Can Social Networks Help Content Websites Predict Traffic and Engagement?

Catarina Sismeiro

Imperial College Business School

Ammara Mahmood

Cass Business School

May 2015

Can Social Networks Help Content Websites Predict Traffic and Engagement?

Abstract

Using individual-level clickstreams data, we study if social network information can help predict traffic to a third party content website. The focal content site is a news provider with an online presence independent of that of Facebook (the social network we study). Users can nevertheless register at the site using their social network account and allow access to their social network information. We estimate a flexible individual-level joint model of site visitation and page requests and find that social network information is valuable for traffic prediction. In our study, measures of an individual's activity on Facebook are predictive of that individual's actions at the news site. In addition, knowing what a user's Facebook *friends* do at the content website provides the greatest benefit in prediction: visitors with active friends view fewer articles (a more directed content consumption) although they are more likely to visit the website. Our study also highlights that visitors' past browsing patterns are important predictors of future content consumption, though social network information significantly improves upon these more traditional behavioral metrics.

Introduction

In the past few years, the market has witnessed the development of online social networks and their rise as one of the most influential online forces. As over a billion users engage with sites like Facebook and Twitter, their importance as a medium to connect to and communicate with customers as become clear. Businesses are fast recognizing that all forms of social media provide benefits beyond engagement and advocacy. "Social analytics" rely on ever more data on user footprints and their actions on social networks. Sentiment analysis based on such data is now used to predict the outcome of elections (Golbeck and Hansen, 2011; Tumasjan et. al. 2010), success of movies at the box office (Rui, Lui and Whinston 2013; Asur and Huberman 2010), the marketability of consumer goods (Shimshoni, Efroni and Matias 2009), and even stock performance (Bollen, Mao and Zeng 2011). Practitioners are also using social media data to predict product performance, whether it is through the analysis of customer satisfaction or through gauging the impact of a competitor's marketing campaign (Bradbury 2013).

The power of social networks lies in the influence of connections and the amount of information that users provide about themselves and their preferences. Hence, social networks collect valuable information about an individual and about her "friends." Because network friends could be interested in the same products as the key user (e.g. homophily; McPherson, Smith-Lovin and Cook 2001) social advertising, social targeting, and social customer scoring on platforms like Facebook have great potential (Hill, Provost and Volinsky 2006; Goel and Goldstein 2013). Such strategies rely on peer influence concepts and/or correlated unobservables. Facebook has been using users general profile information such as age, gender occupation for marketers to identify similar Facebook *friends* to target with personalized ads (Kendall and Zhou 2009). These social ads sell at a premium and are a welcome boon for the online advertising industry.

Despite the importance of social network analysis amongst practitioners, there is currently a dearth of academic studies in marketing that illustrate the value of online social network data in targeting and predicting behavior of connected individuals outside the network (Goel and Goldstein 2013). More importantly, the impact on online content consumption has largely been over looked, even though social networks are inherently a platform to share images, news, videos and experiences and content websites serve exactly that to consumers. Hence, social networks might have a direct impact on user engagement with content consumption at other, third party websites. Despite the neglect by researchers, the importance of the online content sector is evident as it represents, directly and indirectly, a significant portion of economic activity online, and supports the flourishing online advertising market with global advertising spend in excess of $100 billion in 2013 (E-marketer 2013).

Among companies whose business models revolve around content provision, news websites are facing some of the most significant challenges. As 46% of social network users discuss news stories (Anderson and Caumont 2014), it comes as no surprise that Facebook is considering hosting news and Snapchat has already introduced a platform for news and content distribution. This growing importance of social networks has heightened fears among news websites, already burdened by decreasing ad revenues. In a recent NYT article, Somaiya, Isaac and Goel (2015) summarize this tension by stating: “Nothing attracts news organizations like Facebook. And nothing makes them more nervous.”

Perhaps one of the main advantages of embracing the online social network wave by content providers is the possibility to collect valuable information that would otherwise not be accessible to them. As users register with content websites using their social network accounts, they give websites access to their personal information. Websites can then use this information to

predict the behavior of users, making this information extremely valuable to content providers. To the best of our knowledge, no previous study attempts to demonstrate whether this information is of value to content providers nor how much value it adds beyond the already known metrics and predictors.

This paper aims to bridge this gap. Using actual clickstreams of users at a major news website, together with data on users' social network activities and the actions of their friends at the online news site, we study if a user's own Facebook activity and the news consumption of her Facebook friends can help predict traffic and engagement with the focal news site. This is a research question that has not been formally explored using clickstream data from an online content provider. We use data from a panel of Facebook users registered with a major news website (a third-party website independent of Facebook) to jointly model website visitation (traffic) and the number of pages viewed over time (engagement) at the news website. Using our individual level model, which accounts for individual heterogeneity and correlated behaviors, we show that social data allows us to make better predictions of traffic and engagement at the news website.

Considering both in-sample and holdout predictions, we demonstrate that the model incorporating Facebook-related information better predicts both site visits and number of page requests (beyond the already good predictions using only individual level own-browsing data at the news site). Demographic and cross-sectional data (e.g., network size) collected from Facebook seems the least informative (perhaps because the models already adequately account for user heterogeneity); instead, information on a user's dynamic activity at Facebook (e.g., liking Facebook pages) improves predictive ability. More importantly, information on friends' behavior (while navigating the news website) seems to provide the most value to content providers. Of the

Facebook-related variables we tested, variables capturing Friends' activities were the ones that provided the most significant improvement in holdout sample predictions.

This is a very revealing result. Information on the behavior of connected friends, the essence of social networks, provides the greatest benefits to third-party websites. Although this result might depend on the type and structure of the website one studies, it does uncover an important insight in the discussion on the value of Facebook data (and data from other online social networks). It is not surprising that demographic information provided the least value. Previously scholars have noted that demographic information is not as effective as behavioral data in explaining or predicting future behavior offline and online (e.g., Bhatnagar and Ghose 2004). What is notable is the finding that friends' behavioral data provided better predictive improvement than the focal user's own data from Facebook. This is perhaps the result of the limited measures of own-Facebook activity. It should be highlighted that the measures available to us in this study are the measures easily accessible to business whose users register using their Facebook accounts. This makes our results of particular managerial significance as business can compile and use this information to improve predictions of traffic and engagement at little cost.

Our results also provide additional insights. We find that past browsing behavior is an important predictor of future browsing behavior at the news website under study. More importantly, our results suggest that online news consumption is a shared experience, as the activity of social network friends is associated with similar behavior by other network members. We find that news websites can benefit from the popularity of social networks and these networks seem to be an important vehicle to drive traffic to news sites. Hence, news consumption and Facebook activity could be complementary activities as friends and their online actions (through for example shared news on their Facebook pages) might be important in driving traffic to content

websites. Our results are purely correlational, but they supplement existing research on news referrals and highlight the importance of peer influence and correlated unobservables (homophily) in news consumption.

Next, we provide an overview of related literature followed by the description of the data used in the study. We will then outline the modeling approach and the performance of alternative specifications. Finally, we present the results and conclude with a discussion of model findings.

Literature Review

The recent rise in popularity of online communities and online social networks has generated an added interest into the area of social interaction and peer influence. Although it is not easy nor immediate to identify and measure online peer influence and its effects (Aral 2011; Nair, Manchanda, and Bhatia 2010), research on the influence and interaction of individuals in various social contexts online is now well established (Watts and Dodds 2007). Previous research has documented such type of influence in product adoption and diffusion, and even in the formation of product ratings (e.g., Van den Bulte and Lilien 2001; Chevalier and Mayzlin 2006; Godes and Silva 2009; Moe and Trusov 2011). Few recent studies in the marketing literature have also used online social network data to identify the role of influential users (e.g., Trusov, Bodapati and Bucklin 2010) and the impact of advertising via social networks (Tucker 2011).

Establishing the existence of social peer influence per se (as these previous studies do) does not provide an answer to the question of whether social network data can improve targeting and prediction of individual level behavior (Goel and Goldstein 2013). Demonstrating the possibility of peer-influence provides only a limited indication of the value of social media data. On one hand, peer-to-peer transmissions can be so rare as to have little practical value. Goel and Goldstein

(2013) note that in many domains of online diffusion, only a small percentage of adopters are influenced by a peer. On the other hand, it is possible that such information, although predictive, provides no extra value beyond that of data marketers already use (e.g., previous browsing experience is well-documented to be predictive of future online browsing and purchase; see for example the work of Park and Fader 2004 and Bucklin and Sismeiro 2003).

