate of demand the first order conditions have substantial empirical content. Consider for example the case in which $\Pi_j = \tilde{a}V_j - g(\alpha_j)$ for some constant $\tilde{a}$. Then the model implies that

$$g'(\alpha_j) = \tilde{a}\frac{\partial V_j}{\partial \alpha_j} \forall j.$$

An estimate of the demand model implies a value for $\frac{\partial V_j}{\partial \alpha_j}$ and the constant $\tilde{a}$ may be approximated from aggregate data. By plotting $g'(\alpha_j)$ against α_j for all outlets j one can trace out the shape of the cost function for quality. The model also implies that

$$0 = \frac{\partial V_j}{\partial \gamma_j} \forall j. \tag{13}$$

That is, since we have assumed that ideology can be chosen freely, each outlet must be at the visit-maximizing ideology. This is a version of Gentzkow and Shapiro's (2010) test for the optimality of print newspapers' choice of slant.

4.2.4 Equilibrium Number of Outlets

If news outlets are substitutes in demand then in general the profits of all outlets will decline in the number of outlets. A natural way to define the equilibrium number of outlets is then the number of outlets such that the next entering outlet would be unprofitable. For such a number to exist there must be a sunk entry cost. Suppose that this cost is uniform across potential entrants. Then the sunk cost can be bounded above by the operating profit of the least profitable outlet and below by the operating profit that the $J+1$st outlet would earn if it were to enter and choose the optimal position given the positions of the existing J outlets.

5 Estimation and Results

5.1 Empirical Strategy and Identification

Our demand estimator solves the following problem:

$$\min_{\tau_0,\theta,f(),\{\alpha_j,\gamma_j\}_{j=1}^{J}} \quad \ln(L) \tag{14}$$

$$s.t. \qquad c_j = \frac{\int_{\tau_0}^{\infty}\pi_j(\tau)f(\tau)\phi(\tau)\,d\tau}{\int_{-\infty}^{\infty}\pi_j(\tau)f(\tau)\phi(\tau)\,d\tau}\forall j. \tag{15}$$

subject to a normalization of the location of the αs and γs.

Our data include panel microdata on individual households, but to develop intuition for model identification it is useful to imagine data that consist only of the shares c_j and the market shares of each site. Consider the problem of identifying τ_0 and $\{\alpha_j,\gamma_j\}_{j=1}^{J}$ taking as given the parameters governing the number of sites visited by each household.

There are J conservative shares c_j and $J-1$ market shares (these must sum to one): $2J-1$ empirical objects that can vary separately.

Up to an appropriate normalization, there are $J-1$ qualities α_j, $J-1$ site ideologies γ_j, and one reporting cutoff τ_0: $2J-1$ parameters.

We assume that $\tau \sim N(0,1)$. We parameterize $f(\tau)=\kappa$ for some constant κ. This allows us to factor the likelihood into two components: the likelihood for the count model of total visits and the likelihood for the logit model of outlet choice. We exploit this factoring to estimate the model via two-step maximum likelihood, first fitting the count model to the total number of visits T_i, then fitting the logit choice model to each household's individual sequence of visits. In the second step we limit attention to consumers who make 15 or fewer visits to the five sites in our sample. Appendix table 1 presents Monte Carlo evidence on the performance of our estimator.

5.2 Demand Estimates

Table 1 presents estimates of model parameters and their standard errors. We normalize γ so that it has a visit-weighted mean of zero. We normalize α so that it is equal to zero for the least visited site. Estimates are in general very precise; this precision is somewhat overstated as we do not incorporate uncertainty in the constraints in equation (15).

We explore several dimensions of model fit.

Figure 2 shows that the negative binomial model provides a good fit to the distribution of total visits across machines in our panel.

Table 2 shows that the model provides a good fit to the overall size and ideological composition of the sites.

Table 3 shows that the model does an adequate job of replicating the distribution of conservative exposure in the data.

Table 4 shows that the model predicts far more cross-visiting than is observed in the data.

5.3 Supply Estimates

We focus on the supply model's implications for sites' choice of ideology. To get a feel for how the model works we begin with the incentives of a hypothetical news site. Consider a world with $J = 2$ and $\alpha_1 = \alpha_2 = 0$. Suppose that site 1 chooses $\gamma_1 = 0$. Should site 2 stick to the center as well or move out to the extremes?

