AVERTISSEMENT

LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n° 92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992

http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 1 Capitole

Présentée et soutenue par

Ananya SEN

Le 30 septembre 2016

Essais en économie des médias

Ecole doctorale : **Toulouse Sciences Economiques**

Spécialité : **Sciences Economiques - Toulouse**

Unité de recherche :
UT1-GREMAQ : Département de Mathématiques

Thèse dirigée par
Paul SEABRIGHT Jacques CREMER

Jury

Essays in the Economics of the Media and the Internet

Ananya Sen

September 21, 2016

Abstract

This thesis consists of three indepedent and self contained chapters, all of which have the economics of the media or the internet as the common unifying theme. In Chapter I, we ask whether new technologies change the way political markets work in a democracy? We study the impact of adopting a new technology on campaign contributions received by candidates running for the U.S. Congress. To identify the causal impact of joining Twitter, we compare donations just before and just after politicians open an account in regions with high and low levels of Twitter penetration, controlling for politician-month fixed effects. We estimate that opening a Twitter account amounts to an increase of at least 2-3% in donations per campaign. Moreover, this effect holds only for inexperienced politicians who have never been elected to the Congress before. Placebo checks suggest that this impact is not driven by concurrent increase in information about these politicians in newspapers or blogs, TV ads, or campaign expenditures. The gain from opening a Twitter account is stronger for donations coming from new as opposed to repeat donors, for politicians who tweet more informatively, and for politicians from regions with lower newspaper circulation. Overall, our findings suggest that a new communication technology can lower the barriers to entry in political contests by increasing new politicians' opportunities of informing voters and fund-raising. In Chapter II, we ask if clicks received by news stories online, independent of story quality, influence the way editors allocate resources to them, and if so, how? Using a unique online news dataset from a large Indian English daily newspaper, we provide evidence that editors expand coverage of stories which receive more clicks initially. To establish a causal link between clicks and coverage, we use a novel instrumental variables strategy exploiting rainfall and power outages as exogenous shocks to reader access to online news. We find that the newspaper responds asymmetrically to clicks received by hard and soft news stories, giving additional coverage only to popular hard ones providing evidence for hard news crowding out soft news and not vice-versa. Finally, we relate our results to firm strategy and the challenge firms face in handling 'big data'. In Chapter III, we examine whether a technology, such as the internet, which increases the set of products available to decision makers, may make the decision makers worse off. We build a model where there is product heterogeneity and decision makers can choose to screen products at a cost. In equilibrium, an increase in the choice set can lower a decision maker's payoff by raising the number of products which, on average, are of lower quality than those which were available earlier. An additional product can impose a negative externality on the decision maker by adversely affecting the statistical quality of its existing product pool. We discuss applications to the phenomenon of attention congestion through advances in digital technology.

Acknowledgments

I want to thank my advisor Paul Seabright, who over the course of my PhD has been so generous with his time. It would be in his office where I would first come across the most pointed questions about my research which would allow me to develop my work further. His unbelievable breadth of knowledge across different areas, made me feel assured of getting his invaluable advice on any topic I found interesting. More generally, it was his very broad minded approach to research that I hope to always keep in mind.

Jacques Cremer, my co-advisor, was always a pillar of support. Time spent with him would most often be about addressing the most difficult questions whether it was related to research or about different choices in general. One would always somehow come out of meetings with him feeling calm and relaxed. His advice would often make me look at the bigger research picture rather than getting lost in the narrower issues.

Paul and Jacques, together, provided me the support, guidance and encouragement which I am just very grateful for.

I also want to thank Pinar Yildirim, who I first met at a conference in Toulouse (organized by Paul and Jacques!). Having co-authored two out of my three papers which are part of this dissertation with Pinar, provided me with a great learning experience. Her tips, advice and words of motivation kept me going. Her single minded dedication to research is something that always leaves me amazed.

The past six years spent in Toulouse have been extremely enjoyable. The time spent in this beautiful city would not have been anywhere near the same without the people I have met here. Meeting and spending time with people from different parts of the world has been a truly enriching experience. I have been fortunate enough to meet very nice, honest, genuine people and forge friendships which I value a lot. It was the countless number of lunches and dinners, wine in the evening and night outs, gatherings by the Garonne, and trips here and there which have made the six years what they are. I hope this continues whenever and wherever we meet.

Finally, I want to thank my family for always being there. Without their love and support, I would not be the person I am today. I dedicate this thesis to the memory of my Nanaji, who passed away a few months ago. He is sorely missed.

Chapter I: Social Media and Political Donations: New Technology and Incumbency Advantage in the United States

with Maria Petrova*, Pinar Yildirim †

1 Introduction

In a democratic society, electoral competition and low barriers to entry into politics promote good policies and reduce corruption (e.g., Myerson, 1993; Persson et al., 2003; Besley et al., 2010; Ferraz and Finan, 2011; Galasso and Nannicini, 2011). Low political competition and the associated incumbency advantage often emerge because potential challengers don't have enough opportunities to communicate with voters, which limits fund-raising activities as well as awareness about their candidacy and policy positions (Ansolabehere and Snyder Jr, 2000; Prat, 2002; Strömberg, 2004; Prior, 2006). The persistent advantage enjoyed by experienced politicians is well documented in the United States. Incumbents are reported to achieve re-election rates of around 90% (Levitt and Wolfram, 1997). They also receive higher levels of media coverage and endorsements, creating barriers to entry for new politicians. New communication channels such as Twitter allow politicians to access an alternate, relatively cost-effective channel to inform voters about their policies and raise campaign funds, potentially reducing the incumbency advantage in electoral races. Whether using Twitter actually helps to inform voters and to increase the amount of political donations received is an open question.

In this paper, we study the consequences of politicians' adoption of a new communication technology, namely, opening a Twitter account, on the campaign contributions received while running for the U.S. Congress. We evaluate if adopting Twitter helps politicians to inform voters and increase the financial support received from them. Put differently, we test if the contributions politicians receive from individual donors change before and after joining Twitter, comparing regions with low and high Twitter penetration. We use data which includes 1814 politicians who opened a personal Twitter account between 2009 and 2014, their campaign contributions from Federal Election Commission (FEC), and how Twitter use compares to other sites in the politician's region (i.e., Twitter penetration) information from the comScore online browsing panel.

Identifying the causal impact of Twitter on political donations is not trivial, mainly because there can be a host of correlated unobservables which influence both the politician's decision to join Twitter and the amount

*ICREA Research Professor of Economics, Barcelona IPEG, Universitat Pompeu Fabra, Barcelona GSE, New Economic School, maria.petrova@upf.edu

†Assistant Professor of Marketing, The Wharton School, University of Pennsylvania, e-mail: pyild@wharton.upenn.edu.

of political donations raised. Our estimation uses a difference-in-differences strategy to compare donations a politician received before and after joining Twitter, in states with low and high Twitter penetration. We rely on precise timing in our estimation and control for politician-month fixed effects to account for politician-specific unobserved characteristics, such as being more progressive-minded, more tech-savvy, and, in particular, being at a different stage of campaigning. Our identifying assumption is that the differences between contribution flows, unexplained by politician-month fixed characteristics, would be the same in the absence of Twitter entry in states where Twitter has low and high penetration. Put differently, we rely on a parallel trends assumption, but we do not need to assume that politicians' decision to join Twitter is random or exogenous to their fund-raising activities.

The findings from our analysis suggest that adopting Twitter helps politicians to receive more contributions. Weekly aggregate contributions increase after a politician opens an account and starts blogging on Twitter. However, this gain holds only for new politicians who have never been elected to Congress before, and not for the experienced ones. The aggregate political donations of an average new politician increase by at least \$5,773, which corresponds to 2.3% of all donations under \$3000 raised during the campaign. Associated persuasion rate is close to 1%, which is less than the estimated impact of mass media found in the literature (DellaVigna and Gentzkow 2010), but is close to persuasion rates of direct mailing (Gerber and Green, 2000) or political advertising (Spenkuch and Toniatti 2016). The increase in aggregate donations come mostly from new donors (i.e., those who never donated to this politician before) and not from repeat donors. This suggests that politicians may increase awareness about themselves and their policies via Twitter and gain support from those who did not support them before. An analysis of the Tweet content suggests that the effect is stronger for politicians who tweet more informatively. Additional tests demonstrate that the gains from Twitter are higher for the politicians who come from low newspaper circulation areas. Overall, these findings suggest that political contributions respond to politicians' adoption of Twitter, and new information is a likely channel for that. A broader implication of our study is that adoption of Twitter may reduce the gap in fund-raising opportunities between new and experienced politicians, which, in turn, helps to lower the barriers to entry to national politics.

We use a number of placebo tests to ensure that our identifying assumption is plausible. First, we show that there is no discontinuous increase in campaign spending around the time of Twitter entry, across high and low Twitter penetration areas, despite the fact that political donations are closely related to campaign spending in a given week. Second, to control for possible exogenous events which may coincide with Twitter entry and a discontinuity in funds raised, we show that the media and blog coverage of the politicians does not show a significant change around the time of Twitter entry, again that there are no differences between the high and low Twitter penetration areas. Third, we show that Twitter entry does not coincide with the increase in political advertising shown on TV. Finally, we check that Twitter entry does not differentially affect contribution patterns in states with different income, education, political preferences, and racial composition, therefore it is unlikely that Twitter penetration is just a proxy for one of those variables. Overall, while we cannot test the parallel trends assumption directly, the results in all our placebo specifications are consistent with it.

Our study contributes to several streams of literature. First, our paper highlighting how the advent of

social media in general, and Twitter in particular, can potentially intensify political competition by improving opportunities for new candidates to raise funds and inform voters in a cost-effective fashion. We, therefore, complement the literature which documents the positive impact of political competition and lowering barriers for entering politics on good governance and welfare (Myerson, 1993; Persson et al., 2003; Besley et al., 2010; Ferraz and Finan, 2011; Galasso and Nannicini, 2011). Besley et al. (2010) show that low political competition leads to low economic growth, while Galasso and Nannicini (2011) show that electoral competition is good for political selection. Closely related, Ansolabehere and Snyder Jr (2000), as well as Prat (2002) and Prior (2006), study different sources of incumbency advantage, listing the lack of information of voters about the new candidates and lack of funding opportunities.

Next, we contribute to the literature that studies the role of campaign contributions in political processes. Grossman and Helpman (1996, 2001) argue that campaign contributions allow special interest groups to influence policy outcomes. Similarly, theoretical literature on campaign finance regulation and campaign contribution limits is primarily focused on instrumental motivation for contributions (Prat, 2002; Coate, 2004; Ashworth, 2006; Drazen et al., 2007; Cotton, 2009, 2012; Chamon and Kaplan, 2013). In all these models, campaign contribution limits have different implications depending on whether advertising spending reveal some information about the types of politicians (Prat, 2002; Coate, 2004; Cotton, 2012) or enhance incumbency advantage (Ashworth, 2006). Prat et al. (2010) estimate information benefits from private campaign advertising and find that they are small. Our paper contributes to this literature by highlighting that activity on social media can actually raise donations for inexperienced candidates by providing new information about the politicians.

We also contribute to the emerging literature on the impact of social media on various socioeconomic outcomes. Gong et al. (2015) and Seiler et al. (2016) study the impact of advertising of TV content in Chinese micro blogs on subsequent TV series viewership. Enikolopov, Petrova, and Sonin (2016) study the impact of social media on corporate accountability. Acemoglu et al. (2014) and Enikolopov et al. (2016) analyze the effects of social media content and penetration on subsequent protest participation. Qin et al. (2016) study the content and the impact of social media in China for collective action outcomes, while Qin (2013) look at the relationship between a Chinese microblog penetration on drug quality. In contrast to this literature, we focus our investigation on the strategic benefit of entry into an online social network for the politicians, quantifying their financial gain, and investigating information mechanisms behind the impact of this entry in detail.

Lastly, our paper is also related to literature on the impact of information and communication technologies (ICTs) and traditional media on political preferences and policy outcomes. Recent papers have shown that traditional media has an impact on voting behavior (DellaVigna and Kaplan, 2007; Enikolopov et al., 2011; Gentzkow et al., 2011; Chiang and Knight, 2011; Gentzkow et al., 2014), violence and ethnic tensions (Adena et al., 2014; Vigna et al., 2014; Yanagizawa-Drott, 2014), women's status and fertility (Jensen and Oster, 2009; La Ferrara et al., 2012), or policy outcomes (Strömberg, 2004; Eisensee and Strömberg, 2007; Snyder Jr and Strömberg, 2010). We complement this stream of the literature by highlighting a mechanism through which media could influence political outcomes-by providing an efficient channel to raise political donations. A number of earlier studies point to the challenges of measuring the benefits from social media (Lovett and

Staelin, 2012; Bollinger et al., 2013; Culotta and Cutler, 2016; Ma et al., 2015) as well as the relationship between political positioning of media and their revenues (Gal-Or et al., 2012; Yildirim et al., 2013). Our findings suggest concretely that these platforms can generate positive returns.

2 Background

Use of Social Media by Politicians

Until recently, traditional media held the role of being the primary information channel for politicians, so obtaining coverage on newspapers and TV outlets became crucial for electoral success. Candidates further engaged in dissemination of information about their candidacy and policy goals by the speeches they give along the campaign trail and through public appearances (Garcia-Jimeno and Yildirim, 2015). Today, a reported 80% of the politicians around the world use Twitter to communicate with their constituency[1]. The content of this communication is more personal compared to the regular campaign messages and includes information about politicians' lives and activities outside of politics. While politicians who are well known and hold high political positions typically reach out to several million followers on Twitter, lesser known politicians communicate with several thousand individuals. Barack Obama, for instance, in 2016 had over twenty-three million followers while Mike Pence, Orin Hatch and Jared Polis had over thirty thousand accounts following them. According to our data, the number of Congressional candidates who use Twitter increases from 741 in 2009 to 1,024 in 2010, to 1,488 in 2012, and to 1,814 in 2014.

After the 2008 election, scholars predicted increased and targeted web use by political campaigns at the federal and local level (Towner and Dulio, 2012). This included use of Social Networking Services (SNSs), which allow candidates to build profiles and showcase connections within a delimited system (Boyd and Ellison, 2010; Boyd and Marwick, 2011). Among these sites, Twitter is unique due to its confinement to 140-character messages and the lack of restrictions on viewing messages in the form of account-owner permissions. Followers are said to establish connections for the content, rather than the relationships, resulting in numerous ties that span physical and social disparities (Virk et al., 2011). Twitter brings with it new possibilities for candidate-voter interaction as the "@username" function allows candidates to reply directly to other users and promote dialogue. Managing a Twitter audience, therefore, requires constant activity to respond to, monitor, and understand audience interests (Boyd and Marwick, 2011). It is not surprising that this platform is fraught with potential communication challenges. These challenges include an absence of authoritative hierarchies (Metzgar and Maruggi, 2009), the possible loss of message control (Gueorguieva, 2008; Johnson and Perlmutter, 2010) and overall blurring of traditional audience conceptualizations (Marwick et al., 2011). Scholars and pundits also question whether the overall use of SNSs by politicians actually matters when it comes to voting outcomes (Kushin and Yamamoto, 2010; Baumgartner and Morris, 2010; Zhang et al., 2010). Although the number of Twitter users continues to increase, only a fraction of those users report using the site to gather political information (Smith and Rainie, 2008; Smith, 2011). Right now, Twitter and other SNSs are still seen as complementary to traditional outreach mediums (Towner and Dulio, 2012). The true

[1] http://www.adweek.com/socialtimes/world-leaders-twitter/495103

payback may be in organizing volunteers and activists, an aspect some maintain is overlooked (Abroms and Craig Lefebvre, 2009). The primary benefits of the SNS as a campaign tool are said to include low costs, enhanced recruitment of volunteers and contributions, and a space for lesser known candidates (Gueorguieva, 2008). One benefit of all social media is the direct nature of the communication which allows candidates to bypass traditional media outlets (Lassen and Brown, 2010).

There are a number of studies providing correlational evidence on how social media influences campaigns. Metaxas and Eni (2012), for example, comment on the relationship between social media use and elections from the perspective of predicting electoral outcomes, while Hong and Nadler (2011) demonstrate how the use of Twitter correlates with the shifts in polls during election periods. From the perspective of politicians, policy makers and consumers of social media, documenting a robust causal impact of Twitter and showing different mechanisms at play is essential to understanding the role of social media in the electoral process.

Media and Incumbency Advantage

Incumbency advantage is among the best-documented electoral patterns in the United States (Ansolabehere et al., 2006a). Starting with a 1-2% point advantage in the 1940s, incumbents reportedly enjoyed increasing levels of electoral wins, reaching about 8-10% during the 2000s. A rich literature offers explanations for why known and incumbent politicians with experience enjoy this advantage in elections. Higher chances of re-election may simply stem from differences in quality of the candidates. Some of these politicians may be more skilled than their opponents (Jacobson and Kernell, 1982) and hence enjoy higher chances of re-election. The incumbency advantage can also be due to the opportunity the incumbents enjoy to use staff and committee positions to raise campaign funds (Cox and Katz, 1996). Differently from the earlier studies, we find that new politicians see returns from opening an account on social media rather than the more experienced politicians, although experienced politicians may have access to greater resources to maintain an advantage compared to new politicians.