Some authors also contend that many existing studies of social influence present results that confound latent homophily with influence (Shalizi and Thomas 2011). This critique is not relevant from the point of view of prediction: homophily could in fact justify the use of social media data as a predictive variable. It is plausible for the behavior of individuals in a network to move in tandem, despite their being no causal relation between connected individuals. Research in sociology has indicated that social connections in their own right could forecast behavior due to the existence of homophily (McPherson Smith-Lovin and Cook 2001; Lazarsfeld and Merton 1954). Liu and Tang (2011) go further to claim that the effectiveness of prediction of social media data depends on the degree of homophily: categories with a strong homophily effect are more likely to benefit from social data, but the degree of improvement depends on the amount of behavioral information one already has on the targeted users

Our motivation is not to determine influence (and separate this influence from homophily) nor do we intend to study the structure of the network. We care about the prediction of behavior outside the social network using social network data. Whether or not information from social networks is valuable for prediction and targeting does not depend strictly on the true existence of influence and causal effects. Some work in marketing literature attempts to look at this problem at the aggregate level and study how aggregate measures of social network activity correlate to aggregate measures of performance. Such studies have developed mostly in the context of movies

(Dellarocas et al 2007), games (Zhu and Zhang 2010), microlending (Stephen and Galak 2012), and books (Chevalier and Mayzlin 2006).

Researcher using aggregate level data can also be found in other fields including finance (Bollen, Mao and Zeng 2011) and political science (Golbeck and Hansen 2011; Tumasjan et. al. 2010). Interestingly, these studies seem to present some conflicting evidence. For example, some studies find that social media volume data and sentiment analysis was able to predict election outcomes, whereas other studies find that that such data have little predictive ability (Yu and Kak 2012; Metaxas, Mustafaraj and Gayo-Avello 2011). Hence, the use of aggregate level models may have limited predictive accuracy or neglect taking advantage of the many nuances in online dynamics. However, individual-level research that links third-party websites and performance metrics is in not as common.

The current lack of individual-level research is probably due to the difficulty in obtaining data on what an individual does while visiting online social networks, and what that individual does outside the network. Much of what individuals do and share is kept private and not visible to researchers outside the individual's network. As a result, many of the existing studies rely on self-reported data or on small-scale experiments conducted in an artificial environment. Rishika et al. (2013) is an exception. In their work, the authors study the effect of customers' participation in a firm's social media efforts on the intensity of the relationship between the firm and its customers. The authors find that a firm's social media efforts lead to an increase in the frequency of customer visits and customer profitability. However, the "social network" information in this study is limited to participation in the firm's social media campaigns. The private actions of social network users are not used to explain the performance of the third party firm. In addition, content websites are not studied.

Another exception is Goel and Goldstein (2013). The authors use the extensive Yahoo! communications network to establish connections between individuals, and use these connections to predict response to advertisements, frequenting an online store, and membership to the Fantasy football league. The researchers assumed individuals to be part of the same network if they had exchanged emails and IM messaging using their Yahoo! accounts. They find that social data was valuable in identifying individuals who would perform specific actions across all instances studied. Bhatt, Chaoji and Parekh (2010) also use data of contacts from an instant message network to predict the adoption voice-over-IP services and find promising results.

One important distinction of this previous work and our study is the type of social data used. These studies use data that is not readily available to most businesses. Other social media data, like the one we investigate, is more readily available to businesses and costs little to obtain. For example, by allowing users to register using their social media accounts, business can access users' profiles. This is the type of social media data we will be studying, not the type of proprietary data used by previous research. In addition, our focus is on networks formed due to exogenous factors and even the result of offline ties. This type of network differs from product-related networks or the networks at a site like Youtube or Amazon also studied in previous research (e.g., Chu and Park 2009).

Finally, another point of departure of our work is the research context. Notably, content websites, or more specifically news websites, are the focus of a limited number of studies (whether at the aggregate or disaggregate level). Lerman and Hogg (2010) argue that the ability to predict traffic and engagement at content sites using the browsing behavior of social network members is of financial relevance as it can lead to better placement of ads and placement of content. Online news websites are currently under tremendous pressure to increase site visitation and engagement,

and to increase ad revenues. Falling online advertising prices, difficulty to charge for content (after years of giving it away for free), reduced readership of their offline arms, competition from a variety of information (and entertainment) sources, a shift in consumers towards more informal sources of information, are all factors that play a role in news websites' demise.

Despite the lack of research, online content is a context in which social media might play a significant role. Content websites compete directly with social media for the time of online users, and, as a result, have long faced a troubled relationship with online social media. News websites and Facebook (one of the most popular social networks in the world) are a good example of how these competing effects can be significant as they both provide information and entertainment to users. Online social networks are an ideal place for users to share information and links to other websites. Because of such sharing, online social network users could be more likely to visit and interact with third party content websites. This "promotional" hypothesis would be in contrast to the displacement effects typically feared by online content providers.

Irrespective of the type of relationship between content websites and social networks (whether they compete for the time of users or, instead, provide synergetic effects), it is vital for content providers to determine if social media data can help predict what their own users will do and potentially aid with targeting. Previous studies on the link between social network and content consumption are mostly survey based (e.g., Bernoff and Li 2011) and often provide contradicting results depending on the methodology used and data employed (e.g., De Waal and Schoenbach 2010; Dimmick, Chen, and Li 2004; Lee and Leung 2008; Nguyen 2010; Tewksbury 2003). Studies have shown that public endorsement of content serves as a site navigation tool and affects readers' attitudes towards content (see Hallahan 1999, and Johnson and Kaye 2004), and recommendations for online news (including 'most read' and 'most emailed' stories or the

inclusion in Google News) affect individual patterns of news consumption even at external websites (e.g., Thorson 2008; Jeon and Esfahani 2012). However, none of these studies analyze actual behavioral data from online social networks.

To the best of our knowledge, the relation between social network activity and content consumption through online content websites has not been empirically validated or studied using behavioral data tracked from actual websites. As a result, the value of individual-level social media data in predicting the behavior at third party content websites is yet unclear. We bridge this gap in the literature by developing a model of user engagement at a news site and test whether social network activity data and friend's news consumption behavior have predictive value. We will focus on different types of data that can be collected: static cross-sectional data (e.g., demographic information and metrics like number of friends), dynamic data of the actions of focal users at the social media network, and dynamic data on the actions of friends of the focal user. We predict that information on friend's actions will be extremely valuable. For example, Bagherjeiran and Parekh (2008), find that there is a parallel in the ads considered and clicked by online friends and they propose an ensemble classifier to combine both behavioral and social data to improve targeting of ads. Given that social networks are essentially a platform for the creation and dissemination of content (both user generated and otherwise) it should not be surprising that a strong connection exists between the actions of online friends and that of focal users.

Data

We obtained data from a major European online newspaper and we monitored the activity of a panel of online users who registered with this focal news website using their Facebook accounts (when registering, users gave consent to the news website to view part of their actions while on

Facebook). We collected the browsing activity of these users from March 1st, 2012 to March 31st, 2012. For each visitor we recorded the daily number of visits and pages viewed at the news site, how many times the visitor viewed the home page or not, and the content categories viewed in each day.

To capture a user's browsing experience, we wanted to ensure we had as much of the full stream of visitation for a typical 30-day period. Previous work has identified browsing experience as an important predictor of website navigation (e.g., Bucklin and Sismeiro 2003) and the objective was to allow enough time to initialize variables reflecting previous experience (e.g., inter visit time, lagged variables, etc.) to ensure we could use such variables in our baseline model. To determine an adequate number of days to use as initialization, we analyzed the inter visit times for the entire panel. On average, users return to the website after 2.4 days, less than 2% of the users have an inter-visit time of more than 15 days, and only 0.76% of more than 20 days. Given these statistics, we have decided to keep the first week of the data for initialization (from March 1st to March 8th), and the remaining three weeks of data for model estimation (23 days from March 9th to March 31st). We then kept in our sample all users who visited the website during the first week of March. Of these users, we kept those who returned at least once to the website during the estimation period (we do not observe any action of those who do not return).[1]

We collected this information from the news website servers. We also had access to Facebook friendship information for those users in the panel. We know the number of friends, whether friends are registered with the news website and, if registered, their identification number.