Figure 3 plots our three audience size metrics—average visits V_j, share ever visiting S_j, and share visiting exclusively X_j—as a function of site 2's choice of γ_2. We find that site 2 maximizes visits and the share ever visiting by being centrist. In the case of a site maximizing exclusive visits, it is optimal to be slightly to the right or to the left of the center. Moving away from the center attracts viewers who are not attracted to site 1, and hence who are more likely to visit site 2 exclusively.

Figure 4 explores the incentive to differentiate ideologically in the context of the five sites in our data. We take the αs as given at their estimated values. For each site j, we plot our audience size metrics as a function of γ_j, taking as given the estimated γs for the other sites. The plot also shows the estimated position $\hat{\gamma}_j$ for each site.

Whether a given site would increase its audience by moving closer to or further from the center depends on the audience metric of interest. Most sites would get more households to visit at least once by moving to the center. But most would get more exclusive visitors by moving further from the center. Most sites would also increase total visits by becoming more ideologically extreme.

6 Discussion and Conclusions

We propose a model of the demand and supply of online news designed to capture key descriptive features of the market. We estimate the model on data from a panel of Internet users and explore its fit to consumer behavior. We then study the model's implications for the supply of news.

We stop short of a full equilibrium model of the supply of news, but we believe such a model can be estimated with the primitives we propose. A proposed strategy is as follows. From our demand model, it is possible to calculate how much each outlet would gain in terms of audience from increasing its quality. Using a model of equilibrium advertising rates, one can translate this audience gain into a revenue gain. Conditions for a static equilibrium imply that the gain in revenue must equal the cost of additional content. By performing this exercise for a large set of sites, it is in principle possible to trace out the marginal cost of quality at different points in the quality distribution, and hence to recover the shape of the cost function for quality. A similar exercise could in principle yield a cost function for ideology.

Given cost functions and a notion of equilibrium, the model implies a set of equilibrium positions for news outlets under various assumptions. For example, it would be possible to contemplate changes in the value of online audience to advertisers, or changes in fixed costs or other elements of the news production technology. The model will imply a mapping from these primitives to features of consumer demand such as the extent of ideological segregation.

Stepping further back, it may also be interesting to explore how well the same model can perform in rationalizing patterns of demand in other domains. As we note in section 2, many of the descriptive features of news consumption are reminiscent of other domains such as DVD-by-mail rental patterns. Though the conditions of supply likely differ greatly across domains, common features in demand may suggest a similar underlying model of consumer behavior.

Finally, it is important to note that we focus on the supply and demand for news but not its impact on political beliefs or behavior. As technology evolves it will be important to accumulate theory and evidence on how media platforms change politics.

References

Ambrus, Attila, and Markus Reisinger. 2006. Exclusive vs overlapping viewers in media markets. Harvard University mimeograph.

Anderson, Chris. 2006. *The Long Tail: Why the Future of Business is Selling Less of More*. New York: Hyperion.

Anderson, Simon P., Øystein Foros, and Hans Jarle Kind. 2010. Hotelling competition with multi-purchasing: Time magazine, Newsweek, or both? CESIFO Working Paper Industrial Organization No. 3096.

Armstrong, Mark. 2002. Competition in two-sided markets. Nuffield College mimeograph.

Athey, Susan, Emilio Calvano, and Joshua S. Gans. 2013. The impact of the internet on advertising markets for news media. Rotman School of Management Working Paper No. 2180851.

Berry, Steven, and Joel Waldfogel. 2010. Product quality and market size. *Journal of Industrial Economics* 58(1): 1-31.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. 1999. The rise and decline of the American ghetto. *The Journal of Political Economy* 107(3): 455-506.

Della Vigna, Stefano and Ethan Kaplan. 2007. The Fox News effect: Media bias and voting. *Quarterly Journal of Economics* 122(3): 1187-1234.

Edmonds, Rick. 2013. New research finds 92 percent of time spent on news consumption is still on legacy platforms. *Poynter*. Accessed at <http://www.poynter.org/latest-news/business-news/the-biz-blog/212550/new-research-finds-92-percent-of-news-consumption-is-still-on-legacy-platforms/> on August 1 2013.

Elberse, Anita. 2008. Should you invest in the long tail? *Harvard Business Review* 86(7): 88-96.