Another advantage incumbents hold is the disproportionate attention they receive from the media. During elections, traditional media acts as the primary source of information about the candidates and also influences the decisions of the voters through endorsements. Survey-based findings suggest that incumbents enjoy higher levels of media coverage and greater levels of endorsements (Goldenberg and Traugott, 1980; Clarke and Evans, 1983; Ansolabehere et al., 2006a). Traditional media such as TV and newspapers support the better known politicians, and voters are more likely to favor candidates who they can recognize (Jacobsen, 1987). Ansolabehere et al. (2006b) find that endorsements influence the outcome of an election by about 1-5% points. These findings suggest that experience in politics - both through higher public recognition and through holding a public office - can put new politicians at a disadvantage, and discourage entry into political contests (Cox and Katz, 1996), which reduces the competitiveness of electoral races. Less competitive races result in lower responsibility and accountability towards constituents by politicians (Carson et al., 2007). These concerns together suggest that new technologies, which can reduce the incumbency advantage, can help elections to be contested on fairer grounds.

3 A Simple Model of Political Donations

We sketch out a simple partial equilibrium framework of donation decisions by potential political donors. We analyze donation decisions in situations where politicians do and do not use Twitter. In this framework, we abstract away from explicitly modeling the strategic decision of politicians to join Twitter. We use the model to derive some testable predictions on donation decisions which we then take to data.

The Framework

Consider a setting where politicians can be either new or experienced, indexed by $i \in \{e, n\}$. A politician i has a 'type' or quality, $\theta_i \in [0,1]$ interval. The politician knows her θ_i. There is a unit mass of potential donors. We assume that all potential donors want a higher 'quality' politician which, in this context, can be interpreted as competence, honesty, or experience of a politician.[2]

We adopt a separable utility framework for donors similar to Chiang and Knight (2011) and Matejka et al. (2015). An individual donor d has the following utility from donating to politician i:[3]

$$U_{di} = \theta_i - c_d$$

The term $c_d \sim U[0,1]$ captures the cost of donating which affects each donor idiosyncratically. We normalize the outside option of the donors to 0. The donors do not observe θ_i but hold (unbiased) prior beliefs such that

$$\theta_i \sim N(\bar{\theta}_i, \sigma^2_{i0})$$

We assume that $\bar{\theta}_e > \bar{\theta}_n$ which will imply that ex-ante, without Twitter, experienced politicians have an advantage in receiving higher donations relative to newer politicians. We will focus on the case where also $\sigma^2_{n0} > \sigma^2_{e0}$. A higher variance for new politicians implies that ex-ante, the donors place less confidence in their estimate of θ_n relative to θ_e. This structure is in line with the evidence that experienced politicians hold an informational advantage over newer candidates as documented by Anderson (2004) and Oliver and Ha (2007).

If a politician does join Twitter then she can provide information to the donors or could send (rational or non-rational) persuasion similar to advertising messages. The politician can send a message m to the voters such that:

$$m_i = \bar{\theta}_i + \epsilon_i$$

with $\epsilon_i \sim N(0, \sigma^2_{i\epsilon})$.

[2] Analyzing quality instead of ideology is more pertinent in our context, since we analyze donations within states, where ideological differentiation within a party would be limited. In such a situation, information via Twitter is more likely going to be about their integrity, experience and track record. This modeling choice is in line with Durante and Knight (2012) as well as Knight and Chiang (2011).

[3] The linear utility framework is in line with (Chiang and Knight, 2011) and Durante and Knight (2012).Matejka et al. (2015) adopt a more general framework where $u(\theta_i)$ is concave and differentiable. Our main insight would hold in such a setting as well (with more algebra) as well as other frameworks where voters have bi-modal policy preferences.

The Donation Decision

To highlight how joining Twitter affects donations differentially for new and experienced politicians, we analyze the donations received by each type of politician with and without Twitter. If the politician does not join Twitter, donor d will donate if

$$E(\theta_i) \geq c_d$$

Normalizing each donation to 1, the total amount of donations is then given by $\bar{\theta}_i$ since $E(\theta_i) = \bar{\theta}_i$. If a politician does join Twitter then she will send a message m_i which will be used by the donors to update their beliefs about θ_i. The posterior belief after seeing m_i is:

$$E(\theta_i|m_i) = V_{i0}m_i + V_{i\epsilon}\bar{\theta}_i$$

where $V_{i0} = \left(\frac{\sigma_{i0}^2}{\sigma_{i0}^2+\sigma_{i\epsilon}^2}\right)$ and $V_{i\epsilon} = \left(\frac{\sigma_{i\epsilon}^2}{\sigma_{i0}^2+\sigma_{i\epsilon}^2}\right)$. If a politician does join Twitter, donor d will donate if

$$E(\theta_i|m_i) \geq c_d$$

and then the amount of donations received is $E(\theta_i|m_i)$. We define $\Delta_i \equiv E(\theta_i|m_i) - E(\theta_i)$. If $\Delta_i > 0, \forall i$ then we can establish the following proposition.

Proposition 1. *A new politician gains more from joining Twitter relative to a more experienced one if:*

(1) Variance of prior belief is higher for new as compared with an experienced politician, i.e. σ_{n0}^2 *is high relative to* σ_{e0}^2*, and* $\frac{\partial(\Delta_n - \Delta_e)}{\partial(\sigma_{n0}^2 - \sigma_{e0}^2)} \geq 0$.

(2) Twitter message is more informative for new politicians as compared with experienced ones, i.e. $\sigma_{e\epsilon}^2$ *is high relative to* $\sigma_{n\epsilon}^2$*, and* $\frac{\partial(\Delta_n - \Delta_e)}{\partial(\sigma_{e\epsilon}^2 - \sigma_{n\epsilon}^2)} \geq 0$.

Proof. The proof follows straight from writing out the expressions for Δ_i. $E(\theta_i|m_i) - E(\theta_i)$ is simply $\left(\frac{\sigma_{i0}^2}{\sigma_{i0}^2+\sigma_{i\epsilon}^2}\right)\epsilon_i$. This implies that $\Delta_n - \Delta_e = V_{n0}\epsilon_n - V_{e0}\epsilon_e$. The comparative statics in the proposition follow directly. **QED**

The proposition implies new politicians see a bigger increase in political donations from joining Twitter relative to experienced politicians, the larger σ_{n0}^2 is relative to σ_{e0}^2. The condition $\Delta_n \geq \Delta_e$ is also likely to hold if the messages sent by new politicians have higher precision: $\sigma_{e\epsilon}^2 \geq \sigma_{n\epsilon}^2$.

Donations and Twitter Penetration

Till now, we have assumed that there is universal access to Twitter and all donors observe how informative the use of Twitter is for all politicians. As in our empirical model, we assume that there are different geographical regions (states), $s \in \{1, 2, ..., S\}$ with different twitter usage. Each state has a unit mass of potential donors. Moreover, we assume that only a (random) fraction ϕ_s uses Twitter. This assumption is in line with Butters (1977). This penetration coefficient varies across states with $\phi_1 \geq \phi_2 \geq .... \geq \phi_S$.

Assuming that Twitter penetration is the only dimension which varies across regions, we can easily see

that politicians in regions with a higher ϕ_s will receive a bigger increase in donations by joining Twitter:

$$\phi_s \Delta_i \leq \phi_{s-1} \Delta_i$$

This also shows that if $\phi_s = 0$ for some s then in that region there will be an insignificant increase in donations for both experienced and new politicians.

4 Data

Our study uses data from a variety of sources. We compile a list of politicians available from the Federal Election Commission (FEC) which includes those who either registered with the FEC or whose name is mentioned on the state ballot for an election to the U.S Senate or House of Representatives for three election cycles from 2009 to 2014.[4] For each politician, we combine weekly data on campaign contributions with data on Twitter activity. We also acquire information on the campaign expenditures and the number of media mentions on Google News and Google Blogs of each candidate. Finally, we gather data about how Twitter usage compares to the usage of other websites in each US state, using data from the company comScore. Summary statistics for various variables are provided in Tables 1.

Campaign Contributions and Expenditures

The main data source for political donations for our study is the Federal Elections Committee (FEC), which makes data on campaign contributions for each candidate publicly available. Our data focuses on the contributions to candidates, rather than to PACs or other organizations. In most parts of the analysis, we limit our attention to donations under \$1000 [5], as larger donations are more likely to be caused by different motivations. The data details the amount of each contribution, its date, and the name of the donor. For the subsample of donations below \$1000, the average amount of donations per week for a politician is \$516 and the median amount is \$500. We use donations aggregated at the politician-week level.

The source of data for the campaign expenses is the Center for Responsive Politics (opensecrets.org). The site lists the exact date for each piece of expenditure made by each candidate, and we use the aggregated weekly campaign expenses of the candidate as a variable in our analysis.

Twitter Account Opening

For each politician in our list, we collect information on their Twitter activity.[6] We combine an automated script with a manual check to gather information about whether a politician has Twitter account or not and if there is one, we collect a variety of data related to it. We identify the date that the account was first activated and supplement it by information on the number of tweets, re-tweets, the text of all the tweets as well as the number of followers.

[4] Elections are held every two years in even-numbered years.
[5] We also study donations between \$1000 and \$3000, in Section 6.
[6] A detailed description of the data collection process is given in the Data Appendix.

Figure 1 demonstrates the distribution of the date of Twitter account opening for the politicians before 2014 campaign. The distribution shows that entry on Twitter takes place continuously between 2009 and 2014, and there was almost no entry before 2009. Substantial variation in entry dates reduces the concern that politicians' entry may correlate with the timing of a few specific events. Note also that many of entries happen outside traditional campaign periods. To further reduce the concern that donations are influenced by other campaign activities than Twitter account opening, we drop any Twitter account which has "2010", "2012", "2014" or "4" (e.g., "@chip4congress", "@MCarey2012") in the handle string which may indicate that it was started for an upcoming campaign.

Twitter Penetration

We collect information about the use of Twitter relative to other sites by the general population in the U.S. by creating a penetration measure. We construct this metric by using data from comScore internet browsing data. The dataset provides a panel of fifty thousand households who have been tracked in their online activities throughout the period of data collection. Each household's browsing activities are tracked through a machine and all browsing of websites is recorded. This allows us to create a measure of engagement with Twitter relative to all sites, which we refer to as Twitter penetration which is aggregated at the state-year level. Formally:

$$\text{Twitter Penetration} = \frac{\text{Number of Site Visits to Twitter}}{\text{Number of Visits to All Websites}}$$

Twitter penetration plays a significant role in our identification strategy which we will detail subsequently. Note that, we normalize Twitter penetration so that mean penetration is equal to 1 (with the median penetration being 0.99 - close to the mean of the distribution).

Unfortunately, we were not able to obtain data on website visits from mobile devices. However, there are reasons to think that mobile Twitter penetration is strongly correlated with desktop Twitter penetration

News and Blogging Data

For each politician on our list, we collect information on the number of media mentions for a window of ten weeks before and after they started using Twitter. We run a search for the number of times his or her name has appeared in Google News and Google Blogs. We use this information to check whether there are systematically more media mentions of a politician around the time her Twitter account is started. If there are various other events related to a politician's campaign which could affect the amount of donations generated then the number of media mentions should capture it. Thus, we can test whether our results are driven by some other events which happen at the same time as entry to Twitter.

Politician Data

We collect additional data about the politicians using two different data sources. The first source is FEC, and the second is VoteSmart database, which provides information about their age, education, income, and voting history.

Throughout the empirical section, we extensively use the division of politicians into 'new' ones and 'experienced' ones. A politician is 'new' if she was never elected to Congress before. If the politician has already won an election in the past, then she is classified as 'experienced'. We also use the more traditional classification of incumbents and challengers. Note that a challenger could be an experienced politician if she was elected to Congress before. We present summary statistics separately for experienced and new politicians in Table 2.

Other variables

Finally, we also use data on demographics at the state level such as household income, share of rich (i.e. share of households with over 250K income), share with college education, and share of African-American population from the Census, aggregated at the state level. We use data on newspaper circulation per capita from the American Association of Newspapers. We also use data on the vote share received by George W. Bush in 2004 Presidential elections from uselectionatlas.org.

We collected data on dates of the first public post on Facebook for all the politicians in our list. We then create a dummy variable equal to one if a politician had opened a Facebook account before joining Twitter, and zero otherwise (i.e. for politicians with a Twitter account).

5 The Empirical Framework

5.1 Empirical Hypotheses

Based on the simple framework highlighted in the model, a number of hypotheses can be derived.

(1) Politicians, on average, can potentially increase their donations by joining Twitter since it serves as an additional channel of communication.

(2) The gain for new politicians from joining Twitter will be higher relative to more experienced ones due to being relatively unknown initially.

(3) The gain from joining Twitter is higher for politicians who tweet more informatively.

(4) States with higher Twitter penetration will contribute more to politicians joining Twitter.

Our main empirical hypothesis is that politicians who join Twitter gain access to an additional, relatively inexpensive channel of communication with their electorate. As a result, several things could happen. First, Twitter might help to provide new information to the members of politician's constituency. This information channel is more likely to hold for new politicians. Second, Twitter could allow politicians to engage in non-rational persuasion, potentially through repeated interactions through political messages which is more likely to hold for more experienced politicians who have greater access to their potential donors across different media platforms. In all these cases, we expect Twitter to affect the behavior of potential donors in a positive way. To investigate potential channels, we also check whether this hypothesis holds for new politicians, new donors, and in less saturated information environments.

Figure 2 demonstrates how political donations change in high Twitter penetration and low Twitter penetration states, controlling only for politician and week fixed effects, before and after Twitter entry. There

are two takeaway points from this figure. First, donations increase after joining Twitter, but not before, and this effect is stronger in places with high Twitter penetration. Second, there were no significant pre-trends in the difference between high and low Twitter penetration places before joining Twitter. Overall, Figure 2 illustrates our main point: the entry to Twitter helps politicians to raise political donations, and more so in high Twitter penetration places.

5.2 Main Specification

To study how opening a Twitter account influences the amount of political donations received, we use a difference-in-differences approach exploiting the precise timing of entry on Twitter. The main specification we estimate is:

$$DonationOutcome_{it} = \alpha_{im} + \theta_1 Entry_{it} + \theta_2 Entry_{it} \times Penet_{sy} + \theta_3 Entry_{it} \times \mathbf{X_s} + \theta_4 log\left(Expenditures_{it}\right) + \theta_5 t + \epsilon_{it} \quad (1)$$

where i is the index for politicians, t is a week level time index, s is the index for state. $DonationOutcome_{it}$ will represent various ways of measuring donations received by politician i in week t, such as the log of aggregate dollar value of donations and the probability of receiving at least one donation. $Entry_{it}$ is a binary variable equal to 1 if politician i has a Twitter account in week t and 0 otherwise. $Penet_{sy}$ is the level of Twitter penetration or usage in each state s which we aggregate at the annual level (and hence the subscript y). α_{im} is a politician-month fixed effect. $\mathbf{X_s}$ is a set of controls including average education in state, median income, percent rich (i.e., households with annual income of over $250,000 or more), percent voting for Bush in the 2004 elections, and race (percentage of the African-Americans), all interacted with Twitter entry. $Expenditures_{it}$ is campaign expenditures by politician i during week t. We do not include direct effect of $Penet_{sy}$ as it is perfectly collinear with politician-month fixed effects.

We allow for flexible controls in our specifications with politician-month fixed effects. Politician-month fixed effects account for unobserved differences in a politician's ability to attract donations, and we control for this by allowing this ability to fluctuate temporally from month to month. Our identification therefore comes from precise timing, as we effectively look at donations just before and just after Twitter entry. Note that our baseline results remain unchanged qualitatively and quantitatively if we replace linear time trend with week fixed effects.[7] We cluster standard errors at the level of the state, to account for both cross-sectional and time-series variation.

Our main coefficient of interest is θ_2, corresponding to the interaction between entry on Twitter and penetration. If Twitter indeed allows politicians to share new information with the members of constituency, we expect this coefficient to be positive and significant.

We do not claim that the decision to join Twitter is exogenous or taken completely at random, since this decision could be driven by a host of factors which we cannot fully observe. Our identification rather assumes parallel trends, i.e., that the difference in political donations, unexplained by politician-month fixed effects, would remain the same in the absence of Twitter entry across states of high and low Twitter penetration.

We use a variety of placebo checks to check the credibility of this identifying assumption. One potential

[7] See tables A12 and A13 in the Appendix.

issue with our identification strategy might be other events happening simultaneously which would drive both the Twitter entry in areas of high and low penetration as well as donations to the politicians. If this is true and the politician is involved in multiple campaign activities, this is likely to reflect on the campaign expenses reported to FEC. We test whether campaign expenditures show a spike around the time of opening a Twitter account. We use the mandatory campaign expenditure data disclosed to FEC by the candidate. These expenditures may relate to activities on the campaign trail such as visits to towns, or a TV or newspaper advertisement purchases.