[1] We exclude all users who do only one visit to the website during the entire month. These casual browsers correspond to a small percentage of users (17% of users registered using a Facebook account, and 24% of those registered with an email account), and to an even smaller percentage of pages viewed (0.5% of page views by Facebook users, and 0.8% of the page views by users registered using an email account). The amount of traffic not analyzed is negligible.

This information allowed us to monitor the activity of a visitor's friends while at the focal news site. We included in the panel those individuals with at least one Facebook friend also registered with the news website because our goal is to determine the value of Facebook-related information in predicting user behavior at the news site. Finally, we collected visitors' general profile data (e.g., age and gender), and part of their Facebook activity. Though we do not have access to a visitor's daily posts and comments, we know whether they have liked a Facebook page or website in a given day, and when they liked it (we call this a "page like"). Researchers seldom have access to such detailed information.

Our final dataset comprises the actions of 1,562 site visitors during the month of March.[2] There are 35,926 daily individual observations in our estimation sample that correspond to 15,864 website visits. Table 1 summarizes the browsing behavior over the three-week estimation period. On average visitors viewed five pages per day though the number of page views is highly skewed. Figure 1 presents the distribution of daily page views across Facebook users. As it can be seen from Figure 1 a significant number of visitors only request a single page in a day with some heavy users requesting many more pages. It is very typical of visitors to news websites to request only one page (typically the Home Page of the website) as most visitors simply read news headlines without reading the full articles (shallow readers). We note that the more a user visits the site, and the higher the number of pages viewed, the more ad exposures (i.e., impressions) the website can sell to advertisers. Hence, site visits and page views are key determinants of news websites' revenue.

[2] We also collected from the company's servers overall traffic data on users who registered with the website using their email accounts instead of their Facebook accounts. We add the daily traffic to the website by these registered users to the baseline model to account for temporal variability in news interest due, for example, to special events that might attract interest of users and their friends to the website.

Not all visitors visited the site on a daily basis: some visitors made multiple visits to the website during a single day whereas others took several days to return. On average visitors made about 0.6 visits to the news website in a given day, though there were instances in which visitors visited the site 11 times. The type of navigation and content consumed at the site also differs significantly across visitors. For example, whereas some site visitors visited the home page in search for content, other visitors directly visited news articles without going through the home page (these are likely visitors who were sent links or viewed the recommendations of articles while online). About 90% of the visits to the site included home page views, and on average visitors visited the home page twice during a site visit. This seems to confirm that home pages tend to be the preferred navigation tool for users to find content and the latest news (Mitchell, et. al., 2012b). We further breakdown the page views of visitors by category. The most popular categories were "Local", "Sports" and "TV" news; we grouped the remaining very fragmented categories into the category "Other News."

Based the Facebook information available to us we build two types of Facebook variables: descriptive variables that are static across the estimation period, and activity variables that can change over time. Descriptive variables comprise demographics (age and gender) and variables that measure the total number of friends and the total number of "page likes." Table 2 presents the details on these descriptive variables. The majority of visitors in our sample are male (78%) with an average age of 39 years (this demographic profile maps well with the readership audience of the newspaper). Visitors to the site had on average 424 friends, with a minimum of five friends and a maximum of 1,000, and liked an average of 178 Facebook pages.

We also built two focal *activity* variables. First, for each user we monitored whether users liked pages during the month of March 2012. We created a "page like" dummy that takes the value

of one if the user was active liking pages in a given day, and zero otherwise. Second, we built a variable that summarizes the daily actions of a user's *Facebook friends* while visiting the content website. For each visitor included in our panel, we monitored whether their Facebook friends visited the focal news website on a given day. We then built a daily dummy variable called "friend activity" that takes the value of one if at least one friend was active on the focal news website, and zero otherwise.[3] Table 2 provides the summary statistics for these Facebook-related activity variables for the estimation period. On average visitors liked 0.13 new Facebook pages per day (average gap of 2.39 days in between like activities) and friends made an average of 0.15 visits per day to the focal news site. (Table 3 also presents a description of all the key variables tested in our models.)

Modeling Approach

Our objective is to determine if social network information can help predict user behavior at content websites in general, and at a news website in particular. We propose a random effects flexible Poisson Hurdle model to study simultaneously (1) the decision to visit the news website in a given day, and (2) the engagement with the site (measured by the number of pages viewed). These two dependent variables are of significant interest for news websites. About 81.5% of news websites' revenue is derived from advertising (the remaining is generated from subscriptions; Clemons, Gu, and Lang 2002) and for most content websites ad-related revenue opportunities are directly proportional to the traffic and page views a site is able to generate. It is for this reason that

[3] We could not monitor comments nor the specific posts or content sharing. We could not see either when friends interacted with the posts of a specific user. However, our measures of user activity, albeit imperfect, reveals that users were actively engaging with the social network (and choosing to reveal to their friends and community something about themselves). If we can find an effect with such a measure of user actions, it is also likely we could find clearer effects using a better and stronger measure of user activity.

content providers routinely monitor these key variables and wish to build models that can help in predicting site visits and page view decisions.

To help websites in this prediction task, we use a flexible joint model that incorporates not only measures of social network activity and information but also individual level heterogeneity and browsing variables, previously shown as being significant predictors (e.g., Bucklin and Sismeiro 2003). Through our analysis we show the improvement in predictive power beyond what other commonly used variables (not related with social networks) can provide. Although social networks can provide rich information, previous research has demonstrated that if not properly used can lead to less than useful results. For example, (Yu and Kak 2012) show that aggregate level sentiment analysis have mixed results in predicting election outcomes due to methodological issues. Our modeling choice reflects our desire to capture the specific features of visitation and page view data at the individual level to avoid such potential caveats.

Correlated Random Effects Poisson Hurdle Model

Because not all visitors are active every day at the news website, daily count of page views by individual visitors are typically characterized by excessive zeros (more zeros than Poisson models can accommodate) and over dispersion (mean different from its variance, though the Poisson model requires variance and mean to be the same). Excessive zeros and over dispersion in count data is common and if not correctly modeled can lead to incorrect inferences (Ridout, Demétrio, and Hinde 1998). Poisson Hurdle and zero inflated models provide a flexible way to model such data (see Atkins et al. 2012; Hilbe 2011; and Zeileis, Kleiber, and Jackman 2008 for details on zero inflated and Poisson Hurdle models).

We propose a correlated random effects Poisson Hurdle model to jointly study the visit and page view behavior of website visitors. The model has two components: the hurdle component to

study the daily site visitation decision and the truncated count model to study the page view decisions once a visitor is at the site. We allow the two decisions (site visit and page views) to be correlated over time, and we jointly estimate the two components (these could be fit independently for computational flexibility, though we report the results from a joint estimation). We further account for individual differences via a random effects specification.

The correlated random effects specification allows us to capture unobserved heterogeneity across users, the possible non-independence in observations due to their clustering around individuals (Greene 2009), and the non-independence of the two decision processes. To the best of our knowledge, this is the first paper to jointly model site visits and page requests to study online content consumption while accounting for heterogeneity, previous browsing behavior, and social network activity.

We propose a generalized linear mixed model (GLMM) specification to estimate the correlated random effects Poisson Hurdle model (Breslow and Clayton 1993). Such a specification allows an efficient estimation of mixed non-linear models and the adequate handling of individual-level heterogeneity. Following the GLMM framework (Min and Agresti 2005), we define p_{it} as the probability that individual *i* visits the website on day *t* such that:

$$P(Y_{it} > 0) = p_{it} \text{ , and} \quad (1)$$

$$P(Y_{it} = 0) = 1 - p_{it} \text{ .} \quad (2)$$

Following Wang et al. (2007), we adopt a logistic regression specification for the hurdle component of the model and we define the probability of a user visiting the website as:

$$p_{it} = \frac{\exp(V_{it})}{1+\exp(V_{it})}, \quad (3)$$

where V_{it} represents the value associated with individual *i* visiting the news website at time *t*.