Fiorina, Morris P. and Samuel J. Abrams. 2008. Political polarization in the American public. *Annual Review of Political Science* 11: 563-588.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78(1): 35-71.

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

Gentzkow, Matthew, and Jesse M. Shapiro. 2011. Ideological segregation online and offline.

Quarterly Journal of Economics 126(4): 1799-1839.

Greene, William H. 2012. *Econometric Analysis*. New York: Prentice Hall.

McCarty, Nolan, Keith T. Poole and Howard Rosenthal. 2006. *Polarized America: The Dance of Ideology and Unequal Riches.* Cambridge, MA: MIT Press.

Mullainathan, Sendhil, and Andrei Shleifer. 2005. The market for news. *American Economic Review* 95(4): 1031-1053.

Prior, Markus. 2005. News vs. entertainment: How increasing media choice widens gaps in political knowledge and turnout. *American Journal of Political Science* 49(3): 577-592.

Prior, Markus. 2013. Media and political polarization. *Annual Review of Political Science* 16: 101-27.

Shaked, Avner, and John Sutton. 1987. Product differentiation and industrial structure. *The Journal of Industrial Economics* 36(2): 131-146.

Sunstein, Cass R. 2001. *Republic.com.* Princeton, N.J.: Princeton University Press.

Webster, James G. and Thomas B. Ksiazek. 2012. The dynamics of audience fragmentation: Public attention in an age of digital media. *Journal of Communication* 62: 39-56.

White, Michael J. 1986. Segregation and diversity measures in population distribution. *Population Index* 52(2): 198-221.

Table 1: Model parameters

γ	
CNN	-0.0127
	(0.0006)
Drudge Report	0.7229
	(0.0000)
Fox News	0.5320
	(0.0002)
Huffington Post	-0.3645
	(0.0008)
New York Times	-0.2156
	(0.0007)
α	
CNN	4.3252
	(0.0488)
Drudge Report	0.0000
	(0.0000)
Fox News	2.7345
	(0.0475)
Huffington Post	1.8632
	(0.0547)
New York Times	3.6381
	(0.0502)
θ	0.3132
	(0.0000)
κ	3.0259
	(0.0000)
$\Pr(\tau > \tau_0)$	0.5431
	(0.0009)

Notes: The table presents the estimated parameters of the model presented in section 4. Estimates use 2008 comScore data for 5 sites. Estimation is by two-step maximum likelihood, estimating (θ, κ) in the first step and the remaining parameters in the second step. We normalize γ to have a visit-weighted mean of zero across all sites, and α to take value zero for the least-visited site. Asymptotic standard errors are in parentheses.

Table 2: Model fit to size and ideology of news outlets

	Share of total visits		Conservative share of site visits	
	Data	Simulation	Data	Simulation
CNN	0.5297	0.5348	0.5504	0.5604
Drudge Report	0.0113	0.0101	0.9266	0.9270
Fox News	0.1401	0.1339	0.8669	0.8731
Huffington Post	0.0483	0.0488	0.3008	0.3079
New York Times	0.2707	0.2724	0.4027	0.4080

Notes: The table presents, for each site, the share of total visits that each site receives, and the share of visits to each site from conservative consumers, along with analogues from a single simulation at the estimated parameters.

Table 3: Model fit to conservative exposure

Conservative exposure of households visiting at least one site

	Percentile						
	5th	25th	50th	75th	95th	Mean	Standard deviation
Data	0.4027	0.4256	0.5504	0.5504	0.8669	0.5387	0.1360
Simulation	0.4080	0.4842	0.5604	0.5805	0.8213	0.5516	0.1155