As another check to test for other events which may coincide with adopting Twitter, we look at the coverage of politicians in the news media. Any report or feature of the candidate by the traditional media coinciding with the opening of a Twitter may influence donations or be a consequence of some unobserved event simultaneously affecting both. Using data collected from Google News and Google Blogs, we test whether the number of articles or the blogs which mention a candidate increase discontinuously around the time of opening an account.

Our identifying assumption would also be violated if the characteristics of the regions which makes individuals spend a higher proportion of their online visits also correlated with their tendency to donate. To check for any systematic differences, we regress aggregate donations on a set of demographic characteristics included in $\mathbf{X}$, interacted with a dummy for being on Twitter.

6 Baseline Results, Placebos and Mechanisms

6.1 Baseline Results

We begin our analysis with the main specification (1) to evaluate the impact of joining Twitter on the aggregate weekly political donations received. The main independent variable of interest is the politician's presence on Twitter interacted with Twitter penetration. The results of the estimation are presented in Table 3, with several sets of controls included in the estimation. As one can see from this table, our main coefficient of interest, having an account on Twitter interacted with Twitter penetration, is positive and significant in the specification which includes politician-month fixed effects (columns (2)-(5)). The coefficient remains stable in magnitude (0.35-0.38) as controls for campaign expenditures, time trend, and census controls interacted with joining Twitter are added. The direct effect of being on Twitter becomes insignificant once time trend variable is introduced (column 4). This means that in areas with no Twitter penetration, joining Twitter is not associated with an increase in donations, consistent with what our theory predicts. Columns (6) and (7) estimate equation (1) separately for the sub-samples of new and experienced politicians. Column (6) suggests that joining Twitter is especially helpful for new politicians, consistent with our theoretical prediction. However, we do not find any significant impact of joining Twitter on donations for experienced politicians (column (7)) even in areas of higher Twitter usage. A potential explanation for this difference is that new politicians typically are at an informational disadvantage compared to the experienced politicians.

We do some back of the envelope calculations to interpret the magnitudes in our regressions. We include

both the campaign and non-campaign periods in our estimation, and the average donation per candidate per week is $1,534 per week and the average length of time being on Twitter after joining till the end of the month is 2.79 weeks (note that once the month is over, the coefficient that indicates being on Twitter for a politician becomes perfectly collinear with politician-month fixed effect). Using our coefficient, the back of the envelope calculation yields $1,534 \times 0.378 \times 2.79 = $1,618. Here we make the calculations for an average politician in a place with mean Twitter penetration (which is normalized to 1). Note that this number ($1,618) is likely to be an underestimation of the effect of Twitter on aggregate funds raised, as Twitter is likely to continue to help politicians receive donations even after the first month of adoption. Similarly, for new politicians, a similar number is obtained by multiplying $1,077 (average donation per week) with 0.69 (the coefficient from column 6) and with 2.79 weeks, which yields $2,078. Overall, these results suggest that adopting a new communication channel by joining Twitter leads to an average increase of 1.6% (for all politicians) or 2.6% (for new politicians) of the total donations below $1,000 raised over a two year campaign period.

The results in Table 3 suggest that politicians are able to raise more money after joining Twitter. Table 4 tests whether similar results hold at the extensive margin, i.e., if politicians are more likely to receive donations in a given week after joining Twitter. The results in Table 4 suggest that joining Twitter helps to raise the likelihood of receiving donations. In terms of the magnitudes, the probability of at least one donation per week increases by 5.1 percentage points for all politicians (column (2)), and for 8.4 percentage points for new politicians (column (3)). The results for experienced politicians remain insignificant (column (4)), consistent with the results in Table 3. When we analyse the impact of being on Twitter on the number of donations received in a week, we get a similar picture. New politicians joining Twitter in higher penetration areas get a significantly larger number of weekly donations (column (7)) while this effect does not hold for the more experienced politicians (column (8)).

Finally, we also check whether the results for donations hold for different donation sizes. We estimate equation (1) for donations between $1,000 and $3,000. We report the results for both aggregate donations and for the probability of at least one donation in a given category. Table 5 summarizes these results. We find that while there is no average effect of joining Twitter for the sample of all politicians, there is still an impact of joining Twitter for the sample of new, inexperienced politicians (column 3). The size of the interaction coefficient (0.57) is smaller than the size of interaction coefficient for donations below $1,000 (Table 3). In terms of absolute dollars, however, the impact seems to be stronger, as these are larger donations. The average weekly sum of donations from a given category is $2,313. After multiplying $2,313 by 0.57 and 2.79 we find that Twitter entry can account for at least $3,695 extra from donations between $1,000 and $3,000. This constitutes 2.1% of total donations to an average politician in a given category. Similarly, columns (5)-(8) report the results for extensive margin. For an average politician, the probability of receiving at least one donation between $1,000 and $3,000 goes up by 3.3 percentage points, while a similar probability for a new politician goes up by 6.7 percentage points. Thus, the results for the extensive margin are smaller in magnitude than similar results for donations below $1,000.[8] Estimating our main specification for donation

[8] We demonstrate the robustness of our findings by testing different specifications. Varying the window size in the difference-in-differences specification does not alter the estimate of the interaction between being on Twitter and Twitter penetration in

values greater than $3,000 does not show a significant effect of Twitter.[9] This is expected, since donors who make large contributions may have different reasons to contribute compared to the small donors. They may also be less likely to be influenced by a politician's regular Twitter communications.

Overall, the results in Tables 3-5 and in Figure ?? suggest that joining Twitter allows the politicians to raise a larger dollar sum of donations, but only if they are new to politics.

6.2 Persuasion Rates

To be able to compare the magnitudes that we uncover with other studies in the literature, we compute persuasion rates (DellaVigna and Kaplan, 2007; DellaVigna and Gentzkow, 2010). We cannot compute persuasion rates for the general audience, but we can compute persuasion rates for a set of early followers of a subset of these politicians. Unfortunately, we do not observe how the number of followers evolved for every politician from when they opened their accounts.[10] However, we observe the number of the followers for some politicians within 3 months of their account opening for two points in time: at the time of data collection by Halberstam and Knight (2014) and at the time of our data collection. For 29 politicians (21% of the politicians who opened their accounts in 2012), we estimate that the mean number of followers is 150.93 within the first 4 months and 103.83 within the first 3 months of opening an account on Twitter,.[11] Note that we compute persuasion rate for a non random sample of the population, so we cannot assume that persuasion rates based on our calculations apply to the general population. In contrast to most media studies, we are forced to use temporal rather than spatial variation while carrying out this exercise.

More specifically, we would like to estimate a persuasion rate, using the formula $f = \frac{y_t - y_c}{e_t - e_c} \times \frac{1}{1-y_0} \times 100$ from DellaVigna and Gentzkow (2010). Here our treatment is the entry of politician to Twitter. Similar to approach used in DellaVigna and Gentzkow (2010), we assume that $e_t - e_c$ is given by 100%, i.e. that all followers of any given politician observe the treatment (entry to Twitter and subsequent tweets within a first month),. The $y_t - y_c$ is given by the coefficient from column (6) of Table 4, divided by the number of followers, (150 or 103.83, depending on the number of months after account opening). Finally, y_0 is a counterfactual estimate for the number of donations in the absence of Twitter entry. As before, we multiply all magnitudes by 2.7 weeks, the average number of weeks in the month after Twitter entry. Using the numbers above, we obtain that the persuasion rate (under the assumption of 100 and 150 followers) is equal to:

$$f_{100} = (3.518 \times 0.1546162 \times 2.7/103.83)/(1 - .35231644 \times 2.7/103.83) = .01482742 = 1.41\%$$

Table A1 in the Appendix. When we vary the window size for our diff-in-diff specification from ±5, ±10 weeks to up to ±300 weeks, our aggregate estimates stay highly stable at 0.37 and significant at the 5% level throughout.

[9] The impact of Twitter on donations between $3,000 and $5,000, and donations above $5,000, are reported in Tables A10 and A11 of the Appendix.

[10] These data are not made available through Twitter API.

[11] Unfortunately, we do not observe the number of followers within the first two months, so this is the best approximation we have for our exercise.

$$f_{150} = (3.518 \times 0.1546162 \times 2.7/150)/(1 - .35231644 \times 2.7/150) = .01482742 = 0.98\%$$

This persuasion rate is on the lower end of the estimates found in the literature. It is lower than most estimates for persuasion rates of news media (range from 2 p.p. to 20 p.p. for media in the United States), and it is comparable with 1.0 p.p persuasion rate of direct mailing in Gerber and Green (2000), or 0.1-1.0 p.p., persuasion rates of political advertising in Spenkuch and Toniatti (2016). However, we do not find it surprising. The previous persuasion rate literature was focused on voting behavior, while we focus on donations above $200. It is easier to convince people to come to vote than to give a good chunk of money to support the candidate. It is possible that this estimate is an upper bound as some people could get info from twitter without actually becoming a follower. Thus, despite large dollar numbers, these persuasion rates are quite low. Another exercise we can do is to compute a persuasion rate of implied donations. Assume that there are 2 donations below $200 for every donation above $200. Thus, if we look at implied donations, our persuasion rate are 2.94-4.23 p.p., much closer to persuasion rates observed in the literature for newspaper endorsement (Chiang and Knight (2011), 6 p.p. for unexpected endorsements, or 2 p.p. for expected endorsement), or turnout effects of entry of television (4.4 p.p., Gentzkow (2006)).

6.3 Placebo tests

Our identifying assumption is that the difference in political donations, unexplained by politician-month fixed effects, would remain the same in the absence of Twitter entry across areas of high and low Twitter penetration. While we cannot test this assumption directly, we conduct several tests to ensure that the data is indeed consistent with our identifying assumption.

6.3.1 Campaign Expenditures

A first potential threat to identification is the possibility of a correlation between the timing of Twitter entry and other campaign activities which can contribute to funds raised. While we do not have extremely detailed measures of campaign activities, we use campaign spending per week as a proxy. The estimates in Table 3 show that weekly campaign expenditures are significantly correlated with campaign contributions during the same week, indicating that this measure is indeed meaningful. To check for potential simultaneous changes in campaign activities, we test if there is a spike in campaign expenditures around the date politicians start using Twitter. Table 6 shows that controlling for politician-month fixed effects and including a week time trend, joining Twitter does not predict an increase in the campaign expenditures neither in high nor in low Twitter penetration areas. Both the direct and the interaction terms are insignificant for the full sample (column (4)) in explaining campaign expenditures. This result also holds separately for both new (column (5)) and more experienced politicians (column (6)). To the extent that campaign expenditures capture other activities of the politician around the same time with opening a Twitter account, this result provides a reassuring check for our identification strategy.

6.3.2 Political Advertising Expenditure

A particular measure of campaign activities which we use is the amount of political advertising expenditure by the candidates. Like for campaign expenditures above, we check whether there was a spike in political advertising expenditure in particular around the time candidates joined Twitter which could also involve a call for raising more funds. Table 7 shows that as soon as we include politician-month fixed effects (column (2)), joining Twitter does not predict an increase in political advertising in either high or low Twitter penetration areas. Both the direct and the interaction terms are insignificant for the full sample and the full set of controls (column (5)) in explaining political advertising expenditures. This result also holds separately for both new (column (6)) and more experienced politicians (column (7)) though the coefficient is negative and weakly significant at the 10% level for the experienced ones. Overall, we show using a measure particular to political advertising does not correlate with politicians joining Twitter which provides greater confidence in our estimation approach.

6.3.3 News and Blogs Coverage

It is also possible that politicians join Twitter as part of their information campaigns, and opening Twitter accounts coincide with the spikes in coverage of these politicians by traditional media outlets. Media mentions of a politician might capture both additional information shocks voters receive and events a politician is involved in (which may not be reflected in campaign expenditures) which drive donations independently of Twitter. To address this concern, we collect data on the media mentions of a politician. We run a search for each politician's name in Google News and Google Blogs for a ± 10 week window around the time of opening of their twitter account[12]. Figures 2 and 3 demonstrate how the news mentions and Google Blogs coverage of the politicians evolve over this period, and they do not indicate any patterns. Table 8 reports the results of this estimation. We use the total number of mentions in the news as the main dependent variable in our specification. Overall, the estimates in Table 8 suggest that being on Twitter interacted with Twitter penetration is not significantly associated with the number of news mentions (columns (1)-(4)) and this holds for both new (column (3)) and experienced politicians (column (4)). Moreover, we find that these results also hold when we look at the number of blog mentions as the dependent variable in Table 8 (columns (5)-(8)), as the coefficient for Twitter entry and penetration interaction remains insignificant and is actually negative in all the specifications.[13]

6.3.4 Twitter Entry, Twitter Penetration and Demographics

A further concern about our identification strategy is that Twitter penetration merely serves as a proxy for the income, education, or other socioeconomic characteristics of the state, and what we observe is a higher responsiveness to the shock (joining Twitter) in richer, more educated, or more liberal places. To ensure that it is not the case, we conduct another check. In particular, we test whether donations received

[12] We search for the full name of the politician and record the number of hits we find on Google News and Google Blogs.

[13] Another relevant issue to address is a check on politicians' use of other social media platforms such as Facebook. To test the robustness of our results, we collected information on when each politician opened her Facebook account (if she did) and we find no robust relationship between having a Facebook account before and being on Twitter and Twitter penetration interaction (See Table A4 in the Appendix).

can be explained by differential effects of entry on Twitter with different socioeconomic controls, such as the median household income in a state, the share of people who earn over $250,000 annually, the share of people with a college education, the share of people who voted for Bush in 2004 as well as the share of African Americans in the state. We report these results in Table 9. For the ease of comparison, the results with the coefficient for Twitter entry interacted with Twitter penetration (from specification in culmn (4) of Table 3) are reproduced in column (1). Our results suggest that the interaction of being on Twitter with each of controls mentioned above is insignificant both economically and statistically, with coefficients in the opposite sign to what a proxy hypothesis suggests (columns (2)-(6)). [14] Therefore, the data is not consistent with Twitter penetration being a proxy for a major socioeconomic characteristics of a region.

Overall, while we cannot test our identifying assumption directly, the placebo checks in this section suggest that unobserved heterogeneity, and other potentially simultaneous campaign activities in particular, are not driving our results.

6.4 Mechanisms

The main findings suggest that a politician's adoption of Twitter causes an increase in the aggregate donations s/he receives. There could be different channels through which Twitter affects the behavior of donors. The first one is an information channel. Opening a Twitter account allows the politicians to access a new, relatively inexpensive, channel of communication with its potential constituents. Moreover, information on the politicians' ability and policy stance are distributed at a low cost through social media. For donors who do not know about a candidate or are uninformed of her policies, this channel serves to create awareness. An alternate, second mechanism could be a persuasion channel. For potential donors who already know the candidate, communication via Twitter can create repeated exposure and persuade them to contribute more. This channel is akin to persuasive advertising in the Industrial Organization literature.

Our findings demonstrate that social media raises donations only for the new politicians and not for the experienced ones. This is in line with our theoretical framework, which states that the marginal return to information provision through Twitter is likely lower for the experienced candidates, since their quality, experience, and policy positions are already known. For a newcomer, it is cheap to open an account on Twitter, and information dissemination through online word of mouth is possible at a relatively low cost. Our main result that joining Twitter only helps inexperienced politicians is thus consistent with the information mechanism.

In this section, we present a number of additional tests that allow us to check what mechanisms our data is consistent with. First, we check whether our estimates are stronger for new or repeat donors. We classify each donor as new if no donor with the same first and last name has contributed to a particular Congressional candidate before. Next, we check whether Twitter effects are stronger or weaker in places with high newspaper circulation. Finally, we also analyze Tweeting activities by politicians to document how differences in Tweeting activity and content of Tweets affect donations.

[14] We carry out additional checks to find that being on Twitter (interacted with socioeconomic controls) is not driving Twitter penetration. In addition, states with higher Twitter penetration do not see significantly more Twitter account openings. We analyze this in terms of levels and first differences of weekly Twitter penetration but find no economic or statistically significant relationship (Please see Tables A5 and A6 in the Appendix).

6.4.1 New vs Repeated Donors

We conjecture that a politician's presence on social media have two possible ways of influencing donors. First, it is possible that a politician's presence simply changes the amount individuals contribute without altering the donor population. A second plausible argument, in line with an information channel, is expanding the donor base, with new donors hearing about and contributing to the campaign for the first time. When the second explanation holds, being on Twitter will affect the probability of receiving donations as well. We provide evidence that indeed Twitter presence is associated with new donors rather than just a shift in the donation amounts of old donors.

We split the donations received by politicians into those received from new and repeat donors to re-estimate our diff-in-diff specifications. Panel A of Table 10 shows the results for new donors are in line with our information hypothesis. Using Twitter in a high penetration state leads to an increase in aggregate donations received from new donors (columns (1)-(2)) but splitting the sample into new and experienced politicians shows that new donors donate more to only new politicians (column (3)) and not to the experienced ones (column (4)). The same results hold when, instead, we look at receiving at least one donation per week as the dependent variable (columns (5)-(8)). The probability of at least one donation from a new donor goes up by 10.3% for new politicians, but does not increase significantly for new politicians. Panel B of Table 10 shows the estimation for old donors. We do not find any effect of being on Twitter on donations received from old donors for either new or experienced politicians for aggregate donations and for receiving at least one donation per week. This, again, is consistent with the explanation that Twitter is expanding the donor base by providing better information about politicians or their policies. In sum, all the results in this section are consistent with Twitter having effect on new donors rather than repeat donors who presumably already have enough information about a politician.