Given a site visit, we model the probability that visitor *i* views y_{it} (non-zero) pages on day *t* using a truncated Poisson process, such that:

$$P(Y_{it} = y_{it}) = p_{it} \cdot \frac{\exp(\mu_{it})\mu_{it}/y_{it}!}{1-\exp(-\mu_{it})}. \quad (4)$$

The term μ_{it} is the positive, individual, and time-dependent parameter of the Poisson distribution and corresponds to the mean of the (untruncated) Poisson distribution.

To complete the model specification we model V_{it} and μ_{it} as a function of linear predictors assuming a logit and a log link-function, respectively. The linear predictors comprise: (1) "fixed effects" common across individuals (i.e., the associated parameters are not individual-specific), (2) random effects that vary across individuals but are uncorrelated across the two model components, and (3) correlated random effects. Consider the following linear predictors:

$$V_{it} = \log\left(\frac{p_{it}}{1-p_{it}}\right) = \beta Z_{it} + \theta_i Q_{it} + u_{it}, \text{ and} \quad (5)$$

$$\log(\mu_{it}) = \alpha X_{it} + \omega_i W_{it} + v_{it}, \quad (6)$$

where Z_{it} and X_{it} are vectors of covariates and the terms β and α are the associated parameters, common across individuals; Q_{it} and W_{it} are vectors of covariates and the terms θ_i and ω_i correspond to the associated individual-specific parameters. Note that the variables Z_{it} and X_{it} exert the same effect on all individuals, whereas we allow the variables Q_{it} and W_{it} to elicit an individual-specific response. We further assume that the individual level parameters θ_i and ω_i are normally distributed random effects, such that:

$$\omega_i \sim N(\Delta_\omega, \Sigma_\omega), \quad (7)$$

$$\theta_i \sim N(\Delta_\theta, \Sigma_\theta) \quad (8)$$

Finally, we assume that u_{it} and v_{it} are jointly normally distributed random effects with zero mean and variance-covariance matrix Ω, such that:

$$\begin{bmatrix} u_{it} \\ v_{it} \end{bmatrix} \sim N\,(0, \Omega)\text{ , with} \tag{9}$$

$$\Omega = \begin{pmatrix} \sigma_u^2 & \rho\sigma_v\sigma_u \\ \rho\sigma_v\sigma_u & \sigma_v^2 \end{pmatrix}, \tag{10}$$

where σ_u^2 and σ_v^2 are the variances of u_{it} and v_{it}, respectively; ρ is the correlation of u_{it} and v_{it} ($0 \leq \rho \leq 1$); σ_u and σ_v correspond to the standard deviations of the random effects.

There are several important features of this model to note. First, σ_u^2, the variance of the random factor associated with the value function of the logistic model (website visit decision), is not identified. Following previous work, we set this variance to 1 for identification purposes (Hadfield 2010). Second, with this specification, we allow the two model components, visit and page decisions, to correlate. A positive correlation would mean that an idiosyncratic positive shock that increases an individual's likelihood of site visitation simultaneously increases the expected number of page views requested on that day by that same individual. Similarly, an idiosyncratic positive shock that increases the expected number of page views requested on a given day by an individual, also increases the likelihood of site visitation, on that day and by that same individual.

Finally, beyond the correlation of the two model components, through the unobserved random effects, we can also include a variety of observed effects. For example, we make the individual's visitation and page decisions a function of (1) individual characteristics (to capture observed heterogeneity), (2) the individuals' activity while on the social network site, (3) the individual's previous browsing activity on the news site, (4) the activity of online friends on the news site, and (5) exogenous temporal factors influencing news generation and interest. These are included via the covariates added to each model component (that is via X_{it}, Z_{it}, Q_{it} and W_{it}).

We can then define the log-likelihood as:

$$LL = \sum_{i=1}^{N} \sum_{t=1}^{T} \log \int_{\vartheta} (1 - p_{it})^{1-d_{it}} \left[p_{it}\, g(Y_{it} = y_{it})\right]^{d_{it}} d\vartheta\,, \tag{11}$$

where d_{it} takes the value 1 if visitor *i* visits the website on day *t*, and zero otherwise and *g*(.) represents the zero-truncated Poisson distribution. Note that each individual's contribution to the likelihood is the product of the probability of crossing the hurdle and then selecting y_{it} pages to view (when a visit to the site occurs) times the probability of no visit taking place (when no site visit occurs in a day). Due to the inclusion of individual-specific random effects, we need to integrate over their distribution.

Model Estimation

We use a hierarchical Bayesian approach to estimate the proposed model. We adopted conjugated priors for all parameters whenever possible including an inverse Wishart as prior for variance-covariance matrix of the joint normal distribution of all random-effects including u_{it} and v_{it}. We allow for flexibility in the random effects distribution by allowing the covariance of the error terms to be unstructured. No closed form solutions exist for the integral over the random effects distribution, hence we use the Markov Chain Monte Carlo (MCMC) sampling to generate draws from the posterior densities of model parameters (for a discussion of Bayesian estimation of such correlated random effect Poisson hurdle models using a GLMM specification see Draper 2008). We used the first 25,000 iterations for burn-in and checked for chain convergence using standard methods in Bayesian estimation. We then run an additional 50,000 iterations to compute the posterior distributions of all parameters.

Results

To determine the final model specification we tested for the inclusion of variables sequentially and compared model performance in-sample using the deviance information criterion (DIC). We tested the two parts of the model separately as well as simultaneously and kept only those variables that

improved model fit (Table 3 includes a list of the main variables tested). We also compare the different models in holdout to better gauge the predictive power of Facebook information in predicting individual behavior at a third-party website

Predictive Power of Facebook Data: In-Sample Model Comparison

To test the added predictive power of Facebook-related variables we estimate four alternative specifications of the Poisson Hurdle model: (1) Intercept only, (2) No Facebook (includes individual browsing behavior variables but excludes Facebook-related variables), (3) Own Facebook (adds the activity of users on Facebook), and (4) Full Facebook (also adds the activity of Facebook friends while at the news site).

The "No Facebook" model tests the predictive ability of users own browsing behavior at the news site by including individual-specific browsing variables and their non-linear transformation whenever significant. We also added other variables that could predict individual behavior at the focal news site including "Registered User Browsing Activity." This variable captures the interest in news due to external shocks. Finally, all results we report allow for individual random effects for the intercepts and for all parameters associated with time-varying variables. All models further include day-specific dummies to capture daily effects common across individuals (the parameters associated with these daily effects are not individual specific). We do not present the daily dummies parameters but details are available from the authors upon request.

We also tested static cross-sectional variables that firms can easily obtain from the social network for those users who register using their social network account (see Table 3). These include age, gender, total number of likes, and the size of the social network (total number of friends). These variables could capture observed heterogeneity and help predict traffic and engagement. However, these variables did not improve fit and we do not include them in the final

models (detailed results available from the authors upon request). This is already an important finding: static data from the social network could not improve on the predictive ability of a model that incorporates browsing history and accounts for individual heterogeneity. This is perhaps not a surprising result considering the findings from previous literature that show the limited value of demographic variables in predicting behavior when behavioral data is also available (e.g., Bhatnagar and Ghose 2004).

We also tested for the predictive power of other Facebook related variables not included in Table 3 for brevity (e.g., total number of Facebook likes in March for each individual, average gap between Facebook like activity, total daily page views by Facebook friends, news categories viewed by Facebook friends). These variables are also static Facebook variables and they did not improve fit and, as a result, were not included in the final analysis (again, details available from the authors upon request).