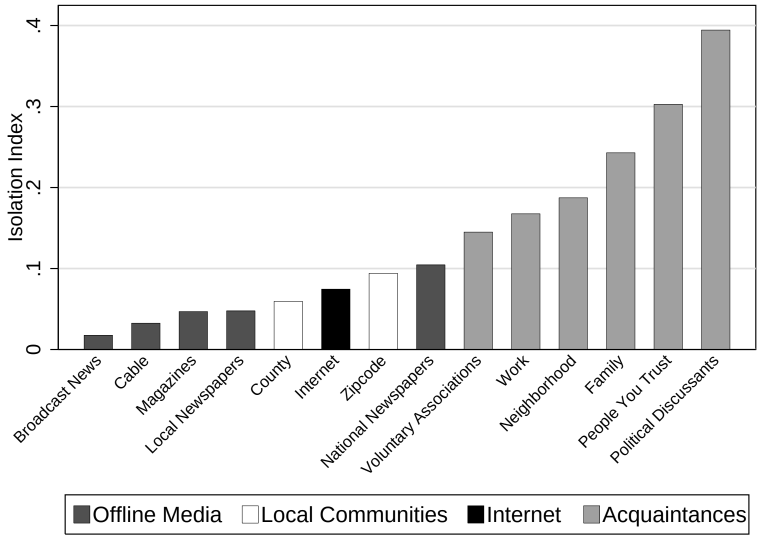
Notes: The table presents statistics of the distribution of conservative exposure in the data and in a single simulation at the estimated model parameters. A consumer's conservative exposure is the visit-weighted average share conservative across the sites visited by the consumer.

Table 4: Model fit to cross-visiting patterns

Share of visitors to site:		Also visiting site:				
		CNN	Drudge Report	Fox News	Huffington Post	New York Times
CNN	Data	.	0.0087	0.1635	0.0711	0.3027
	Simulation	.	0.0406	0.3254	0.1781	0.5667
Drudge Report	Data	0.4131	.	0.2278	0.0656	0.2857
	Simulation	0.8495	.	0.6905	0.1153	0.5133
Fox News	Data	0.4774	0.0140	.	0.0826	0.2996
	Simulation	0.8019	0.0814	.	0.1485	0.5684
Huffington Post	Data	0.4640	0.0090	0.1847	.	0.3556
	Simulation	0.8442	0.0261	0.2857	.	0.7363
New York Times	Data	0.4472	0.0089	0.1516	0.0805	.
	Simulation	0.7896	0.0342	0.3213	0.2164	.

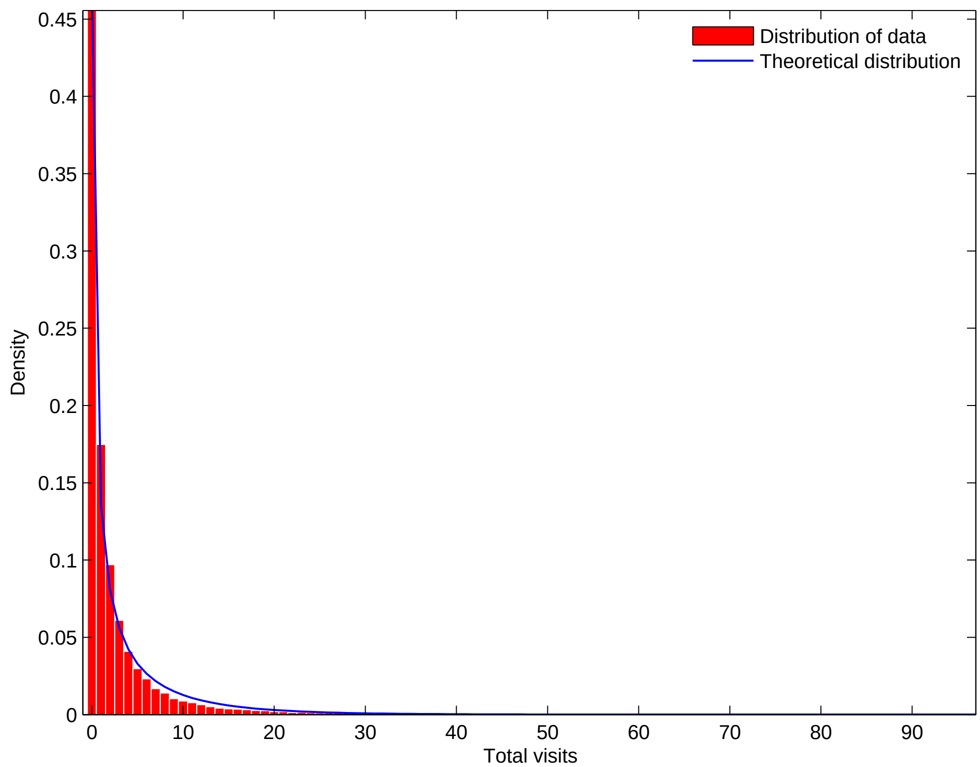
Notes: For each site, the table shows the share of visitors to that site who also visit each of the other sites, both for the empirical data and for a single simulation at the estimated parameters.