6.4.2 Other means of communication

Blogging on Twitter could be especially useful when voters' other channels to receive information about new politicians and their policies are limited. In this subsection, we re-estimate our benchmark specification considering newspaper circulation of the region the politician is from, separating low and high newspaper circulation regions.[15]

Panel A of Table 11 shows the estimates for states which have newspaper circulation per capita lower than the median while Panel B captures the effects for states with higher than median circulation. In Panel A, one can see that new politicians using Twitter in higher penetration and low newspaper circulation areas receive a significantly higher amount of aggregate donations (columns (1)-(4)) as well as a higher probability of receiving at least one donation per week (columns (5)-(8)). Panel B shows that the effects hold for new politicians (columns (3) and (7)) in high newspaper circulation states but the effects are weaker statistically and quantitatively. The results are suggestive that information from social and traditional media are substitutes.

Overall, these results suggest that Twitter acts as a channel increasing awareness about new politicians

[15] Low (high) circulation refers to circulation per capita below (above) the median circulation per capita across states.

and their policies particularly when other channels of information, such as traditional media, are limited.

6.4.3 Tweeting Activity and Tweet Content

The results indicate that Twitter activity benefits new politicians by attracting new donors, and more experienced politicians do not see a significant return from opening an account on Twitter. To document why new politicians might be attracting more donations, we analyze their Tweeting activity along with the content of their Tweets, which allows us to test Hypothesis 3 formulated above.

To this end, we focus on the coefficient on the triple interaction term between being on Twitter x Twitter penetration x different measures of Tweeting activity within a 25 week window of the politicians joining Twitter. We use the number of Tweets as the intensity with which politicians use Twitter and the number of Retweets as a measure of popularity of their Tweets. We report these results in Table 12. We find that triple interaction term for tweets is positive and significant for the sample of new (column 3) politicians, and the effect is indeed stronger for politicians who Tweet the most after joining Twitter in areas of higher Twitter penetration. The same coefficient is much smaller and is not significant for the experienced politicians as seen in column (4). Columns (5)-(8) show similar results when we use the number of Retweets as a measure of popularity. In line with intuition, new politicians who got more Retweets are likely to get a larger increase in donations in states with high Twitter penetration (column 3, coefficient significant at 1%), while a similar triple interaction coefficient is not significant for experienced politicians (column 4).

The potential social contagion effect of Retweeting activity working only for new politicians is in line with our model where the increase in information has a higher payoff to the relatively lesser known candidates.[16]

Next, we move on to analyzing the content of the politicians' Tweets. It is possible that newer politicians use Twitter as a channel to inform their supporters of their positions and plans, or tell them to take part in the campaigning activities. We find that about 2-3% of the total Tweets include a hyperlink/URL. We find that new politicians who use hyperlinks more often have significantly (at 1%) higher donations in high penetration areas (column (3) in Table 13), while the corresponding coefficient is not significant for experienced politicians. Similarly, we find that politicians who use more 'inclusive' pronouns such as 'We' more often are getting higher donations in high Twitter penetration states, while it does not help the experienced politicians (columns (7) and (8) of Table 13).[17]

We also carry out sentiment analysis based on the content of the Tweets by the politicians in our sample. In particular, we use the Linguistic Inquiry and Word Count methods which analyze the use of articles and pronouns highlighting personality traits of individuals. The scale we use is developed by James Pennebaker and is intended to measure the time-invariant personal characteristics of an individual. The measure takes the recent tweets from a given Twitter handle to compute a score (between 0 to100) of the account owner's 'social' and 'thinking' styles. In particular, we analyze how being 'plugged in' (social style) and 'analytic' (thinking style) correlate with donation levels. New politicians are on average more plugged in and analytic than experienced politicians (see Table 2). A higher plugged in score as a proportion of the total social style

[16] We find that more experienced politicians, on average, send higher number of Tweets and receive higher number of Retweets (see Table 2).

[17] We show that using less inclusive words (e.g., 'I') has no impact on raising donations (See Table A7 columns (5)-(8) in the Appendix).

scores (whichconsists of categoeries plugged in, personable, arrogant, and spacy) correlates with getting higher donations for the new politicians and not the experienced ones (Table 14 columns (3) and (4)).[18] Eventhough new politicians score higher on analytic in thinking style, it is not associated with higher donations (See Table A7 in the Appendix).

Overall, the results in this subsection are consistent with the theoretical prediction that using Twitter more informatively is associated with a greater increase in donations received following opening a Twitter account.

7 Robustness Checks

7.1 Heterogeneous Effects Between Democrats and Republicans

Republican and Democratic voters have traditionally differed in demographic characteristics. Democratic voters are generally ethnically more diverse, have higher education, are religiously unaffiliated, and have lower income. One or more of these characteristics may correlate with internet or social media use, implying that candidates registered with the Democratic Party may have higher returns from being on Twitter because the medium appeals to their constituents. We test whether Twitter has an asymmetric effect on candidates from the two parties.

Panel A of Table A9 (in the Appendix) shows the estimates for the Democrats while Panel B demonstrates the effects for the Republicans. In Panel A, one can see that new Democratic politicians using Twitter in higher penetration areas receive a significantly (at the 1% level) higher amount in aggregate donations (columns (1)-(4)) and have higher probability of receiving at least one donation per week (columns (5)-(8)). But experienced Democratic politicians do not show a gain. In Panel B, the effects hold for new Republican politicians (columns (3) and (7)) as well, but at substantially weaker statistical levels (significant at the 10% level). Overall, these results suggest that Twitter adoption has heterogeneous effects across the two party candidates and Democrats gain substantially more from it, possibly because of the demographic differences between the target audiences of the two parties.

7.2 Excluding the Year 2009

One concern related to our baseline estimation is the disproportionate number of accounts opened in 2009. While allowing for politician-month fixed effects and a week time trend (or week fixed effects) to account for any idiosyncrasies of a particular time period, we would not want our estimates to be driven by only one year's worth of data. Hence, as a robustness check, we exclude any accounts opened in 2009 and re-estimate our baseline specification.

Table A2 in the Appendix shows that the results remain in line with our baseline estimates with the whole sample both qualitatively and quantitatively. Being on Twitter in a high penetration state leads to higher aggregate donations (columns (1) to (4)) as well the probability of getting at least one donation per

[18] It is also interesting to note that other social styles do not correlate with donations. We also analyze how emotional styles correlate with donations, focusing on the impact of 'worried emotional style'. A worried style does not correlate with donations raised for new politicians but it is negatively correlated for experienced politicians (see Table A8 in the Appendix).

week (columns (5) to (8)) but these effects hold only for new politicians and not the more experienced ones. This is exactly the takeaway from our main estimates with the full sample.

7.3 Excluding Campaign Periods

The main concern associated with our identification is other events such as campaign activities happening at the same time with opening a Twitter account which might be driving donations. While our placebo checks provide confidence that this indeed is not the case, we conduct another test to check the robustness of our results. Since elections take place in even numbered years (2010, 2012 and 2014 in our sample), it is likely that in the first half of each of the odd numbered years (2009, 2011, 2013) campaign activities would be limited. Hence, we re-estimate our diff-in-diff specifications with only the first six months of 2009, 2011 and 2013.

In Table A3 in the Appendix, we find that even when we focus only on this disconnected 18 month period, the effect of using Twitter in a high penetration state persists for new politicians (columns (2) and (5)) while results remain insignificant for the experienced ones (columns (3) and (6)). Putting both types of politicians together (columns (1) and (4)) leads to insignificant results, presumably because of a lack of power and limited variation since we only use a quarter of the entire data.

8 Conclusion

Electoral campaigns in the past couple years have seen a significant change in the communication channels used by the candidates to reach out to the electorate.[19] A notable change during this period was the intensified use of social media platforms to reach out and inform voters, partially eliminating dependence on the traditional media outlets such as newspapers and television. The essential question remains, does the use of social media accounts by politicians fundamentally alter any aspect of electoral politics? More broadly, can innovations in communication technologies change the way political markets operate? In this study, we document that a politician's entry on Twitter can help her to attract new donations. Overall, results imply that social media can help to democratize electoral politics by reducing the barriers for new politicians to raise money from the public.

Many avenues of future research lie at the intersection of adoption of new communication technologies and political outcomes. Future studies can expand the findings from our study to investigate the extent of substitution between the new and traditional media channels. For instance, we do not study how political advertising and use of social media may be complements or substitutes in delivering information about the candidates and their policies to voters. Further, a unique feature of social media is enabling two-way communication. One additional feature of social media may be the ability to listen to citizens' concerns and respond by policy proposals. Our results are not directly related to political polarization, but since new politicians are likely to avoid strong statements, they can explain why social media does not seem to be associated with much higher political polarization (Gentzkow et al., 2011; Halberstam and Knight, 2014).

[19] Andrews, Natalie and Rebecca Ballhaus, "Twitter Courts U.S. Presidential Campaigns With New Donations Service," *Wall Street Journal*, 2015.

Finally, in our study we focus on the effect of opening a new channel of communication on candidates' fund raising, but being on Twitter will also influence the politicians who are in the office. Some of the activities in office may be influenced by politicians' presence on channels like Twitter, since accounts which allow citizens to engage in communication may force the politicians to be more accountable. All of the listed are important questions, and future studies may consider addressing them.

References

Lorien C Abroms and R Craig Lefebvre. Obama's wired campaign: Lessons for public health communication. *Journal of health communication*, 14(5):415–423, 2009.

Daron Acemoglu, Tarek A Hassan, and Ahmed Tahoun. The power of the street: Evidence from egypt's arab spring. Technical report, National Bureau of Economic Research, 2014.

Maja Adena, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. Radio and the rise of the nazis in prewar germany. *Available at SSRN 2242446*, 2014.

Stephen Ansolabehere and James M Snyder Jr. Old voters, new voters, and the personal vote: Using redistricting to measure the incumbency advantage. *American Journal of Political Science*, 44(1):17–34, 2000.

Stephen Ansolabehere, Rebecca Lessem, and James M Snyder. The orientation of newspaper endorsements in us elections, 1940–2002. *Quarterly Journal of Political Science*, 1(4):393–404, 2006a.

Stephen Ansolabehere, Erik C Snowberg, and James M Snyder. Television and the incumbency advantage in us elections. *Legislative Studies Quarterly*, 31(4):469–490, 2006b.

Scott Ashworth. Campaign finance and voter welfare with entrenched incumbents. *American Political Science Review*, 100(01):55–68, 2006.

Jody C Baumgartner and Jonathan S Morris. Who wants to be my friend? obama, youth and social networks in the 2008 campaign. *Communicator-in-chief: How Barack Obama used new media technology to win the White House*, pages 51–66, 2010.

Timothy Besley, Torsten Persson, and Daniel M Sturm. Political competition, policy and growth: theory and evidence from the us. *The Review of Economic Studies*, 77(4):1329–1352, 2010.

Bryan K Bollinger, Michael Andrew Cohen, and Jiang Lai. Measuring asymmetric persistence and interaction effects of media exposures across platforms. *Available at SSRN 2342349*, 2013.

Danah Boyd and Alice Marwick. Social privacy in networked publics: Teens' attitudes, practices, and strategies. In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. International Communication Association, Boston, MA, 2011.

Danah M Boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. *In Online Communication and Collaboration: A Reader (Eds. Helen Donelan, Karen Kear, and Magnus Ramage)*, pages 261– 281, 2010.

Jamie L Carson, Erik J Engstrom, and Jason M Roberts. Candidate quality, the personal vote, and the incumbency advantage in congress. *American Political Science Review*, 101(02):289–301, 2007.

Marcos Chamon and Ethan Kaplan. The iceberg theory of campaign contributions: Political threats and interest group behavior. *American Economic Journal: Economic Policy*, 5(1):1–31, 2013.

Chun-Fang Chiang and Brian Knight. Media bias and influence: Evidence from newspaper endorsements. *The Review of economic studies*, 78(3):795–820, 2011.

Peter Clarke and Susan H Evans. *Covering Campaigns: Journalism in Congressional Elections*. Stanford University Press, 1983.

Stephen Coate. Political competition with campaign contributions and informative advertising. *Journal of the European Economic Association*, 2(5):772–804, 2004.

Christopher Cotton. Should we tax or cap political contributions? a lobbying model with policy favors and access. *Journal of Public Economics*, 93(7):831–842, 2009.

Christopher Cotton. Pay-to-play politics: Informational lobbying and contribution limits when money buys access. *Journal of Public Economics*, 96(3):369–386, 2012.

Gary W Cox and Jonathan N Katz. Why did the incumbency advantage in us house elections grow? *American Journal of Political Science*, 40(2):478–497, 1996.

Aron Culotta and Jennifer Cutler. Mining brand perceptions from twitter social networks. *Marketing Science, forthcoming*, 2016.

Stefano DellaVigna and Matthew Gentzkow. Persuasion: Empirical evidence. *Annu. Rev. Econ*, 2:643–69, 2010.

Stefano DellaVigna and Ethan Daniel Kaplan. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.

Allan Drazen, Nuno Limão, and Thomas Stratmann. Political contribution caps and lobby formation: Theory and evidence. *Journal of Public Economics*, 91(3):723–754, 2007.

Ruben Durante and Brian Knight. Partisan control, media bias, and viewer responses: Evidence from berlusconi's italy. *Journal of the European Economic Association*, 10(3):451–481, 2012.

Thomas Eisensee and David Strömberg. News droughts, news floods, and us disaster relief. *The Quarterly Journal of Economics*, 122(2):693–728, 2007.

Ruben Enikolopov, Maria Petrova, and Ekaterina Zhuravskaya. Media and political persuasion: Evidence from russia. *American Economic Review*, 101(7):3253–3285, 2011.

Ruben Enikolopov, Alexey Makarin, and Maria Petrova. Social media and protest participation: Evidence from russia. *Available at SSRN 2696236*, 2016.

Claudio Ferraz and Frederico Finan. Electoral accountability and corruption: Evidence from the audits of local governments. *American Economic Review*, 101(4):1274–1311, 2011.

Esther Gal-Or, Tansev Geylani, and Pinar Yildirim. The impact of advertising on media bias. *Journal of Marketing Research*, 49(1):92–99, 2012.

Vincenzo Galasso and Tommaso Nannicini. Competing on good politicians. *American Political Science Review*, 105(01):79–99, 2011.

Camilo Garcia-Jimeno and Pinar Yildirim. Matching pennies on the campaign trail: An empirical study of senate elections and media coverage. *University of Pennsylvania working paper*, 2015.

Matthew Gentzkow. Television and voter turnout. *The Quarterly Journal of Economics*, pages 931–972, 2006.

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

Matthew Gentzkow, Jesse M. Shapiro, and Michael Sinkinson. Competition and ideological diversity: Historical evidence from us newspapers. *American Economic Review*, 104(10):3073–3114, 2014.

Alan S Gerber and Donald P Green. The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, 94(03):653–663, 2000.

Edie N. Goldenberg and Michael W. Traugott. Congressional campaign effects on candidate recognition and evaluation. *Political Behavior*, 2(1):61–90, 1980.

Shiyang Gong, Juanjuan Zhang, Ping Zhao, and Xuping Jiang. Tweeting increases product demand. *MIT working paper*, 2015.

Gene M. Grossman and Elhanan Helpman. Electoral competition and special interest politics. *The Review of Economic Studies*, 63(2):265–286, 1996.

Gene M. Grossman and Elhanan Helpman. *Special interest politics*. MIT Press, Cambridge MA and London UK, 2001.

Vassia Gueorguieva. Voters, myspace, and youtube the impact of alternative communication channels on the 2006 election cycle and beyond. *Social Science Computer Review*, 26(3):288–300, 2008.

Yosh Halberstam and Brian G Knight. Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *NBER Working Paper*, (w20681), 2014.

Sounman Hong and Daniel Nadler. Does the early bird move the polls?: The use of the social media tool'twitter'by us politicians and its impact on public opinion. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, pages 182–186. ACM, 2011.

Gary C Jacobsen. The marginals never vanished: Incumbency and competition in elections to the us house of representatives, 1952–1982. *American Journal of Political Science*, 31(1):126–141, 1987.

Gary C Jacobson and Samuel Kernell. Strategy and choice in the 1982 congressional elections. *PS: Political Science & Politics*, 15(03):423–430, 1982.

Robert Jensen and Emily Oster. The power of tv: Cable television and women's status in india. *The Quarterly Journal of Economics*, 124(3):1057–1094, 2009.

Thomas J Johnson and David D Perlmutter. Introduction: the facebook election. *Mass Communication and Society*, 13(5):554–559, 2010.

Matthew James Kushin and Masahiro Yamamoto. Did social media really matter? college students' use of online media and political decision making in the 2008 election. *Mass Communication and Society*, 13(5): 608–630, 2010.