Table 4 reports all model results for the final specification and the in-sample (DIC) fit measures. Of all the models estimated the Full Facebook specification is the best fitting in-sample: the DIC of the Full Facebook model is 103,811.0 compared to 111,508.4 for the Own Facebook model, 111,507.6 for the No Facebook model, and 115,149.9 for the Intercept Only model. This provides a first clear indication of the usefulness of Facebook-related in predicting traffic and news consumption at the focal website.[4]

[4] We further note that jointly estimating the two dependent variables provide better fit and that the two dependent variables are positively (and significantly) correlated. The posterior mean of the covariance between site visit and pages views is 0.68 (with a probability interval of [0.60,0.75]) and the posterior mean of the page views variance is 0.71 (with a probability interval of [0.68,0.73]). This means that an unexpected positive shock that increases the likelihood of visitation will also likely mean an increase in the average number of page views, once a visit takes place (i.e., in a given day a higher likelihood of site visitations is associated with a higher expected number of page views).

Holdout Model Results

A stringent test of the superiority of prediction using Facebook data also requires a holdout-sample model comparison. To compare the models in holdout, we re-estimated the four alternative specifications after removing the last two days of data from the original sample (these correspond to the two last days of March 2012). We then built a holdout sample that included the page views and visit information made by 1,559 site visitors during the days excluded from estimation.[5] For each estimated model, we predicted the number of page views and the likelihood of site visit. Table 5 presents a summary of the holdout model performance when predicting page views.

Comparing the holdout model performance one can see that the "Full Facebook" model (which includes own Facebook activity and the activities of Facebook friends) has the lowest mean squared error (MSE) and outperforms all other models. The proposed final model predicts an average of three page views per visit, which closely corresponds to the actual average number of page views (3.06). Models without any own-Facebook information and models that do not consider friends' activities perform worse out-of-sample and tend to over-predict individual page views. Clearly, incorporating social network data improves the prediction of site visit and page views decisions. Hence, Facebook-related information can be valuable for content websites when modeling individual consumer behavior.

Figure 2 presents the holdout lift charts when predicting site visitation for the "Full Facebook", "No Facebook", "Own Facebook", and the "Intercept Only" model specifications. To create the charts we sorted the holdout observations by predicted visit probabilities. We then took

[5] We removed three visitors with no activity prior to the last two days in March. After their visits to the website during the first week of March (the initialization period) these three visitors returned to the website only on the two last days used for holdout. We can predict page views and site visitation for these users using the population means of the parameters. However, if we do so, the overall performance results do not present any significant variation from the ones presented here which exclude these visitors.

10% of all (holdout) observations with the highest predicted probability and computed the percentage of actual visits associated to these observations. We repeat this procedure for 20% of the observations, 30%, and so on. We then plotted the fraction of visits that each model would have been able to capture at different targeting percentages. As we can see from the graphs, the lift line of the Full Facebook model is always above the others: this is the best model in predicting whether a user will visit the site or not in a given day (considering the holdout sample).

The value of Facebook-related data becomes clearer from the holdout results. Note that demographic information obtained from the Facebook profile of each user was not predictive in-sample and is not included in the holdout model tests. Hence, demographic and cross-sectional variables collected from Facebook (including the total number of friends for each user) were of little value for both visit and page views prediction. This is likely due to the adequate heterogeneity controls included in all models. Considering site visits prediction (Figure 2 lift charts) we see that Facebook friend's information clearly adds predictive power, more than own-Facebook information. The social component of the data (that takes advantage of the social network of each individual) provides the greatest fit improvement. Although own social media activity adds some value to prediction, it is clearly outperformed by the inclusion of friend's actions at the focal news website. Similarly, considering predictions of page views, friends' actions are again the Facebook-related variable that adds most of the predictive power. However, own-social media related variables also provide some noticeable gains in page views predictive ability.

Discussion

Our results show that Facebook related information can have strong predictive power in models of third-party website visitation and content consumption (engagement). The predictive ability of these social variables further improve already strong models that use own-browsing behavior at

the website to predict site visitation and page views. Indeed, we find that past browsing behavior is an important predictor of future browsing behavior at the news website under study. Including time-variant variables that capture previous navigation at the website significantly improves model performance, confirming previous literature findings (e.g., Park and Fader 2004).

For example, individuals who allow a long period to pass in between their visits are less likely to visit the site (the longer users stay away from the website, the less likely those users are to return). In addition, visitors who visit the site repeatedly in a day are more likely to return to the site the following day. This means that once users stay away from the news website for a significant amount of time, they might forget and lose the habit of visiting the website for daily news consumption, which in turn indicates that it is important to remain salient in users' minds and be part of their daily choice for news consumption. Our results also support theories of involvement and selective exposure (see Dutta-Bergman 2004) whereby a high level of involvement in a particular subject area is positively associated with information seeking related to that subject. Visitors who read categories like Sports, Local News, and check TV related content are more likely to return to the site. Knowing that visitors are not shallow visitors, that is, knowing that they are reading specialized content and not simply browsing for headlines while at the home page, helps predict future visits to the site. The results also suggest that Facebook registered users follow the behavior of other readers (not registered using their Facebook account) as the number of page views of other registered users is positively correlated with site visit and content consumption of Facebook users.

However, the main contribution of our work lies in showing that Facebook-related information can help predict the behavior of users at a third-party website (beyond models of own-browsing information). Detailed analysis of the results suggests that when users are active on

Facebook they are also more likely to visit the focal news website (the coefficient of daily like activity in the logit component of the model is positive and its 95% probability interval does not cover zero). This implies that the decision to visit the news website is positively correlated to the user's social network activity on a specific day, perhaps indicating that when users have time to consume news they also have time to connect with their online friends.

To demonstrate the predictive value of this variable, we simulate the impact of Facebook activity. We find that, when someone is active on Facebook on a given day, the median increase of daily visit probability is about 22%.[6] This effect is substantial and demonstrates that collecting information on the actions of individuals while on social networks can help predict the actions of those same individuals on content websites. In addition, considering the Poisson component of the model, our results further suggest that daily Facebook page like activity is also positively correlated with the number of articles read. On days in which users are *active on Facebook*, they are *more likely to visit* the news site (as seen previously) and they *read more articles* once they are at the site. By simulating the effect of the daily like variable, we find that the median number of articles read when a visitor is active on Facebook is 7.15 versus 6.23 when visitors are not active. Hence, in the case of our study, we find that social networks do not necessarily divert traffic from news providers nor do they necessarily reduce engagement. This further suggests that news websites should not view social networks necessarily as a threat. However, further studies that go beyond mere correlation effects (as the ones we measure) are required to understand the true nature of these relationships.

The power of Facebook data is not limited to the observed actions or information of the focal user. The impact of friends' actions on site visitation and page views provides additional

[6] Because this is a dataset characterized by an excessive number of zeros, we report median values for these effects because they represent better the central tendency of the effect.

evidence of how Facebook data can be of value. Holdout sample results suggest that friend's actions result in substantive improvements in predictive power. This is because when Facebook friends are active on the news website, a user is more likely to visit the website (positive effect of friends' actions on the probability of site visitation). By simulating the impact of friends' actions on visit probability we find that a visitor is four times more likely to visit the news website when we observe her friends have also been active at the news site, compared to when they are not (the median change in visitation likelihood is 422%, which corresponds to a median absolute change in visit probabilities of 0.81; see Table 6). It is likely that this positive effect on site visitation is due to article recommendations made by friends and to the links friends add to their timeline as they read news articles. However, it is beyond the scope of this study to identify these relations, although our results do suggest that the potential impact of such actions can be significant.

Even though friends' actions are significant in predicting page views, Contrary to the results of site visitation, friends' actions correlate negatively with page views by the focal user. Visitors tend to request fewer pages from the website when their friends are active compared to when they are inactive (from Table 4 we can see that the impact of friends' actions on the page view model is negative and the 95% probability interval does not include zero). By simulating the impact of friends' actions on page views we find that visitors view about two fewer pages when their friends are active compared to when they are not, which corresponds to a reduction of one third in page views (median change of page views of -2.39; see Table 6). We note that this negative effect is conditional on a visit-taking place. Because the impact of friends' actions on site visitation is positive and strong, the net effect on page views, unconditional on a visit-taking place, could also be positive.

To further explore this possibility we predict page views unconditional on site visitation and simulate what would happen if friends were active on the news website on a given day. Table 6 reports the unconditional median change in page views due to friends' activities. As we can see from the table, when friends are active the significant increase in visit probability is associated with a net positive effect on page views: we find that an increase in friends' activities has a positive net increase in page views (almost two more page views per day). This positive net effect is due to an increase in traffic to the website, despite the more directed navigation once at the site. These effects hint at the possibility of recommendations and posts of friends on Facebook influencing browsing behavior (though our results are only correlational).