Figure 1: Ideological segregation across domains



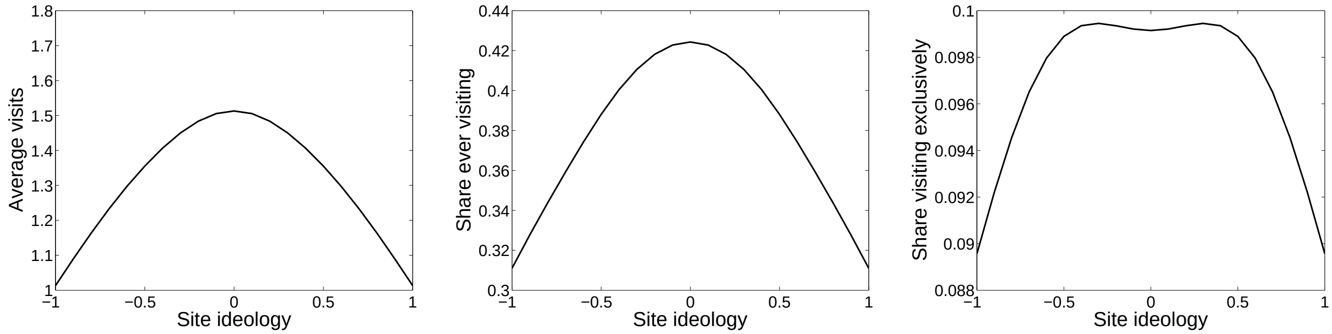
Notes: Internet data are from 2009 comScore Media Metrix and Plan Metrix. County, zipcode, and offline media data are from the 2007 and 2008 Mediamark Research and Intelligence Surveys of the American Consumer. Voluntary associations, work, neighborhood, family, and "people you trust" data are from the 2006 General Social Survey. Political discussants data are from the 1992 Cross-National Election Study. Figure is reprinted from Gentzkow and Shapiro (2011).

Figure 2: Fit of model to total visit counts



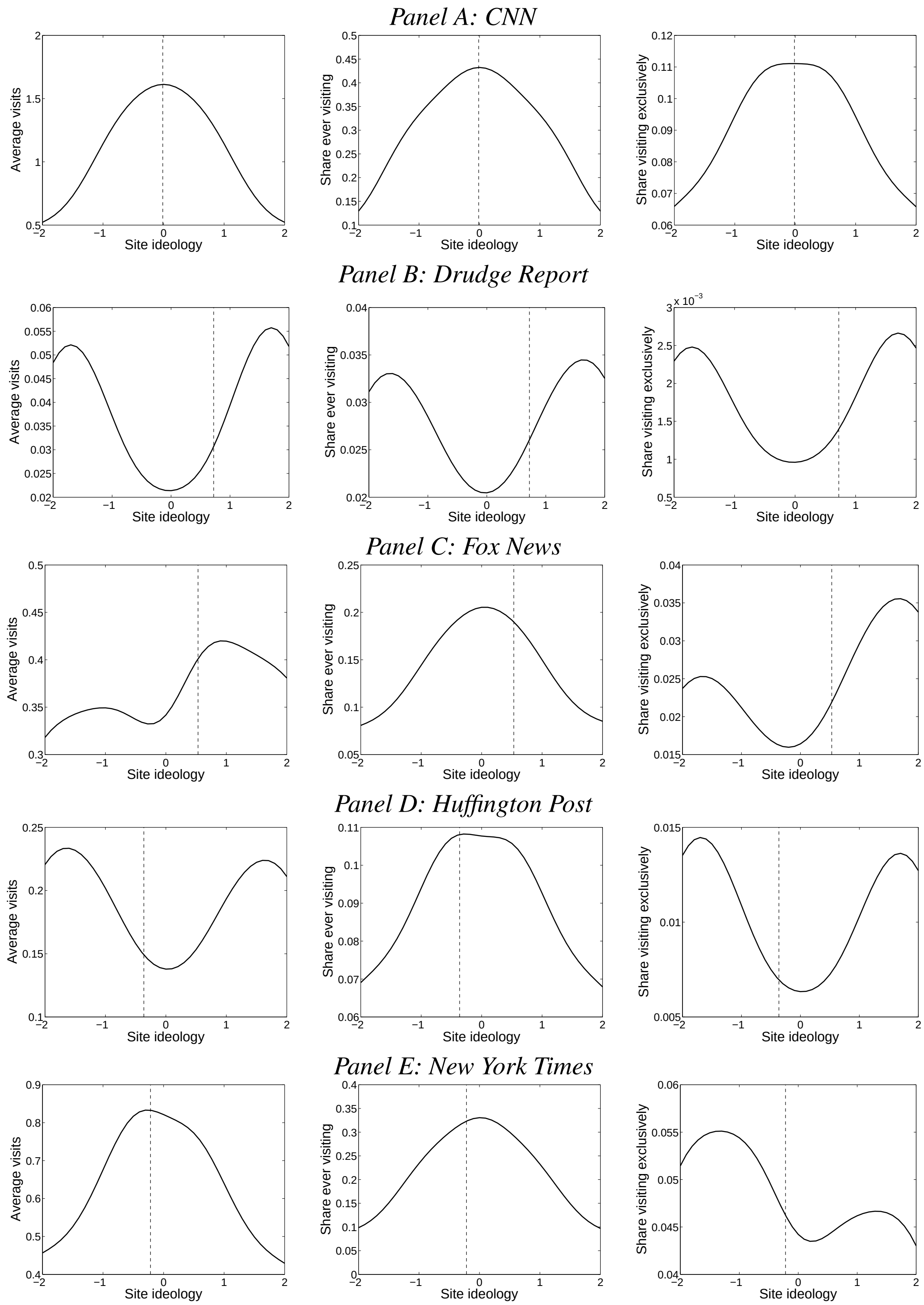
Notes: Plot shows total visits to the 5 sites in our sample in 2008 for each machine in the panel and the density predicted from our estimated model.