Eliana La Ferrara, Alberto Chong, and Suzanne Duryea. Soap operas and fertility: Evidence from brazil. *American Economic Journal: Applied Economics*, 4(4):1–31, 2012.

David S Lassen and Adam R Brown. Twitter: The electoral connection? *Social Science Computer Review*, page 0894439310382749, 2010.

Steven D Levitt and Catherine D Wolfram. Decomposing the sources of incumbency advantage in the us house. *Legislative Studies Quarterly*, 22(1):45–60, 1997.

Mitchell J Lovett and Richard Staelin. The role of paid and earned media in building entertainment brands: Reminding, informing, and enhancing enjoyment. 2012.

Liye Ma, Baohong Sun, and Sunder Kekre. The squeaky wheel gets the grease - an empirical analysis of customer voice and firm intervention on twitter. *Marketing Science*, 34(5):627–645, 2015.

Alice E Marwick et al. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133, 2011.

Filip Matejka, Guido Tabellini, et al. Electoral competition with rationally inattentive voters. Technical report, The Center for Economic Research and Graduate Education-Economics Institute, Prague, 2015.

Panagiotis T Metaxas and Mustafaraj Eni. Social media and the elections. *Science*, 338(6106):472–473, 2012.

Emily Metzgar and Albert Maruggi. Social media and the 2008 us presidential election. *Journal of New Communications Research*, 4(1):141–165, 2009.

Roger B Myerson. Incentives to cultivate favored minorities under alternative electoral systems. *American Political Science Review*, 87(04):856–869, 1993.

Torsten Persson, Guido Tabellini, and Francesco Trebbi. Electoral rules and corruption. *journal of the European Economic Association*, 1(4):958–989, 2003.

Andrea Prat. Campaign advertising and voter welfare. *The Review of Economic Studies*, 69(4):999–1017, 2002.

Andrea Prat, Riccardo Puglisi, James M Snyder, et al. Is private campaign finance a good thing? estimates of the potential informational benefits. *Quarterly journal of political science*, 5(3):291–318, 2010.

Markus Prior. The incumbent in the living room: The rise of television and the incumbency advantage in us house elections. *Journal of Politics*, 68(3):657–673, 2006.

Bei Qin. Chinese microblogs and drug quality. *Institute for International Economic Studies working paper*, 2013.

Bei Qin, David Strömberg, and Yanhui Wu. The political economy of social media in china. *Working paper*, 2016.

Stephan Seiler, Song Yao, and Wenbo Wang. The impact of earned media on demand: Evidence from a natural experiment. *Stanford University working paper*, 2016.

A. Smith. Why americans use social media. *Pew Internet & American Life Project*, 2011. URL `http://www.pewinternet.org/Reports/2011/WhyAmericans-Use-Social-Media/Main-report.aspx`.

A. Smith and L. Rainie. The internet and the 2008 election. *The Review of Economic Studies*, 2008. URL `http://pewinternet.org/~/media/Files/Reports/2008/PIP_2008_election.pdf`.

James M Snyder Jr and David Strömberg. Press coverage and political accountability. *Journal of Political Economy*, 118(2):355–408, 2010.

Jörg L Spenkuch and David Toniatti. Political advertising and election outcomes. *Available at SSRN 2613987*, 2016.

David Strömberg. Mass media competition, political competition, and public policy. *The Review of Economic Studies*, 71(1):265–284, 2004.

Terri L Towner and David A Dulio. New media and political marketing in the united states: 2012 and beyond. *Journal of Political Marketing*, 11(1-2):95–119, 2012.

Stefano Della Vigna, Ruben Enikolopov, Vera Mironova, Maria Petrova, and Ekaterina Zhuravskaya. Cross-border media and nationalism: Evidence from serbian radio in croatia. *American Economic Journal: Applied Economics*, 6(3):103–32, 2014.

Amardeep Virk et al. Twitter: The strength of weak ties. *University of Auckland Business Review*, 13(1): 19, 2011.

David Yanagizawa-Drott. Propaganda and conflict: Evidence from the rwandan genocide. *The Quarterly Journal of Economics*, 129(4):1947–1994, 2014.

Pinar Yildirim, Esther Gal-Or, and Tansev Geylani. User-generated content and bias in news media. *Management Science*, 59(12):2655–2666, 2013.

Weiwu Zhang, Thomas J Johnson, Trent Seltzer, and Shannon L Bichard. The revolution will be networked the influence of social networking sites on political attitudes and behavior. *Social Science Computer Review*, 28(1):75–92, 2010.

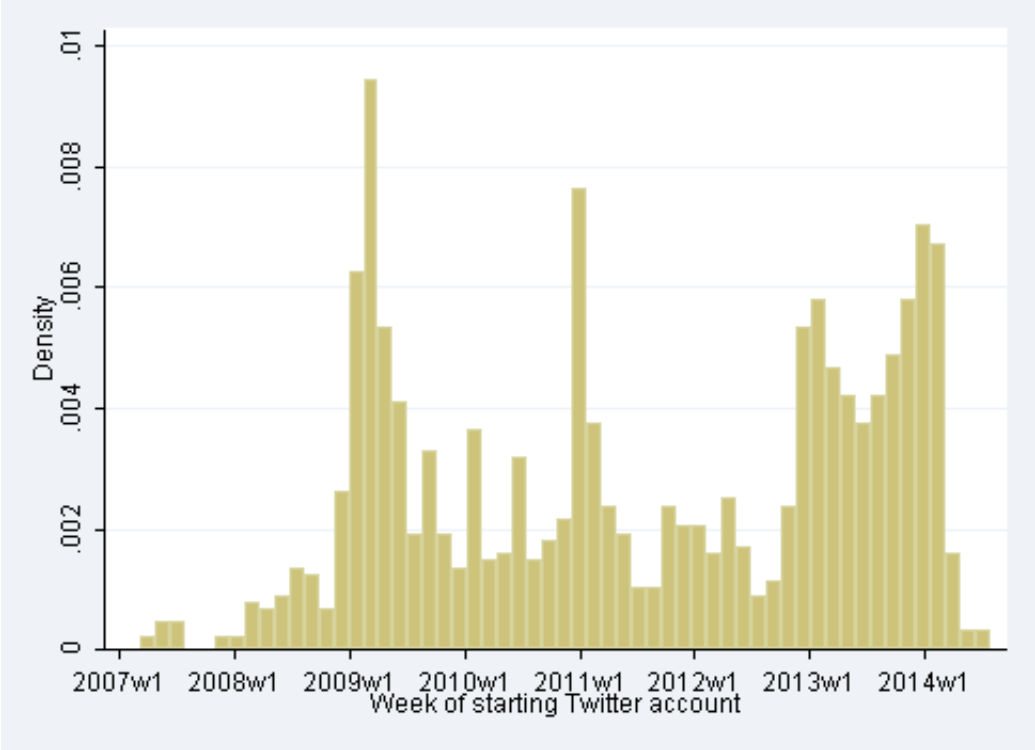

Figure 1: Date of Opening Accounts on Social Media

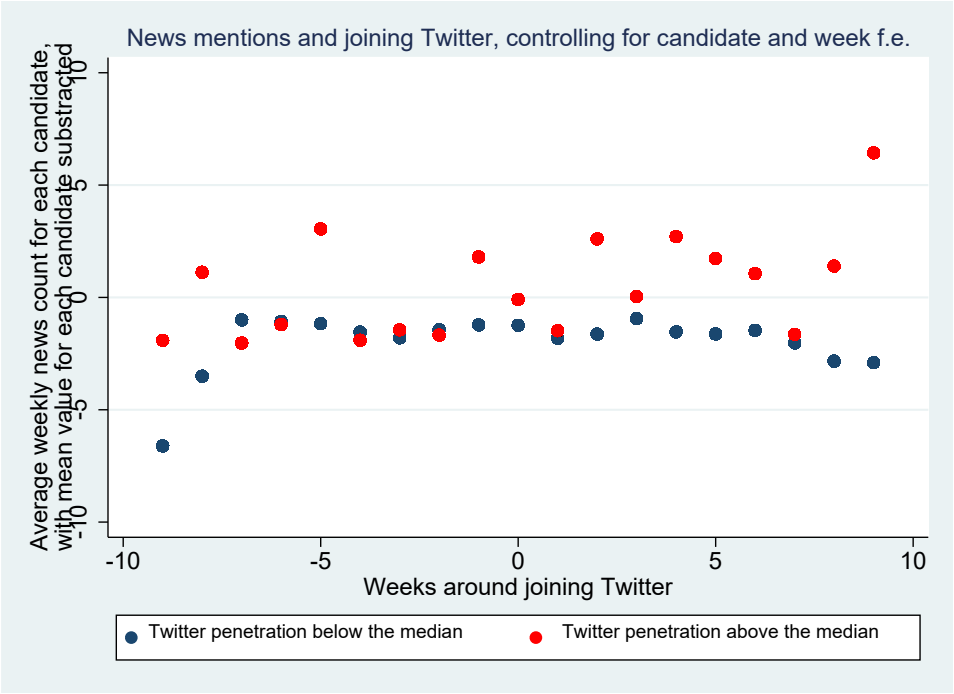

Figure 2: Number of News Mentions and Twitter Penetration

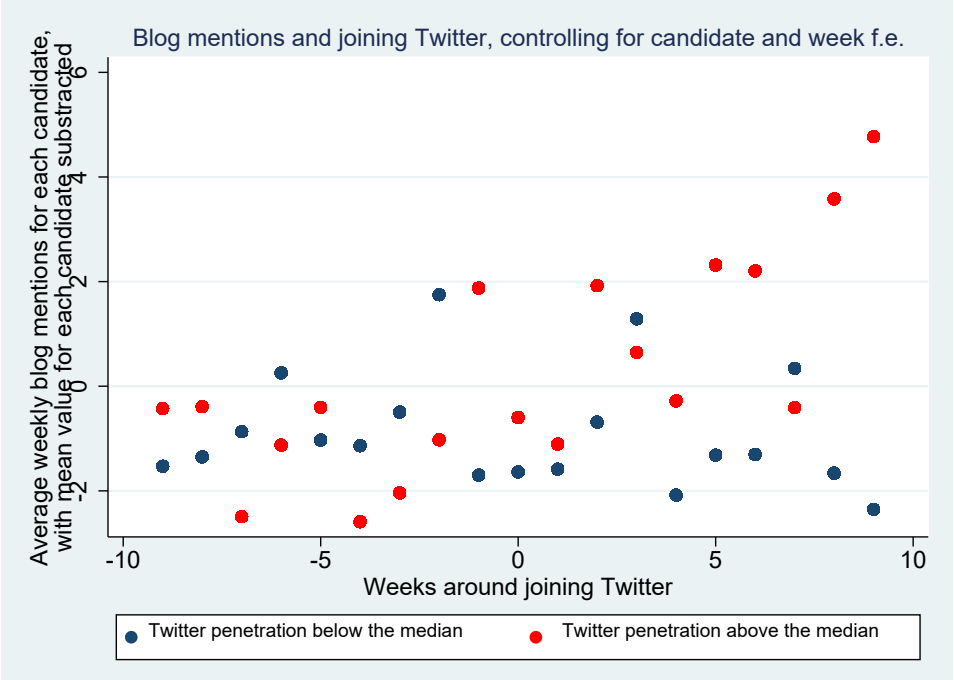

Figure 3: Number of Blog Mentions and Twitter Penetration

Table 1: Summary Statistics: All Politicians

Variable	Observations	Mean	Std. Dev.	Min	Max
	All Politicians				
Log(Aggregate Donations)	1,834	1.99	2.03	0	9.48
Probability of Donations	1,834	0.25	0.24	0	0.99
Log (Campaign Expenditures)	1,834	2.53	2.74	0	11.24
Number of News Mentions	1,834	10.52	265.31	0	11,281.43
Number of Blog Mentions	1,834	6.99	158.13	0	6641.90
Facebook Account Before	1,834	0.02	0.14	0	1
Log(Number of Tweets)	1,834	0.11	0.28	0	1.98
Log(Number of Retweets)	1,834	0.12	0.40	0	4.91
Log(Number of Favorites)	1,834	0.04	0.21	0	3.66
Log(Proportion of URLs)	1,834	0.03	0.07	0	0.52
Log(Proportion of words)	1,834	0.003	0.008	0	0.09

Table 2: Summary Statistics: New and Experienced Politicians

Variable	N	Mean	Std. Dev.	Min	Max	N	Mean	Std. Dev.	Min	Max	Difference
	New					Experienced					(Experienced-New)
Log(Aggregate Donations)	1,230	1.34	1.50	0	8.30	604	3.30	2.32	0	9.48	1.96 (0.90)***
Probability of Donations	1,230	0.17	0.18	0	0.89	604	0.41	0.28	0	0.99	0.24 (0.10)***
Log(Campaign Expenditure)	1,230	1.59	1.86	0	8.42	604	4.46	3.20	0	11.24	2.87(0.11)***
Number of News Mentions	1,230	4.93	37.77	0	946.85	604	30.41	526.51	0	11281.43	18.83(13.17)*
Number of Blog Mentions	1,230	3.91	36.43	0	780.52	604	18.77	310.72	0	6641.90	10.81(7.85)*
Facebook Account Before	1,230	0.01	0.09	0	0.98	604	0.05	0.20	0	1	0.03(0.006)***
Log(Number of Tweets)	1,230	0.09	0.26	0	1.89	604	0.13	0.32	0	1.98	0.03(0.01)**
Log(Number of Retweets)	1,230	0.09	0.30	0	2.50	604	0.18	0.54	0	4.91	0.08(0.01)***
Log(Number of Favorites)	1,230	0.02	0.11	0	1.16	604	0.09	0.32	0	3.66	0.06(0.01)***
Log(Proportion of URLs)	1,230	0.02	0.07	0	0.50	604	0.03	0.09	0	0.52	0.009(0.003)*
Log(Proportion of Words)	1,230	0.003	0.008	0	0.09	604	0.004	0.01	0	0.66	0.001(0.0004)***
Log(Proportion of 'I')	1,230	0.01	0.001	0	0.32	604	0.02	0.002	0	0.27	0.005(0.002)***
Log(Proportion of 'We')	1,230	0.003	0.009	0	0.07	604	0.004	0.01	0	0.08	0.0008(0.0004)**
'Plugged In' Score	998	51.59	18.66	0	98	647	54.17	19.62	0	100	-2.579(0.961)***
'Analytic' Score	998	40.01	15.84	8	100	647	40.78	18.24	8	100	0.77(0.874)

Table 3: Joining Twitter and Aggregate Donations: Baseline Estimates

	Log (Aggregate donations)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	All Politicians					New	Experienced
on Twitter x Twitter penetration	-0.340***	0.359**	0.353**	0.349**	0.378**	0.692***	-0.217
	(0.117)	(0.148)	(0.147)	(0.147)	(0.144)	(0.169)	(0.256)
on Twitter	1.312***	0.435***	0.406***	0.161	0.700	-3.185	7.496*
	(0.116)	(0.104)	(0.103)	(0.100)	(2.404)	(3.268)	(4.450)
Log (campaign expenditure)			0.094***	0.091***	0.091***	0.121***	0.079***
			(0.004)	(0.004)	(0.004)	(0.007)	(0.004)
Politician-Month Fixed Effects		Yes	Yes	Yes	Yes	Yes	Yes
Time trend				Week	Week	Week	Week
Baseline controls x on Twitter					Yes	Yes	Yes
Observations	565,968	565,968	565,764	565,764	565,764	236,700	329,064
R-squared	0.019	0.820	0.821	0.823	0.823	0.885	0.787

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1) - (5) include all politicians while column (6) includes only new and column (7) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 4: Joining Twitter and the Probability and Number of Donations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
on Twitter x Twitter penetration	0.047**	0.051**	0.084***	-0.014	0.0563	0.0633	0.155***	-0.107
	(0.019)	(0.019)	(0.024)	(0.034)	(0.0411)	(0.0418)	(0.0450)	(0.0701)
on Twitter	0.021	0.023	-0.476	0.876	0.0469*	0.158	-0.931	2.092*
	(0.014)	(0.343)	(0.451)	(0.617)	(0.0276)	(0.546)	(0.708)	(1.147)
Log (campaign expenditure)	0.011***	0.011***	0.015***	0.009***	0.0265***	0.0265***	0.0353***	0.0232***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.00101)	(0.00101)	(0.00208)	(0.000973)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	565,764	565,764	236,700	329,064	565,764	565,764	236,700	329,064
R-squared	0.788	0.788	0.847	0.752	0.840	0.840	0.902	0.802

Note: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the probability of receiving at least one donation in a week in columns (1) - (4) while it is the total number of donations in a week in columns (5)-(8). Columns (1) - (2) and (5)-(6) include all politicians while columns (3) and (6) includes only new and columns (4) and (8) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 5: Joining Twitter (Donations between $1000 and $3000)