Conclusion

The purpose of this research has been to establish whether information obtained from social networks could be a valuable predictor of traffic and engagement with online content websites. In recent years, news websites have pushed for the registration of visitors using a Facebook account instead of the traditional registration with the site using an email or the creation of a username. There are obvious benefits that the access to Facebook accounts provides. The collection of rich personal information from user profiles and their social interactions, which allows the selling of premium-targeted ads, is one of the most important immediate benefits.

Less obvious to content providers is the value of the information they collect from Facebook in predicting user behavior at their website. Can social media information be of value for the prediction of traffic and page requests? This is an important question for content websites whose ad revenues depend directly on the traffic and page views they are able to generate. Indeed, being able to adequately predict user behavior including site visitation and number of page views

is central for the daily operations of these websites. Previous research has relied on survey data to study the behavior of social network users within the network and little research exists that studies the impact of the network, or its predictive power, on the individual's behavior on third party websites. Finally, little research has focused on online content providers and their users as most studies tend to be focusing on movies, games, microlending and electoral outcomes.

We focus on two browsing decisions at a leading news website: (1) a user's decision to visit the content website, and (2) the decision on how many pages to view. The corresponding browsing variables, visit decision and number of page views, have a direct bearing on the revenue generated by content providers. We adopt a flexible modeling approach and jointly model these two variables (visit and page views) using a random coefficients Poisson Hurdle model (estimated using a hierarchical Bayesian approach). We allow for site visit and content consumption decisions to be correlated (visitors who are more likely to visit the website may also read more news and consume more content) and correct for potential temporal variation.

We fit the proposed model on detailed browsing data from a panel of 1,562 Facebook users who registered with a leading newspaper website and have visited the website during the month of March 2013. We show that the predictive ability of our model of content engagement increases when we added social network data into a baseline model that accounts for browsing information. Our results also show that knowledge of the browsing actions of network friends provide the greatest improvement in predictive accuracy to a model that already included own browsing history and individual and time-specific effects. In contrast, and not surprisingly, demographic and cross-sectional information collected from Facebook did not improve predictive accuracy. Own-social media data, on the other hand, improved predictive accuracy but these data added value mostly to the modeling of page views and less so in the prediction of site visits.

Our findings provide valuable insights regarding the dynamics of visitors' content consumption and the interrelation between demand for content and social network activity. For example, based on actual browsing patterns of consumers of online content and their exogenous social network ties, we conclude that visitors are more likely to visit the content website when their friends have also visited the website, suggesting that the sharing of news and news recommendations play an important role in attracting traffic. However, the navigation within the site exhibits a directed search, as users with active friends request fewer pages once they are at the site. The net effect of friends' actions is nevertheless positive.

Our findings have several implications for content managers. Given the importance of ad impressions for website managers, managers can increase traffic flow to the content website by targeting active social network users and making article recommendation easy and seamless (once social network users are online and active, they are also more likely to visit the content website and read more articles). Many websites seem to be adopting this policy by requiring Facebook users to register by default with the website Facebook page and app to be able to read articles recommended (or even just read) by friends. This provides an additional intake of registered users. In addition, many news websites are prominently promoting Facebook registration on their websites.

Despite the relevance of our findings, our research suffers from certain limitations. First, the social media data we considered is limited, as it does not include all possible behaviors of users while using the network. However, we believe that access to more detailed actions (e.g., frequency and content of comments) could provide even stronger results. We further note that the data we used is the data that business have readily available once users use Facebook accounts to register with their websites. Hence, the analysis we provide is realistic and can be easily

implemented by business at little extra cost. Other more detailed information (which might be proprietary) could be significantly more expensive to obtain.

Second, our findings suggest that news consumption and Facebook activity could be complementary. We find that users active on Facebook in a given day are also more likely to visit news websites and consume more content at the news site. In addition, visits to the focal news website seem more likely when a user's Facebook friends also visit the news website. Because we have not properly accounted for homophily we cannot go beyond a mere correlational explanation. More work is required to determine whether news sites and Facebook can indeed be complementary instead of competing.

Finally, in this research we have not considered metrics of network centrality and influence. Future research could also explore how influential users affect the flow of traffic to content providers and how one can identify the most influential users with respect to traffic generation. If certain users who share news stories are more influential in channeling traffic to content sites, then managers could target these focal network members with the latest news updates. Similarly, content sites could encourage greater content consumption by highlighting the stories read by focal network members.

References

Alfò, Marco, and Antonello Maruotti (2010), "Two-part Regression Models for Longitudinal Zero-Inflated Count Data," *Canadian Journal of Statistics*, 38(2), 197–216.

Anderson, Monica and Andrea Caumont (2014), "How Social Media is Reshaping News", *Pew Research Center*.State of the Media News.

Aral, Sinan (2011), "Identifying Social Influence: A Comment on Opinion Leadership and Social Contagion in New Product Diffusion," *Marketing Science*, 30(2), 217–223.

Asur, Sitaram, and Bernardo A. Huberman (2010), "Predicting the Future with Social Media," *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (1),* 492-499).

Atkins, David C., Scott A. Baldwin, Cheng Zheng, Robert J. Gallop, and Clayton Neighbors (2012), "A Tutorial on Count Regression and Zero-Altered Count Models for Longitudinal Substance Use Data," *Psychology of Addictive Behaviors*, 27(1), 166.

Bagherjeiran, A. and R. Parekh (2008), "Combining Behavioural and Social Network Data for Online Advertising," *Proceedings of IEEE International Workshop on Data Mining for Design and Marketing (DMDM'08),* 837–846.

Bernoff, Josh, and C. Li (2011). *Groundswell: Winning in a World Transformed by Social Technologies.* Boston: Harvard Business Review Press.

Bhatt, Rushi, Vineet Chaoji, Rajesh Parekh (2010), "Predicting Product Adoption in Large-Scale Social Networks," *Proceedings of the 19th International Conference on Information and Knowledge Management (CIKM*) .

Bhatnagar, Amit, and Sanjoy Ghose (2004), "A Latent Class Segmentation Analysis of E-Shoppers," *Journal of Business Research*, 57, 758 – 767.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011), "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2(1), 1–8

Bradbury, Danny (2013), "Effective Social Media Analytics", (accessed 14 April), [available at, http://www.theguardian.com/technology/2013/jun/10/effective-social-media-analytics]

Breslow, Norman E., and David G. Clayton (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88 (421), 9–25.

Bucklin, Randy.E. and Catarina Sismeiro (2003), "A Model of Website Browsing Behavior Estimated on Clickstream Data," *Journal of Marketing Research*, 40(3), 249–267.

Chu, Wei, and Seung-Taek Park (2009), "Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models," in *Proceedings of the 18th international conference on World wide web*, 691–700.

De Waal, Ester, and Klaus Schoenbach (2010), "News Sites' Position in the Mediascape: Uses, Evaluations and Media Displacement Effects Over Time," *New Media and Society*, 12(3), 477–496.

Dellarocas, Chrysanthos, Xiaoquan Michael Zhang, and Neveen F. Awad (2007), "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures," *Journal of Interactive Marketing*, 21(4), 23–45.

Draper, David (2008), *Bayesian Multilevel Analysis and MCMC*, in Handbook of Multilevel Analysis, 77–139. Springer.

Dutta-Bergman, Mohan. J. (2004), "Complementarity in Consumption of News Types Across Traditional and New Media," *Journal of Broadcasting & Electronic Media*, 48 (1), 41–60.

E-marketer (2013), "Digital to Account for One in Five Ad Dollars," (accessed July 27, 2013), [available at, http://www.emarketer.com/Article/Digital-Account-One-Five-Ad-Dollars/1009592]

Godes, David, and Jose Silva (2009), "The Dynamics of Online Opinion," *Working Paper, R.H. Smith School of Business, University of Maryland.*

Goel, Sharad, and Daniel G. Goldstein. (2013) "Predicting Individual Behavior with Social Networks." *Marketing Science* 33(1), 82-93.