Figure 3: Audience size and ideology: hypothetical news site



Notes: The figure shows objects computed from our model using the values of the parameters θ and κ in table 1. In each plot we assume that $J = 2$, that $\alpha_1 = \alpha_2 = 0$ and that $\gamma_1 = 0$, and we plot measures of the size of the audience for outlet $j = 2$ as a function of its ideology γ_2. "Average visits" is the number of visits V_2 made by the average consumer to site 2 across all consumers. "Share ever visiting" is the share of consumers S_2 who visit site 2 at least once. "Share visiting exclusively" is the share of consumers X_2 who visit site 2 and only site 2. See text for formal definitions. Audience size metrics are approximated using Gaussian quadrature.

Figure 4: Audience size and ideology: actual news sites



Notes: The figure shows objects computed from our model using the values of the parameters γ, α, θ and κ in table 1. In each plot we show measures of the size of the audience for outlet j as a function of its ideology γ_j, holding constant all other parameters. "Average visits" is the number of visits V_j made by the average consumer to site j across all consumers. "Share ever visiting" is the share of consumers S_j who visit site j at least once. "Share visiting exclusively" is the share of consumers X_j who visit site j and only site j. See text for formal definitions. Audience size metrics are approximated using Gaussian quadrature. The dashed line indicates the site's estimated ideology $\hat{\gamma}_j$.

Appendix Table 1: Monte Carlo experiments

Parameter	Baseline estimate	Average estimate across simulations	Asymptotic standard errors	Bootstrap standard errors
γ				
CNN	-0.0127	-0.0127	0.0006	0.0000
Drudge Report	0.7229	0.7230	0.0000	0.0003
Fox News	0.5320	0.5321	0.0002	0.0002
Huffington Post	-0.3645	-0.3645	0.0008	0.0001
New York Times	-0.2156	-0.2157	0.0007	0.0001
α				
CNN	4.3252	4.3264	0.0488	0.0267
Drudge Report	0.0000	0.0000	0.0000	0.0000
Fox News	2.7345	2.7389	0.0475	0.0237
Huffington Post	1.8632	1.8663	0.0547	0.0303
New York Times	3.6381	3.6393	0.0502	0.0249
θ	0.3132	0.3132	0.0000	0.0000
κ	3.0259	3.0259	0.0000	0.0000
$\Pr(\tau > \tau_0)$	0.5431	0.5432	0.0009	0.0003

Notes: The table reports the results of Monte Carlo experiments in which we first simulate 10 datasets from our model at the parameter values shown in the first column, then re-estimate our model on each simulated dataset with the starting parameters set at the estimated values.