	Log(Aggregate donations)				At least one donation per week			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
on Twitter x Twitter penetration	0.205	0.236	0.573***	-0.380	0.030	0.033*	0.067***	-0.029
	(0.151)	(0.147)	(0.180)	(0.316)	(0.019)	(0.019)	(0.022)	(0.041)
on Twitter	0.262***	-1.125	-2.778	2.489	0.027**	-0.143	-0.317	0.223
	(0.097)	(2.346)	(2.368)	(4.194)	(0.013)	(0.269)	(0.275)	(0.515)
Log (campaign expenditure)	0.098***	0.098***	0.130***	0.086***	0.010***	0.010***	0.014***	0.009***
	(0.004)	(0.004)	(0.007)	(0.004)	(0.000)	(0.000)	(0.001)	(0.001)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	565,764	565,764	236,700	329,064	565,764	565,764	236,700	329,064
R-squared	0.759	0.759	0.826	0.722	0.727	0.727	0.791	0.690

Note: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in columns (1)-(4) and the probability of getting at least one donation in columns (5)-(8). Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) includes only new ones and columns (4) and (8) has the experienced politicians. Baseline controls, at the level of the state, are interacted with the politician being on Twitter, include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 6: Joining Twitter and Campaign Expenditures

	Log (campaign expenditures)					
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	All politicians				New	Experienced
on Twitter x Twitter penetration	-0.297**	0.036	0.035	0.063	0.233	-0.261
	(0.141)	(0.156)	(0.157)	(0.154)	(0.220)	(0.227)
on Twitter	1.481***	0.352***	0.269**	1.177	3.039	-1.230
	(0.145)	(0.105)	(0.106)	(2.550)	(3.321)	(2.810)
Politician-Month Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time trend			Week	Week	Week	Week
Baseline controls x on Twitter				Yes	Yes	Yes
Observations	565,764	565,764	565,764	565,764	236,700	329,064
R-squared	0.022	0.888	0.888	0.888	0.896	0.876

Notes. Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of total campaign expenditures incurred in a week. Columns (1)-(4) includes all politicians while column (5) includes only new ones while column (6) has the experienced politicians. Baseline controls, at the level of the state, are interacted with the politician being on Twitter, include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 7: Joining Twitter and Political Advertising

	Log (Political Advertising)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	All Politicians					New	Experienced
on Twitter x Twitter penetration	-0.068***	-0.052	-0.053	-0.053	-0.060	0.025	-0.218*
	[0.021]	[0.037]	[0.038]	[0.038]	[0.038]	[0.054]	[0.126]
on Twitter	0.065***	0.033	0.030	0.014	-0.802	-1.200	0.075
	[0.017]	[0.026]	[0.026]	[0.026]	[0.592]	[0.790]	[0.674]
Log (campaign expenditure)			0.008***	0.008***	0.008***	0.010***	0.008***
			[0.001]	[0.001]	[0.001]	[0.002]	[0.001]
Politician-Month Fixed Effects		Yes	Yes	Yes	Yes	Yes	Yes
Time trend				Week	Week	Week	Week
Baseline controls x on Twitter					Yes	Yes	Yes
Observations	565,968	565,968	565,764	565,764	565,764	236,700	329,064
R-squared	0.000	0.813	0.814	0.814	0.814	0.830	0.802

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1) - (5) include all politicians while column (6) includes only new and column (7) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 8: News and Blogs Coverage

	#News				#Blogs			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
on Twitter x Twitter penetration	-0.735	-0.627	0.217	-2.094	-0.518	-0.595	-0.0740	-1.528
	(1.051)	(0.966)	(0.254)	(2.984)	(1.033)	(0.918)	(0.151)	(2.776)
on Twitter	-0.164	8.073	0.0824	29.85	-0.179	1.638	-1.984	13.30
	(0.167)	(12.68)	(3.048)	(34.11)	(0.671)	(10.35)	(1.327)	(30.33)
Log (campaign expenditure)		-0.208	-0.586	0.132		-0.0177	-0.0626	0.0248
		(0.277)	(0.729)	(0.150)		(0.108)	(0.207)	(0.0843)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Baseline controls x on Twitter	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	47,375	47,356	28,947	18,409	47,375	47,356	28,947	18,409
R-squared	0.825	0.825	0.514	0.865	0.935	0.935	0.654	0.946

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the number of news mentions in columns (1)-(4) and the number of blog mentions in columns (5)-(8). Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) include only new and columns (4) and (8) have the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 9: Demographic Characteristics and Donations

	Log (aggregate donations per week)					
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
on Twitter x penetration	0.349**					
	(0.147)					
on Twitter x median household income		-0.017				
		(0.015)				
on Twitter x share of rich			-0.031			
			(0.083)			
on Twitter x share of those with college education				-2.175		
				(1.839)		
on Twitter x vote share of Bush in 2004					0.164	
					(0.858)	
on Twitter x share of African Americans						0.279
						(1.024)
on Twitter	0.161	1.112*	0.480**	2.161	0.323	0.374***
	(0.100)	(0.622)	(0.216)	(1.489)	(0.417)	(0.129)
Log(campaign expenditures)	0.091***	0.091***	0.091***	0.091***	0.091***	0.091***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week
Observations	565,764	565,764	565,764	565,764	565,764	565,764
R-squared	0.823	0.823	0.823	0.823	0.823	0.823

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of total donations received in a week.

Table 10: Joining Twitter and donations from new and old donors

Panel A. Donations from new donors.

	Log (aggregate donations)				At least one donation per week			
	All	All	New	Experienced	All	All	New	Experienced
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
on Twitter	0.335***	0.343***	0.725***	-0.364*	0.045***	0.045***	0.090***	-0.039
	(0.119)	(0.119)	(0.159)	(0.208)	(0.016)	(0.016)	(0.022)	(0.029)
on Twitter x Twitter penetration	0.129	-0.929	-3.585	4.497	0.016	-0.172	-0.582	0.617
	(0.086)	(2.239)	(3.289)	(3.735)	(0.013)	(0.310)	(0.441)	(0.525)
Log (campaign expenditures)	0.069***	0.069***	0.094***	0.060***	0.008***	0.008***	0.011***	0.007***
	(0.004)	(0.004)	(0.007)	(0.004)	(0.001)	(0.001)	(0.001)	(0.001)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline Controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	565,764	565,764	236,700	329,064	565,764	565,764	236,700	329,064
R-squared	0.786	0.786	0.874	0.734	0.749	0.749	0.838	0.699

Panel B. Donations from new donors.

	Log (aggregate donations)				At least one donation per week			
	All	All	New	Experienced	All	All	New	Experienced
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
on Twitter	-0.085	-0.106	0.022	-0.319	-0.010	-0.013	0.002	-0.041
	(0.149)	(0.135)	(0.079)	(0.293)	(0.021)	(0.019)	(0.012)	(0.040)
on Twitter x Twitter penetration	0.108	2.679	-0.422	7.157*	0.018	0.380	0.047	0.838
	(0.100)	(2.046)	(1.547)	(3.778)	(0.014)	(0.279)	(0.242)	(0.516)
Log (campaign expenditures)	0.071***	0.071***	0.078***	0.068***	0.009***	0.009***	0.011***	0.009***
	(0.003)	(0.003)	(0.005)	(0.003)	(0.000)	(0.000)	(0.001)	(0.000)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	565,764	565,764	236,700	329,064	565,764	565,764	236,700	329,064
R-squared	0.764	0.764	0.810	0.737	0.731	0.731	0.766	0.705

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of total donation in a week from new donors in Panel A and from old donors in Panel B. In both panels, Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) include only new and columns (4) and (8) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 11: Joining Twitter, Donations, and Newspaper Circulation

Panel A. Donations in Low Circulation States

	Log (aggregate donations)				At least one donation per week			
	All	All	New	Experienced	All	All	New	Experienced
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
on Twitter x Twitter penetration	0.456**	0.489**	0.835***	-0.015	0.062**	0.063**	0.103***	0.003
	(0.224)	(0.229)	(0.244)	(0.433)	(0.030)	(0.031)	(0.034)	(0.060)
on Twitter	0.065	2.330	0.181	7.122	0.005	0.229	0.036	0.670
	(0.187)	(4.370)	(5.645)	(7.064)	(0.026)	(0.616)	(0.804)	(0.984)
Log (campaign expenditures)	0.089***	0.089***	0.120***	0.078***	0.010***	0.010***	0.014***	0.009***
	(0.006)	(0.006)	(0.012)	(0.007)	(0.001)	(0.001)	(0.002)	(0.001)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	229,556	229,556	95,831	133,725	229,556	229,556	95,831	133,725
R-squared	0.809	0.809	0.882	0.768	0.774	0.774	0.845	0.732

Panel B. Donations in High Circulation States

	Log (aggregate donations)				At least one donation per week			
	All	All	New	Experienced	All	All	New	Experienced
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
on Twitter x Twitter penetration	0.234	0.248	0.544**	-0.347	0.031	0.034	0.064*	-0.029
	(0.234)	(0.236)	(0.269)	(0.416)	(0.032)	(0.032)	(0.038)	(0.056)
on Twitter	0.248	0.271	-3.052	7.426	0.034	0.057	-0.451	1.067
	(0.168)	(3.194)	(3.692)	(5.996)	(0.023)	(0.448)	(0.529)	(0.810)
Log (campaign expenditures)	0.092***	0.092***	0.122***	0.080***	0.011***	0.011***	0.015***	0.010***
	(0.005)	(0.005)	(0.008)	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	336,208	336,208	140,869	195,339	336,208	336,208	140,869	195,339
R-squared	0.832	0.832	0.887	0.799	0.797	0.797	0.848	0.766

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of total donation in a week from new donors in Panel A and from old donors in Panel B. In both panels, Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) include only new and columns (4) and (8) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 12: Politicians' Tweets and ReTweets

	Log (aggregate donations)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
on Twitter x Twitter penetration x log(tweets)	0.603	0.465	4.285**	0.262				
	(0.546)	(0.566)	(1.644)	(0.623)				
on Twitter x Twitter penetration x log(retweets)					0.111	-0.0271	3.087***	0.0192
					(0.365)	(0.338)	(0.749)	(0.390)
on Twitter	0.371***	0.391	-3.710	7.392	0.423***	0.532	-3.473	7.416
	(0.103)	(2.422)	(3.260)	(4.481)	(0.104)	(2.455)	(3.327)	(4.498)
on Twitter x Twitter penetration	0.399**	0.417***	0.748***	-0.199	0.363**	0.379**	0.697***	-0.213
	(0.150)	(0.145)	(0.175)	(0.253)	(0.149)	(0.145)	(0.171)	(0.259)
Log(campaign expenditure)		0.105***	0.119***	0.0916***		0.105***	0.119***	0.0916***
		(0.00720)	(0.0124)	(0.00954)		(0.00718)	(0.0124)	(0.00954)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend		Week	Week	Week		Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	78,107	78,082	46,964	31,118	78,107	78,082	46,964	31,118
R-squared	0.792	0.797	0.843	0.727	0.792	0.797	0.843	0.727

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) include only new and columns (4) and (8) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 13: Tweet Content: Number of URLs and Number of Use of Word "We"

	Log (aggregate donations)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
on Twitter x Twitter penetration x log(links)	-0.544	-1.207	13.45***	-1.198				
	(1.065)	(1.063)	(1.545)	(1.201)				
on Twitter x Twitter penetration x log(#we)					1.778	1.986	73.56***	-9.190
					(11.30)	(11.33)	(2.590)	(6.633)
on Twitter x Twitter penetration	0.382**	0.399***	0.719***	-0.196	0.360**	0.377**	0.693***	-0.212
	(0.150)	(0.145)	(0.171)	(0.258)	(0.150)	(0.145)	(0.171)	(0.258)
on Twitter	0.398***	0.502	-3.395	7.299	0.432***	0.590	-3.296	7.322
	(0.104)	(2.422)	(3.314)	(4.478)	(0.105)	(2.432)	(3.301)	(4.494)
log(campaign expenditure)		0.105***	0.119***	0.0917***		0.106***	0.119***	0.0915***
		(0.00721)	(0.0124)	(0.00956)		(0.00718)	(0.0124)	(0.00952)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend		Week	Week	Week		Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	78,107	78,082	46,964	31,118	78,107	78,082	46,964	31,118
R-squared	0.792	0.797	0.843	0.727	0.792	0.797	0.843	0.727

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) include only new and columns (4) and (8) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table 14: Tweet Sentiment: "Plugged In"

	Log (aggregate donations)			
	(1)	(2)	(3)	(4)
VARIABLES	All	All	New	Experienced
on Twitter x Twitter penetration x log(pluggedin)	0.612	0.565	0.632*	0.553
	(0.386)	(0.384)	(0.370)	(0.697)
on Twitter x log(pluggedin)	0.168	0.190	0.0134	0.373
	(0.363)	(0.367)	(0.312)	(0.741)
on Twitter x Twitter penetration	1.176*	1.128*	1.528**	0.604
	(0.625)	(0.631)	(0.622)	(1.099)
on Twitter	0.700	2.586	-2.316	9.806**
	(0.556)	(2.572)	(3.463)	(4.586)
log(campaign expenditure)		0.107***	0.121***	0.0937***
		(0.00742)	(0.0131)	(0.00987)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes
Time trend		Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes
Observations	70,260	70,239	41,999	28,240
R-squared	0.790	0.794	0.845	0.720

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1)-(2) include all politicians while column (3) includes only new and column (4) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A1: Joining Twitter, Aggregate Donations, and Different Window Size Specifications

VARIABLES	Log (aggregate donations)				
	(1)	(2)	(3)	(4)	(5)
Window size	±5 weeks	±10 weeks	±25 weeks	±50 weeks	±300 weeks
on Twitter x penetration	0.371**	0.373**	0.376**	0.377**	0.378**
	(0.152)	(0.148)	(0.145)	(0.144)	(0.144)
on Twitter	0.347	0.496	0.598	0.635	0.702
	(2.586)	(2.501)	(2.436)	(2.416)	(2.404)
Log (campaign expenditure)	0.144***	0.137***	0.106***	0.097***	0.091***
	(0.016)	(0.010)	(0.007)	(0.006)	(0.005)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week
Baseline controls x on Twitter	Yes	Yes	Yes	Yes	Yes
Observations	14,562	30,341	75,203	144,110	507,537
R-squared	0.761	0.767	0.796	0.805	0.818

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of total donations received in a week. Baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A2: Joining Twitter and Donations without 2009

	Log (aggregate donations per week)				At least one donation per week			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
on Twitter x Twitter penetration	0.589 ***	0.706***	0.932***	0.163	0.083***	0.098***	0.123***	0.037
	(0.183)	(0.187)	(0.230)	(0.348)	(0.024)	(0.026)	(0.033)	(0.049)
on Twitter	-.183	-2.649	-8.584*	7.128	-0.030	-0.531	-1.350**	0.784
	(0.186)	(3.505)	(4.547)	(6.443)	(0.027)	(0.513)	(0.608)	(0.882)
Log (campaign expenditure)	.089***	0.089***	0.116***	0.079***	0.010***	0.011***	0.014***	0.009***
	(0.004)	(0.005)	(0.009)	(0.005)	(0.0006)	(0.001)	(0.001)	(0.001)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	471,467	471,467	172,989	298,478	471,467	471,467	172,989	298,478
R-squared	0.826	0.827	0.886	0.797	0.79	0.79	0.847	0.762

Note: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate weekly donations in columns (1)-(4) and the probability of receiving at least one donation in columns (5)-(8). This considers a sub-sample without 2009. Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) include only new and columns (8) and (8) only the experienced politicians. Baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A3: Joining Twitter Outside Campaign Periods

				Log (aggregate donations)		
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	All	New	Experienced	All	New	Experienced
on Twitter	0.505	1.124***	-0.185	0.049	0.126**	-0.033
	(0.318)	(0.405)	(0.454)	(0.046)	(0.057)	(0.065)
on Twitter x penetration	2.353	2.048	3.030	0.155	0.255	0.128
	(2.780)	(2.953)	(5.406)	(0.422)	(0.376)	(0.781)
Log (campaign expend)	0.095***	0.181***	0.081***	0.011***	0.021***	0.009***
	(0.007)	(0.019)	(0.006)	(0.001)	(0.003)	(0.001)
Politician-Month FE		Yes	Yes	Yes	Yes	Yes
Time trend				Week	Week	Week
Baseline controls					Yes	Yes
Observations	141,424	61,981	79,443	141,424	61,981	79,443
R-squared	0.794	0.881	0.757	0.761	0.841	0.721

Note: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate weekly donations. This considers politicians joining Twitter outside of campaign periods. Columns (1) and (4) include all politicians while columns (2) and (4) include only the new and columns (3) and (6) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A4: Joining Twitter and Facebook Accounts

	Joined Facebook Before						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	All Politicians					New	Experienced
on Twitter x penetration	0.00708	-0.00160	-0.00160	-0.00161	-0.00220*	-0.000424	-0.00532
	(0.00498)	(0.00105)	(0.00105)	(0.00105)	(0.00130)	(0.000503)	(0.00350)
on Twitter	0.0298***	0.00246*	0.00245	0.00213	0.00960	0.00103	0.0244
	(0.00770)	(0.00146)	(0.00146)	(0.00146)	(0.0121)	(0.00505)	(0.0325)
Log (campaign expend)			2.62e-05	2.16e-05	2.17e-05	5.98e-05	7.59e-06
			(2.21e-05)	(2.18e-05)	(2.18e-05)	(5.95e-05)	(1.98e-05)
Politician-Month Fixed Effects		Yes	Yes	Yes	Yes	Yes	Yes
Time trend				Week	Week	Week	Week
Baseline controls x on Twitter					Yes	Yes	Yes
Observations	565,968	565,968	565,764	565,764	565,764	236,700	329,064
R-squared	0.012	0.996	0.996	0.996	0.996	0.993	0.997

Note: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is whether the politician joined Facebook before joining Twitter. Columns (1)-(5) include all politicians while column (6) includes only the new and column (7) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A5: Twitter Entry, Demographics, and Twitter Penetration: Levels

	Twitter Penetration: Levels					
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
onTwitter	-0.00003	-0.000145	-2.11e-05	-0.000184	-2.59e-05	1.43e-05
	(0.00003)	(0.0002)	(7.66e-05)	(0.0005)	(0.0001)	(4.17e-05)
on Twitter x median household income		4.16e-06				
		(5.28e-06)				
on Twitter x share of rich			2.27e-05			
			(3.15e-05)			
on Twitter x share of those with college education				0.0002		
				(0.0006)		
on Twitter x vote share of Bush in 2004					0.0001	
					(0.0002)	
on Twitter x share of African Americans						(0.0001)
						(0.0002)
Log (campaign expenditures)	1.80e-06	1.80e-06	1.80e-06	1.80e-06	1.80e-06	1.80e-06
	(1.50e-06)	(1.50e-06)	(1.50e-06)	(1.50e-06)	(1.50e-06)	(1.50e-06)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week
Observations	565,764	565,764	565,764	565,764	565,764	565,764
R-squared	0.929	0.929	0.929	0.929	0.929	0.929

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of weekly Twitter penetration.