Golbeck, Jennifer, and Derek Hansen (2011), "Computing Political Preference Among Twitter Followers," *in Proceeding of Human Factors in Comp. Sys.*

Greene, William (2009), "Models for Count Data with Endogenous Participation," *Empirical Economics*, 36, 133–173.

Hadfield, Jarrod D. (2010), "MCMC Methods for Multi–Response Generalised Linear Mixed Models: The MCMCglmm R package," *Journal of Statistical Software*,33(2), 1–22.

Hallahan, Kirk (1999), "No, Virginia, it's Not True What they Say about Publicity's "Implied Third-Party Endorsement" Effect," *Public Relations Review*. 25 (3), 331–350.

Hilbe, Joseph (2011), *Negative Binomial Regression*, Cambridge University Press, 2011.

Hill, Shawndra, Foster Provost, Chris Volinsky (2006), "Network-based Marketing: Identifying Likely Adopters via Consumer Networks," *Statistical Science* 21 (2) 256-276.

Jeon, Doh-Shin, and Nikrooz Nasr Esfahani (2012), "News Aggregators and Competition Among Newspapers in the Internet," *Technical report, NET Institute*.

Johnson, Thomas J., and Barbara K. Kaye (2004), "Wag the Blog: How Reliance on Traditional Media and the Internet Influence Credibility Perceptions of Weblogs Among Blog Users," *Journalism & Mass Communication Quarterly*, 81(3), 622–642.

Kendall, Timothy, and Ding Zhou (2009), "Leveraging Information in a Social Network for Inferential Targeting of Advertisements," *US Patent Application* 12/419,958.

Lee, Paul SN, and Louis Leung (2008), "Assessing the Displacement Effects of the Internet," *Telematics and Informatics*, 25(3), 145–155.

Lerman, Kristina, and Tad Hogg (2010), "Using a Model of Social Dynamics to Predict Popularity of News," *In Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM), May*.

Liu, Kun, and Lei Tang (2011), "Large-Scale Behavioral Targeting with a Social Twist," in Proceedings of the 20th ACM international conference on Information and knowledge management, 1815–1824.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook (2001), "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415–444.

Metaxas, Panagiotis T., Eni Mustafaraj, and Daniel Gayo-Avello (2011), "How (Not) To Predict Elections," *IEEE Third International Conference on Social Computing, Boston (MA)*, 165– 171.

Min, Yongyi, and Alan Agresti (2005), "Random Effect Models for Repeated Measures of Zero-inflated Count Data," *Statistical Modelling* 5(1), 1-19.

Mitchell, Amy, Tom Rosenstiel, and Leah Christian (2012), "State of the media: Mobile Devices and News Consumption: Some Good Signs for Journalism,", *Pew Research Center*.State of the Media News.

Moe, Wendy W., and Michael Trusov (2011), "The Value of Social Dynamics in Online Product Ratings Forums," *Journal of Marketing Research*, 48(3), 444–456.

Nair, Harikesh S., Puneet Manchanda, and Tulikaa Bhatia (2010), "Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders," *Journal of Marketing Research*, 47 (5), 883–895.

Nguyen, An (2010), "Harnessing the Potential of Online News: Suggestions from a Study on the Relationship Between Online News Advantages and its Post-adoption Consequences," *Journalism*, 11(2), 223–241.

Nielsen (2011). "State of the Media: The Social Media Report," *Technical Report*, Nielsen Q3.

Oestreicher-Singer, Gal, and Arun Sundararajan (2012) "The visible hand? Demand Effects of Recommendation Networks in Electronic Markets," *Management Science*, 58(11), 1963–1981.

Park, Young-Hoon, and Peter S. Fader (2004), "Modeling Browsing Behavior at Multiple Websites," *Marketing Science*, 23(3), 280–303.

Ridout, Martin, Clarice GB Demétrio, and John Hinde (1998), "Models for Count Data with Many Zeros," *in Proceedings of the XIXth International Biometric Conference*, 19, 179–192.

Rishika, Rishika, Ashish Kumar, Ramkumar Janakiraman, and Ram Bezawada (2013), "The effect of customers' Social Media Participation on Customer Visit Frequency and Profitability: An Empirical Investigation," *Information Systems Research*," 24(1), 108-127.

Rui, Huaxia, Yizao Liu, and Andrew Whinston (2013), "Whose and What Chatter Matters? The Effect of Tweets on Movie Sales," *Decision Support Systems*, 55(4), 863-870.

Shalizi, Cosma Rohilla and Andrew C. Thomas (2011), "Homophily and Contagion Are Generically Confounded in Observational Social Network Studies," *Sociological Methods Research*, 40(2), 211–239.

Shimshoni, Yair, Niv Efron, and Yossi Matias (2009), " On the Predictability of Search Trends," *Google Research Blog.*

Sismeiro, Catarina. and Randy E. Bucklin (2004), "Modeling Purchase Behavior at an E-commerce Web Site: A Task-Completion Approach," *Journal of Marketing Research*, 41(3), 306–323.

Somaiya, Ravi, Mike Isaac, and Vindu Goel (2015), "Facebook may Host News Sites' Content," *New York Times*, March 23, 2015, (accessed on April 1, 2015) [available at www.nytimes.com/2015/03/24/business/media/facebook-may-host-news-sites-content.html]

Stephen, Andrew T., and Jeff Galak (2012) "The Effects of Traditional and Social Earned Media on Sales: A Study of a Microlending Marketplace," *Journal of Marketing Research*, 49(5), 624–639.

Tewksbury, David (2003), "What do Americans Really Want to Know? Tracking the Behavior of News Readers on the Internet," *Journal of Communication*, 53(4), 694–710.

Thorson, Emily (2008), "Changing Patterns of News Consumption and Participation," *Information, Communication & Society*, 11(4), 473–489.

Trusov, Michael, Anand V. Bodapati, and Randolph E. Bucklin (2010), "Determining Influential Users in Internet Social Networks," *Journal of Marketing Research*, 47(4), 643–658.

Tucker, Catherine (2011), "Social advertising," *Working Paper*, available at SSRN 1975897.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe (2010), "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment," in *Proc. of 4th ICWSM. AAAI Press*, 178–18.

Van den Bulte, Christophe, and Gary L. Lilien (2001), "Medical Innovation Revisited: Social Contagion Versus Marketing Effort," *American Journal of Sociology*, 106(5), 1409–1435.

Yu, Sheng, and Subhash Kak (2012), "A Survey of Prediction Using Social Media," *Arxiv paper*.

Wang, Kui, Kelvin KW Yau, Andy H. Lee, and Geoffrey J. McLachlan (2007), "Two-component Poisson Mixture Regression Modeling of Count Data with Bivariate Random Effects," *Mathematical and Computer Modelling*, 46(11), 1468–1476.

Watts, Duncan J., and Peter Sheridan Dodds (2007), "Influentials, Networks, and Public Opinion Formation," *Journal of Consumer Research*, 34(4), 441–458.

Zeileis, Achim, Christian, and Simon Jackman (2008), "Regression Models for Count Data," *Journal of Statistical Software*, 27(8), 1–25.

Zhu, Feng and Xiaoquan (Michael) Zhang (2010), "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics," *Journal of Marketing*, 74 (March), 133–148.

Table 1: Summary Statistics of Browsing Behavior (Estimation Sample)

	Mean	Standard Deviation	Minimum	Maximum
Page Views	4.98	13.20	0	295
Site Visits	0.59	0.83	0	11
Inter-visit Time (in days)	3.72	4.05	1	28
Home Page Views	1.97	7.49	0	295
Page Views for Local News	0.38	1.84	0	146
Page Views for Sports News	0.37	3.21	0	162
Page Views for TV News	0.69	3.54	0	71
Page Views for Other News	0.40	2.43	0	108

Note: All variables are daily variables.

Table 2: Summary Statistics for Facebook Profiles and Activity

Facebook Profile	**Mean**	**Std. Dev.**	**Min**	**Max**
Gender	0.22	0.41	0	1
Age	38.61	13.17	16	80
Total Number of Likes	178.29	137.93	0	551
Total Number of Friends	424.36	293.45	5	1,000
Facebook Activity (During the Last Three Weeks of March)				
Daily Like Dummy	0.13	0.34	0	1
Daily Friend Activity Dummy	0.15	0.35	0	1

Table 3: Variable Description*

(Variables used in the main model specification)

Browsing-Related Variables	
Page Views	Total number of pages viewed at the news site in given day by each user (it takes the value zero if a user does not visit the website on a specific day)
Lag Page Views	Total number of pages viewed at the news site by each user during the previous day that user visited the focal news site
Home Page Views	Total number of home-page views user requested from the focal news site
Inter Visit Time	Number of days since a user's last visit to the focal news site (this variable measures how many days have passed since a user's last visit to the focal site)
Lag Daily Site Visits	Total number of site visits during a user's last visit to the site (following previous research, a page view is assumed to start a new site visit after the user is idle for at least 30 minutes)
Lag Page Views (by Category)	Number of pages viewed (by category) during a user's last visit to site; we will have four variables, one for each main categories (Local News, Sports News, TV News, and Other News)
Lag Registered User Browsing Activity	Number of pages requested from the news site by users who registered with the site using their email account (this variable does not include the activity of the users in our sample, who registered using their Facebook accounts)
Daily Dummies	Daily dummy variables that take the value one for a given day and zero otherwise (we created 22 daily dummies to account for daily specific effects; the estimation period comprises 23 days)
Facebook-Related Variables	
Friend Activity	Dummy variable that takes the value one if a Facebook friend visited the news site on a specific day, and zero otherwise (i.e., for each user and for each day, if at least one Facebook friend is active on the focal news site, this variable takes the value one; if no friend is active, it takes the value zero)
Like Activity	Dummy variable that takes the value one if a user liked a page on Facebook on a given day, and zero otherwise (i.e., if a user is active liking pages on Facebook in a given day, this indicator variable takes the value of one and zero if the user is not active on Facebook that day)
Age	Age (in years) of the user as the user self-reports on his/her Facebook page
Gender	Dummy variable that takes the value of one if the user reports to be a female on her Facebook page
Number of Likes	Total number of Facebook pages a user liked; these are page likes that are visible on a user's profile page
Number of Friends	Total number of Facebook friends for each user

*We tested for the inclusion of other variables including number of likes of each user during the month of March (only), average gap between Facebook like activity, total daily page views by Facebook friends, news categories viewed by Facebook friends. Variables not reported, or that we do not clearly state as omitted from Table 7 (the results table), did not improved fit and were excluded from the estimation.

Table 4: Comparison of Results (posterior means and 95% probability intervals)

Model Component/Variable	No Facebook	Own Facebook	Full Facebook
Count Model - Page views			
Intercept	0.360*** [0.153, 0.556]	0.333*** [0.140, 0.562]	0.447*** [0.235, 0.661]
Home Page Views	1.439*** [1.362, 1.516]	1.448*** [1.357, 1.520]	1.400*** [1.312, 1.511]
Registered User Browsing Activity	0.008* [-0.001, 0.018]	0.008 [-0.009, 0.011]	0.002 [-0.008, 0.011]
Like Activity		0.124*** [0.078, 0.173]	0.138*** [0.090, 0.185]
Friend Activity			-0.405** [-0.569, -0.229]
Friend Activity*Home Page Views			0.219*** [0.068, 0.377]
Logit Model - Visit Decision			
Intercept	-152.467*** [-191.153, -111.487]	-155.769*** [-195.603, -111.843]	-140.953*** [-189.315, -92.348]
Inter Visit Time	-0.408*** [-0.445, -0.371]	-0.408*** [-0.446, -0.366]	-0.400** [-0.447, -0.352]
Inter Visit Time Squared	0.020*** [0.018, 0.023]	0.020*** [0.018, 0.023]	0.020*** [0.017, 0.023]
Lag Daily Site Visits	0.425*** [0.342, 0.496]	0.426*** [0.337, 0.505]	0.353*** [0.257, 0.444]
Lag Page Views for Local News	0.353*** [0.213, 0.482]	0.354*** [0.229, 0.495]	0.323*** [0.187, 0.483]
Lag Page Views for Sports News	0.278*** [0.149, 0.431]	0.307*** [0.164, 0.473]	0.260*** [0.151, 0.473]
Lag Page Views for TV News	0.223** [0.057, 0.385]	0.220*** [0.048, 0.386]	0.110 [-0.316, 0.083]
Lag Page Views for Other News	0.244*** [0.130, 0.380]	0.230*** [0.113, 0.352]	0.225*** [0.074, 0.361]
Registered User Browsing Activity	7.112*** [5.204, 8.943]	7.276*** [5.216, 9.148]	6.554*** [4.336, 8.890]
Facebook Like Activity		0.357*** [0.239, 0.455]	0.330*** [0.217, 0.456]
Facebook Friend Activity			12.447*** [11.438, 13.942]
DIC	**111,507.6**	**111,458.4**	**103,811.0**

Note: '***' means significance at the 0.1%, '**' at 1% and '*' at the 5% level. In the interest of space, we omit here the daily dummies included (for identification purposes and considering their significance not all dummies are estimated; details available from the authors). Finally, we omit the base model with only individual-specific intercepts in both model components. The DIC for that model is 115,149.9. We are reporting population averages.

Table 5: Out-of-Sample Predictive Performance - Unconditional Page Views

	Mean Predicted Page Views	Mean Squared Error	Normalized Mean Squared Deviation
Base Model	4.52	68.92	2.92%
No Facebook	3.25	50.95	2.51%
Own Facebook	3.27	50.18	2.49%
Full Facebook	3.01	46.14	2.39%

Note: 3.06 is the average observed page views.

Table 6: Median Change in Page Views and Site Visit Due to Friends' Activity

	Absolute Change	Relative Change
Visit Probability	0.81	4.22
Conditional Page Views (given a site visit)	-2.39	-0.33
Unconditional Page Views	1.66	3.37

Figure 1: Distribution of Page views

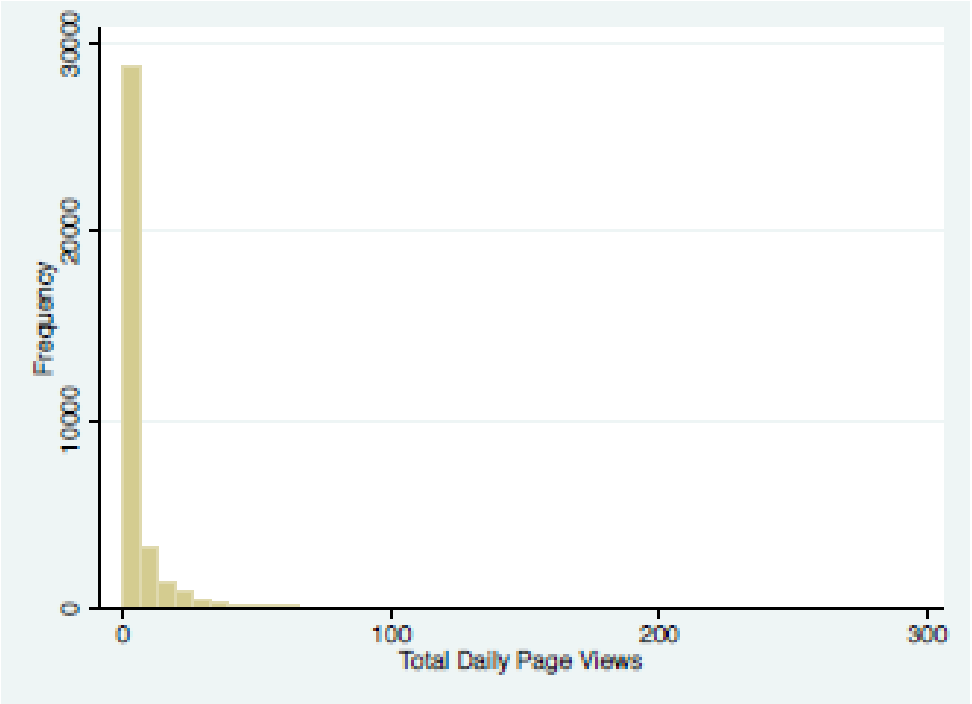

Figure 3: Out-of-Sample Lift Charts