Table A6: Twitter Entry, Demographics and Twitter Penetration: First Differences

	Twitter Penertation: First Difference					
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
on Twitter	-0.00001	-9.08e-05	2.36e-05	-0.0007	-0.0002	-4.84e-05
	(0.009)	(0.0002)	(0.0001)	(0.0007)	(0.00003)	(6.68e-05)
on Twitter x median household income		1.77e-06				
		(6.70e-06)				
on Twitter x share of rich			-1.63e-05			
			(4.30e-05)			
on Twitter x share of those with college education				0.0008		
				(0.0009)		
on Twitter x vote share of Bush in 2004					0.0003	
					(0.0004)	
on Twitter x share of African Americans						0.0002
						(0.0004)
Log (campaign expenditures)	1.61e-06	1.61e-06	1.61e-06	1.61e-06	1.61e-06	1.62e-06
	(1.72e-06)	(1.74e-06)	(1.74e-06)	(1.74e-06)	(1.74e-06)	(1.74e-06)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week
Observations	563,951	563,951	563,951	563,951	563,951	563,951
R-squared	0.104	0.105	0.105	0.105	0.105	0.105

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the weekly Twitter penetration in first differences.

Table A7: Politician Tweets: Sentiment analysis and Tweet Content

	Log (aggregate donations)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
onTwitter x Twitter penetration x log(analytic)	0.374	0.474	0.285	0.884				
	(0.652)	(0.663)	(0.739)	(1.111)				
onTwitter x Twitter penetration x log(#I)					-1.108	-1.509	-2.332	-2.227
					(1.717)	(1.714)	(3.611)	(2.412)
onTwitter x log(analytic)	-0.412	-0.441	-0.361	-0.620				
	(0.445)	(0.436)	(0.562)	(0.746)				
onTwitter x log(#I)					2.023	2.474	1.643*	3.335
					(1.748)	(1.845)	(0.843)	(2.663)
Twitter penetration x log(#I)					1.030	1.389	3.444	1.936
					(1.685)	(1.698)	(2.960)	(2.345)
log(#I)					-1.724	-2.180	-1.843**	-2.795
					(1.699)	(1.816)	(0.766)	(2.536)
onTwitter	0.0138	1.155	-3.777	9.125*	0.424***	0.555	-3.309	7.308
	(0.467)	(2.724)	(3.618)	(4.586)	(0.104)	(2.431)	(3.296)	(4.511)
onTwitter x Twitter penetration	0.707	0.838	0.936	0.757	0.363**	0.380**	0.691***	-0.209
	(0.638)	(0.653)	(0.745)	(1.053)	(0.150)	(0.145)	(0.171)	(0.261)
Log(campaign expenditure)		0.106***	0.121***	0.0922***		0.106***	0.119***	0.0917***
		(0.00749)	(0.0131)	(0.00975)		(0.00718)	(0.0124)	(0.00954)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend		Week	Week	Week		Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	70,954	70,932	42,405	28,527	78,107	78,082	46,964	31,118
R-squared	0.789	0.794	0.844	0.720	0.792	0.797	0.843	0.727

Note: Robust standard errors clustered at the level of the state in parenthesis. {*}{*}{*} p$<$0.01, {*}{*} p$<$0.05, {*} p$<$0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) includes only new ones and columns (4) and (8) has the experienced politicians. Baseline controls, at the level of the state, are interacted with the politician being on Twitter, include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over \$250,000 a year, the median household income as well as the share of the population with a college degree.

Table A8: Tweet Sentiment: 'Worried'

	Log (aggregate donations)			
	(1)	(2)	(3)	(4)
VARIABLES	All	All	New	Experienced
on Twitter x Twitter penetration x log(worried)	-0.458	-0.453	0.479	-1.674*
	(0.542)	(0.521)	(0.568)	(0.838)
onTwitter x log(worried)	0.116	0.108	-0.309	0.646
	(0.358)	(0.346)	(0.352)	(0.723)
on Twitter x Twitter penetration	-0.469	-0.437	1.439	-3.033**
	(0.870)	(0.842)	(0.942)	(1.409)
on Twitter	0.651	1.579	-3.662	9.939**
	(0.588)	(2.833)	(3.442)	(4.782)
log(campaign expenditure)		0.106***	0.121***	0.0924***
		(0.00750)	(0.0131)	(0.00974)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes
Time trend		Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes
Observations	70,954	70,932	42,405	28,527
R-squared	0.789	0.794	0.844	0.720

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1)-(2) include all politicians while column (3) includes only new and column (4) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A9: Joining Twitter and Donations: Democrats vs. Republicans

Panel A. Donations to Democrats.

	Log (aggregate donations)				At least one donation per week			
	All	All	New	Experienced	All	All	New	Experienced
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
on Twitter x Twitter penetration	0.648**	0.605**	1.024***	-0.154	0.085***	0.075**	0.119***	-0.005
	(0.251)	(0.255)	(0.282)	(0.458)	(0.029)	(0.030)	(0.039)	(0.052)
on Twitter	0.071	-1.995	-6.551	5.312	0.003	-0.412	-0.988	0.486
	(0.185)	(4.052)	(4.924)	(6.135)	(0.024)	(0.554)	(0.676)	(0.761)
Log (campaign expenditures)	0.087***	0.087***	0.117***	0.077***	0.010***	0.010***	0.014***	0.009***
	(0.006)	(0.006)	(0.011)	(0.007)	(0.001)	(0.001)	(0.002)	(0.001)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline Controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	234,823	234,823	92,218	142,605	234,823	234,823	92,218	142,605
R-squared	0.827	0.827	0.904	0.787	0.790	0.790	0.869	0.748

Panel B. Donations to Republicans.

	Log (aggregate donations)				At least one donation per week			
	All	All	New	Experienced	All	All	New	Experienced
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
on Twitter x Twitter penetration	0.133	0.188	0.414*	-0.248	0.021	0.029	0.054*	-0.019
	(0.214)	(0.215)	(0.221)	(0.410)	(0.030)	(0.030)	(0.032)	(0.056)
on Twitter	0.212	2.755	-1.257	10.294*	0.031	0.372	-0.168	1.398
	(0.154)	(3.643)	(4.171)	(6.083)	(0.022)	(0.525)	(0.611)	(0.862)
Log (campaign expenditures)	0.093***	0.093***	0.123***	0.081***	0.011***	0.011***	0.015***	0.010***
	(0.005)	(0.005)	(0.008)	(0.006)	(0.001)	(0.001)	(0.001)	(0.001)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	330,941	330,941	144,482	186,459	330,941	330,941	144,482	186,459
R-squared	0.818	0.818	0.871	0.785	0.785	0.785	0.832	0.754

Note: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of total donation in a week for Democratic candidates in Panel A and for Republicans in Panel B. In both panels, Columns (1)- (2) and (5)- (6) includes all politicians while columns (3) and (7) includes only new ones and columns (4) and (8) has the experienced politicians. Baseline controls, at the level of the state, are interacted with the politician being on Twitter, include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A10: Joining Twitter and Aggregate Donations ($3000-$5000)

VARIABLES	(1) All	(2) All	(3) New	(4) Experienced	(5) All	(6) All	(7) New	(8) Experienced
on Twitter x Twitter penetration	0.018	0.037	0.033	0.039	0.003	0.005	0.005	0.005
	(0.048)	(0.049)	(0.061)	(0.095)	(0.006)	(0.006)	(0.007)	(0.011)
on Twitter	-0.001	-0.856	-1.305	-0.143	-0.001	-0.085	-0.132	-0.012
	(0.045)	(0.703)	(0.957)	(0.870)	(0.005)	(0.077)	(0.102)	(0.102)
Log (campaign expenditure)	0.012***	0.012***	0.019***	0.010***	0.001***	0.001***	0.002***	0.001***
	(0.001)	(0.001)	(0.003)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	565,764	565,764	236,700	329,064	565,764	565,764	236,700	329,064
R-squared	0.539	0.539	0.572	0.523	0.518	0.518	0.547	0.502

Note: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in columns (1)-(4) and the probability of getting at least one donation in columns (5)-(8). Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) include only the new and columns (4) and (8) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter, include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A11: Joining Twitter and Aggregate Donations (Above $5000)

VARIABLES	(1) All	(2) All	(3) New	(4) Experienced	(5) All	(6) All	(7) New	(8) Experienced
on Twitter x Twitter penetration	0.011	0.024	0.008	0.046	0.002	0.003	0.001	0.006
	(0.045)	(0.047)	(0.054)	(0.095)	(0.005)	(0.005)	(0.006)	(0.011)
on Twitter	-0.009	-1.048	-1.091	-1.055	-0.001	-0.102	-0.111	-0.097
	(0.043)	(0.877)	(1.055)	(1.144)	(0.005)	(0.094)	(0.115)	(0.117)
Log (campaign expenditure)	0.016***	0.016***	0.031***	0.011***	0.002***	0.002***	0.003***	0.001***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	565,764	565,764	236,700	329,064	565,764	565,764	236,700	329,064
R-squared	0.591	0.591	0.604	0.584	0.566	0.566	0.583	0.558

Note: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in columns (1)-(4) and the probability of getting at least one donation in columns (5)-(8). Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) include only the new and columns (4) and (8) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A12: Joining Twitter and Aggregate Donations: Baseline Estimates with Week Fixed Effects

	Log (Aggregate donations)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	All Politicians					New	Experienced
on Twitter x Twitter penetration	-0.340***	0.359**	0.353**	0.354**	0.374**	0.520***	0.155
	(0.117)	(0.148)	(0.147)	(0.146)	(0.142)	(0.172)	(0.238)
on Twitter	1.312***	0.435***	0.406***	0.148	0.915	-2.934	7.154
	(0.116)	(0.104)	(0.103)	(0.0978)	(2.430)	(3.280)	(4.280)
Log (campaign expenditure)			0.0941***	0.0852***	0.0852***	0.118***	0.0693***
			(0.00407)	(0.00426)	(0.00426)	(0.00741)	(0.00441)
Politician-Month Fixed Effects		Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects				Week	Week	Week	Week
Baseline controls x on Twitter					Yes	Yes	Yes
Observations	565,968	565,968	565,764	565,764	565,764	236,700	329,064
R-squared	0.019	0.820	0.821	0.825	0.825	0.886	0.791

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1) - (5) include all politicians while column (6) includes only new and column (7) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A13: Joining Twitter and the Probability and Number of Donations with Week Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
on Twitter x Twitter penetration	0.0478**	0.0506**	0.0665***	0.0256	0.0564	0.0600	0.0927**	0.0137
	(0.0194)	(0.0193)	(0.0241)	(0.0325)	(0.0413)	(0.0422)	(0.0455)	(0.0664)
on Twitter	0.0178	0.0430	-0.452	0.846	0.0509*	0.250	-0.823	1.978*
	(0.0140)	(0.345)	(0.447)	(0.597)	(0.0279)	(0.546)	(0.719)	(1.141)
Log (campaign expenditure)	0.0103***	0.0103***	0.0144***	0.00836***	0.0245***	0.0245***	0.0339***	0.0196***
	(0.000618)	(0.000617)	(0.00103)	(0.000632)	(0.00106)	(0.00106)	(0.00199)	(0.00103)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Week	Week	Week	Week	Week	Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	565,764	565,764	236,700	329,064	565,764	565,764	236,700	329,064
R-squared	0.788	0.788	0.847	0.752	0.840	0.840	0.902	0.802

Note: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the probability of receiving at least one donation in a week in columns (1) - (4) while it is the total number of donations in a week in columns (5)-(8). Columns (1) - (2) and (5)-(6) include all politicians while columns (3) and (6) includes only new and columns (4) and (8) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A14: Politician Tweets and Re-Tweet: Full Table

	Log (aggregate donations)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
onTwitter x Twitter penetration x log(tweets)	0.374	0.474	0.285	0.884				
	(0.652)	(0.663)	(0.739)	(1.111)				
onTwitter x Twitter penetration x log(retweets)					-1.108	-1.509	-2.332	-2.227
					(1.717)	(1.714)	(3.611)	(2.412)
onTwitter x log(tweets)	-0.637	-0.459	-1.902***	-0.294				
	(0.628)	(0.671)	(0.358)	(0.757)				
onTwitter x log(retweets)					-0.0651	0.0728	-1.502***	-0.0338
					(0.454)	(0.414)	(0.373)	(0.473)
Twitter penetration x log(tweets)	-0.808	-0.671	-4.433**	-0.368				
	(0.553)	(0.573)	(1.680)	(0.620)				
Twitter penetration x log(retweets)					-0.327	-0.170	-3.308***	-0.0977
					(0.340)	(0.313)	(0.237)	(0.362)
log(tweets)	0.946	0.755	2.229***	0.435				
	(0.637)	(0.680)	(0.366)	(0.753)				
log(retweets)					0.398	0.222	1.894***	0.145
					(0.409)	(0.373)	(0.141)	(0.429)
onTwitter	0.0138	1.155	-3.777	9.125*	0.424***	0.555	-3.309	7.308
	(0.467)	(2.724)	(3.618)	(4.586)	(0.104)	(2.431)	(3.296)	(4.511)
onTwitter x Twitter penetration	0.707	0.838	0.936	0.757	0.363**	0.380**	0.691***	-0.209
	(0.638)	(0.653)	(0.745)	(1.053)	(0.150)	(0.145)	(0.171)	(0.261)
Log(campaign expenditure)		0.106***	0.121***	0.0922***		0.106***	0.119***	0.0917***
		(0.00749)	(0.0131)	(0.00975)		(0.00718)	(0.0124)	(0.00954)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend		Week	Week	Week		Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	70,954	70,932	42,405	28,527	78,107	78,082	46,964	31,118
R-squared	0.789	0.794	0.844	0.720	0.792	0.797	0.843	0.727

Note: Robust standard errors clustered at the level of the state in parenthesis. {*}{*}{*} p$<$0.01, {*}{*} p$<$0.05, {*} p$<$0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) includes only new ones and columns (4) and (8) has the experienced politicians. Baseline controls, at the level of the state, are interacted with the politician being on Twitter, include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over \$250,000 a year, the median household income as well as the share of the population with a college degree.

Table A15: Tweet Content: Number of URLs and Number of Use of Word "We": Full

	Log (aggregate donations)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	All	All	New	Experienced	All	All	New	Experienced
on Twitter x Twitter penetration x log(links)	-0.544	-1.207	13.45***	-1.198				
	(1.065)	(1.063)	(1.545)	(1.201)				
on Twitter x Twitter penetration x log(#we)					1.778	1.986	73.56***	-9.190
					(11.30)	(11.33)	(2.590)	(6.633)
onTwitter x log(links)	0.998	1.521	-4.329***	1.480				
	(1.276)	(1.376)	(0.550)	(1.588)				
onTwitter x log(#we)					-0.352	-0.334	-45.61***	10.14
					(11.05)	(11.08)	(0.923)	(7.069)
Twitter penetration x log(links)	-0.119	0.478	-14.33***	0.623				
	(1.029)	(1.031)	(0.811)	(1.067)				
Twitter penetration x log(#we)					-2.190	-2.370	-71.65***	9.093
					(11.39)	(11.43)	(0.222)	(6.628)
log(links)	-0.0547	-0.599	5.391***	-0.768				
	(1.220)	(1.331)	(0.318)	(1.371)				
log(#we)					0.993	0.950	45.71***	-10.16
					(11.12)	(11.16)	(0.0485)	(7.058)
on Twitter x Twitter penetration	0.382**	0.399***	0.719***	-0.196	0.360**	0.377**	0.693***	-0.212
	(0.150)	(0.145)	(0.171)	(0.258)	(0.150)	(0.145)	(0.171)	(0.258)
on Twitter	0.398***	0.502	-3.395	7.299	0.432***	0.590	-3.296	7.322
	(0.104)	(2.422)	(3.314)	(4.478)	(0.105)	(2.432)	(3.301)	(4.494)
log(campaign expenditure)		0.105***	0.119***	0.0917***		0.106***	0.119***	0.0915***
		(0.00721)	(0.0124)	(0.00956)		(0.00718)	(0.0124)	(0.00952)
Politician-Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend		Week	Week	Week		Week	Week	Week
Baseline controls x on Twitter		Yes	Yes	Yes		Yes	Yes	Yes
Observations	78,107	78,082	46,964	31,118	78,107	78,082	46,964	31,118
R-squared	0.792	0.797	0.843	0.727	0.792	0.797	0.843	0.727

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1)-(2) and (5)-(6) include all politicians while columns (3) and (7) include only new and columns (4) and (8) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Table A16: Joining Twitter and Aggregate Donations: Politician and Week Fixed Effects

	Log (Aggregate donations)					
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	All Politicians				New	Experienced
on Twitter x Twitter penetration	0.0861	0.303**	0.218**	0.389***	0.542***	-0.0551
	(0.133)	(0.144)	(0.107)	(0.121)	(0.164)	(0.141)
on Twitter	0.976***	0.147	0.133	-2.303	-5.373**	2.858
	(0.103)	(0.113)	(0.0897)	(1.406)	(2.065)	(2.244)
Twitter penetration	0.202	0.0106	0.0570	-0.0172	0.341	0.0402
	(0.150)	(0.522)	(0.350)	(0.354)	(0.378)	(0.399)
Log (campaign expenditure)			0.307***	0.306***	0.378***	0.169***
			(0.0166)	(0.0168)	(0.0187)	(0.0167)
Politician Fixed Effects		Yes	Yes	Yes	Yes	Yes
Week Fixed Effects		Yes	Yes	Yes	Yes	Yes
Baseline controls x on Twitter				Yes	Yes	Yes
Observations	30,354	30,354	30,341	30,341	17,907	12,434
R-squared	0.022	0.572	0.606	0.606	0.656	0.565

Notes: Robust standard errors clustered at the level of the state in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logarithm of aggregate donations in a week. Columns (1) - (4) include all politicians while column (5) includes only new and column (6) only the experienced politicians. State level baseline controls interacted with the politician being on Twitter include the share of the population who voted for Bush in 2004, the percentage of African-American population, the share of population which earns over $250,000 a year, the median household income as well as the share of the population with a college degree.

Data Appendix

Notes on Data Collection from Twitter

We provide guidelines for Twitter data collection here. Twitter allows researchers and developers to pull data from API in two different forms.

1. **REST API.** The API allows researchers to look up any user or tweet from the past conditional on a unique identifier (i.e. a user's Twitter handle, a tweet's ID, etc). However, Twitter places pretty tight constraints on the amount of data one can get in a given window of time. Due to the limitations in data gathering, we use the REST API to collect information about the politicians and their tweets.

2. **Streaming API.** This API is the most commonly used tool for gathering Twitter data in academic research. The Streaming API allows researchers to tap into 1% of all incoming tweets in a random fashion and without the data extraction limits of the REST API. Via the Streaming API, we are unable to obtain every tweet posted on Twitter, but we obtain a consistent random sample of them. We use this API when we need massive amounts of data: the followers' profile information and their tweeting activity data.

Verification of Politician Twitter Accounts. After data collection, a research assistant who is blind to the research question manually verified the politician accounts. The verification of the politician accounts could also partially be handled via the Twitter API field verified which shows whether or not an account is verified. However, some congressman hold unverified accounts, although from the posted information on the profiles, it is plausible to assume the accounts are authentic.

Searching for a Candidate's Account. The search for a candidate account on Twitter is initiated by searching for each candidate's name via the Twitter API, and deduced which handle was his or hers algorithmically and subsequently checked manually by an RA.

Chapter II: Clicks and Editorial Decisions: How does Popularity Shape Online News Coverage?

with Pinar Yildirim*

"*The New York Times (NYT) editors that spend their entire day deciding what goes on NYT.com don't make any of their decisions based on web stats*" - Jim Roberts[1]

1 Introduction

There is overwhelming evidence that news affects a wide variety of outcomes ranging from voting (Gentzkow et al. 2011), judicial outcomes (Lim et al. (2015)), policy (Eisensee and Stromberg (2007)) to financial decisions (Fang and Peress (2009)). Despite the large body of research, how editors decide on the coverage for stories remains essentially a 'black box'. In this paper, we attempt to open up this 'black box' with a focus on the role of popularity of online news stories. We ask whether news stories which get a larger number of clicks initially, due to reasons independent of story 'quality' or 'importance', get covered more by follow-up articles. Moreover, we ask whether the clicks received by a 'hard' political or business news story are treated symmetrically to those received by 'soft' entertainment stories. In particular, is there crowding out of 'hard' news due to the clicks received by 'soft' news?

Estimating the causal impact of popularity on editorial coverage decisions is difficult for a number of reasons. First, some stories are simply seen as more interesting than others. These stories will attract a larger number of clicks, and will likely be followed by additional articles covering the

*Assistant Professor of Marketing, Marketing Department, The Wharton School, University of Pennsylvania, e-mail: pyild@wharton.upenn.edu.

[1]The Assistant Managing Editor of the NYT until 2014. See more at http://www.businessinsider.com/nytcom-front-page-editors-dont-know-what-stories-readers-are-clicking-on-2009-7?IR=T

same interesting event. This makes it difficult to disentangle the cause and effect: are editors choosing articles they know readers will like, or are both readers and editors responding to the intrinsic 'quality' of the story? We address this issue by instrumenting for clicks using shocks to page views which are unrelated to the characteristics of stories published on a given day. These shocks come from two different sources: days with rain and ordinary power outages. On rainy days, readers are constrained in their outdoor activities and may be more likely to read the news. Power shortages, on the other hand, limit individuals' ability to use electronic devices or connect to the internet. Second, testing the relationship between clicks and editorial coverage requires disaggregated data at the article level. To this end, we acquire proprietary information on clicks at the level of a URL from a leading English language Indian national daily newspaper.

Next, we examine the impact of clicks-based coverage on the type of content provided to readers. We classify the stories as 'hard' and 'soft' to analyze whether the editor responds in the same fashion to the clicks received by the two different types of news stories.[2] We focus on the breakout of the biggest corruption scandals and celebrity scandals during the year, as big hard and soft news events, respectively. We analyze if on celebrity scandal days hard news is crowded out or whether on days of the breakout of corruption scandals soft news articles are crowded out as a result of the clicks.

Whether or not news editors systematically follow up on clicks and how they allocate resources across different types of stories has wide ranging implications. There is a growing concern about the 'dumbing down' of news content online due to editors focusing on clicks rather than story 'quality' or 'importance'.[3] Best (2009) shows theoretically, in the context of 'hard' news, that selective coverage of stories due to their 'sensational appeal' can distort the beliefs of readers and make them demand a sub-optimal policy. [4] Moreover, Prat and Stromberg (2013) highlight, in a comprehensive survey of the media economics literature, that identifying whether news stories are chosen because of demand side (popularity) reasons is potentially important for the design

[2]We group stories in the National, International, Business and Opinion sections as 'hard' news while stories in the Entertainment, Technology, Lifestyle and Sports sections are grouped as 'soft' news as in Cage (2015). Hard news is considered important for providing information related to public decision making while soft news is said to be for pure consumption or entertainment value (Boczkowski (2010)).

[3]See for example `http://www.bbc.com/news/uk-wales-34209894` and `http://www.usnews.com/news/national/articles/2008/08/28/a-digital-dumbing-down`.

[4]This concern is echoed by a recent survey by the Royal Statistical Society which shows that the British public holds incorrect perceptions about statistics related to issues including crime, immigration and government spending. One reason stated for this is the way media sensationalizes coverage of different topics. See http://www.independent.co.uk/news/uk/home-news/british-public-wrong-about-nearly-everything-survey-shows-8697821.html for more.

of media regulation.[5]

These implications assume greater importance given the public good aspect of news which differentiates it from a standard product. We capture this public good dimension of news in a simple model of editorial decisions. We show that a clicks based editorial strategy does not maximize intrinsic story 'quality' which in turn feeds into reader decisions and leads them to make sub-optimal decisions.[6] It is perhaps for this reason that various news outlets have distanced themselves from a clicks based strategy determining editorial decisions. The following are some examples:

"Vox.com don't give their journalists access to analytics- it is more important to pay attention to content right now".[7]

"The whole mission of BuzzFeed is to get people to share. That is not the mission of The New York Times. The mission of The New York Times is about the best journalism in the world and giving people accurate, timely information."[8]

"Even now, when cutbacks are affecting USA TODAY- as they are all newspapers- I've seen no evidence that the paper or its Web site have succumbed to the temptation to snare readers with scare headlines".[9]

The journalism and communications literature validate this general claim made by newspaper editors. Singer et al. (2011) conducted a survey of the local press in Britain and concluded that "despite excruciatingly detailed 'hit log' data, online audience preferences do not seem to be having a notable agenda-setting impact on local editors."[10] According to Singer et al. (2011) "...[there is] the general finding in journalism studies that, amid opportunities for audience

[5]The Federal Communications Commission (FCC), for example, have ignored demand side incentives while proposing regulations for media ownership assuming a positive correlation between ownership of the media outlet and the viewpoint expressed by it. This idea of content restrictions was alluded to in an article related to the coverage of terrorism in The Guardian. See http://www.theguardian.com/media/2015/aug/01/media-coverage-terrorism-further-violence for more. Of course, any media regulation or policy comes with a big caveat of not infringing on the freedom of the press.

[6]Moreover, we show that if editorial strategy is based on specific political preferences then it leads to identical inefficient outcomes as in a situation where editors have a clicks based strategy.

[7]This quote was by Melissa Bell who was one of the co-founders of Vox.com. See more at http://ijnet.org/blog/voxs-melissa-bell-journalists-should-step-away-click-bait-mentality. This policy is followed by The Verge as well.

[8]This is attributed to Alex MacCallum who is currently the assistant managing editor for audience outreach. See https://gigaom.com/2015/01/14/news-flash-for-the-nyt-you-and-buzzfeed-arent-that-different/ for more.

[9]Read more on http://www.washingtonpost.com/blogs/capital- weather-gang/wp/2014/12/18/perspective-how-the-internet-changed-the-weather- journalism-industry/

[10]Boczkowski (2010) comes to similar conclusions after surveying various newsrooms in South America finding that "journalists generally held their ground against the encroachment of audience desires."

participation and inclusion in the news process, journalists keep at the margins those outside influences that might reshape their values or practices."

We provide three key insights on the impact of digital data on editorial policy. First, we show that an increase in the views of the first article of a story, independent of its intrinsic quality, significantly increases the amount of follow-up coverage provided by the newspaper on the same story. Specifically, a one standard deviation increase in clicks extends the duration of a story by 3 days and leads to 3 additional articles being written as part of the story. We relate the magnitude of these estimates to the literature looking at the impact of news coverage on various outcomes as well as the one which quantifies the impact of political preferences on editorial coverage decisions.

Second, estimating the model by splitting the sample into hard and soft news stories, we show that clicks have a positive and significant impact on coverage only for hard news. Contrary to conventional wisdom, the newspaper gives additional coverage based only on the clicks received by hard news and not to those received by soft news. Complementing this result, we find that the proportion of clicks to hard news on days when there is a celebrity scandal is lower by about 7.5%. This reduction in clicks for hard news, however, does not result in a decline in the proportion of hard news articles published on those particular days. On the other hand, days on which news breaks out about a corruption scandal, clicks on hard news articles increase by about 4.5% which also leads to an increase in the proportion of hard news articles by more than 2% on those days.

Third, we quantify the extent of editorial misinterpretation due to the additional clicks on rainy days by comparing coverage decisions given to stories on rainy days relative to those published on national holidays and weekends.[11] We find that national holidays and weekends lead to a statistically significant increase in clicks of a magnitude similar to rainy days but do not lead to an increase in coverage, suggesting that editors are aware of these exogenous increases in reader attention. As an additional check, we find that the clicks received by the second article of a story whose first article was published on a rainy day are lower by 4-6% relative to the views received by the second article of a story whose first article was not published on a rainy day.[12]

To our knowledge, this is the first study to quantify the impact of popularity on online editorial coverage decisions, and furthermore the first to conduct this analysis at story level. Our paper is

[11] We do this, additionally, to rule out a channel such that it could be rational for the editor to provide more coverage to stories published on rainy days.

[12] This relationship, though, is not robust possibly due to the reduction in the sample size.

related to the literature on media bias (Mullainathan and Shleifer (2005), Gentzkow and Shapiro (2010)) and agenda setting (Larcinese et al. (2011), Katona et al. (2013)). Focusing on economic issues, Larcinese et al. (2011) find that newspapers with pro-Democratic endorsement systematically publish fewer articles on negative economic reports when the president is Democratic. Gentzkow and Shapiro (2010) provide evidence for slanting news content towards the political ideology of readers of a newspaper in order to maximize profits, but do so at the level of the newspaper aggregated annually while we explicitly take into account the feedback effect of popularity. Differently from these papers, we focus on the decisions of an editor at an individual story level without making assumptions about the story's political bias or the newspaper's ideology.

Our paper also contributes to the debate on whether basing editorial decisions on online real time metrics is detrimental to the quality of published content.[13] We provide evidence that the newspaper does not respond to the clicks received by soft news stories, allaying concerns about excessive editorial focus on pageviews. This is in line with the evidence provided by Salami and Seamans (2014) who find that the quality of news content has improved with the introduction of the internet.

Finally, this study provides suggestive evidence on how well newspapers are able to handle and interpret 'big data'.[14] We highlight the possibility of crowding out of newsworthy stories because of the newspaper not recognizing the different sources of variation in readers such as those caused by rainfall and outages. This has negative consequences for information provision to readers, as well as for newspaper's profits. The newspaper misinterpreting the source of clicks is more generally related to the literature on firms (mistakenly) making suboptimal decisions, documented in the cases of American football (Romer (2006)) and Cricket (Bhaskar (2009)).

The rest of the paper proceeds as follows. Section 2 presents a simple model of media coverage decisions, Section 3 describes the data and methodology, while in Section 4 we discuss identification and the empirical strategy. Section 5 reports our baseline estimation results. In Sections 6 and 7, we investigate what types of news stories are crowded out. In Section 8 we carry out a series of checks to analyze the robustness of our baseline results and Section 9 concludes.

[13] For more see the four part piece on this topic in the American Journalism Review http://ajr.org/2014/03/27/pay-per-visit-debate-chasing-viral-traffic-hurting-journalism/

[14] This was manifested in the Newspaper Association of America conference in February 2014 on "The Exploding World of Data: How can newspaper media – from journalists to advertisers to publishers – use big data?"

2 The News Industry in India: A Brief Overview

In this section, we briefly provide some details about the newspaper whose data we analyze and the news industry in India overall.

We analyze the data from a leading Indian national daily which is not only amongst the top english language dailies in the country in terms of daily circulation (or an online audience) but also amongst the most respected outlets when it comes to content. The news company does not only publish in English but also in other regional Indian languages. It has a long and credible history as it was established before India got independence from British colonial rule and it played an important role in the country's struggle for independence. Post independence, the newspaper has developed a reputation for high quality content often highlighted for upholding traditional journalistic values and integrity. It is for these reason that comparision with the New York Times in the US context is an apt one. Quotes similar to those mentioned in the Introduction related to the New York Times are actually publicly available for the newspaper we analyse but we cannot point towards them since it would compromise the identity of the newspaper. Moreover, it has also been amongst the first mainstream newspapers to embrace different strategies to capture and grow a sizeable online audience.

More generally, The news industry in India was valued at $4.37 billion with 107 million copies of newspapers being sold daily in the year 2009-2010.[15] The online readership has seen sharp annual increases with 9.4 million daily unique viewers to the category in 2013 up by 34% compared to the previous year.[16] The Indian newspaper industry is a composed of a large number of newspapers in regional languages along with a set of national English language dailies. Even though the print circulation numbers of the English language dailies are not as high as those published in regional languages, they fare much better when it comes to online readership since internet penetration is limited to metropolitan cities who are also more likely to consume their news in English. The leading English language dailies in the country are The Times of India, Hindustan Times, The Indian Express and The Hindu amongst others. Even though they might not have the lion's share of the circulation numbers, since they cater to the educated and high income tiers of society, they definitely do set the agenda in terms of what issues are to be considered of national importance.

[15] See http://www.bbc.com/news/business-14362723 for more.

[16] See a report based on comScore data http://www.comscore.com/Insights/Press-Releases/2013/10/Indias-Daily-Readership-of-Online-News-and-Information-Jumps-34-Percent-in-the-Past-Year.

3 A Simple Model of Editorial Decisions

3.1 The Newspaper

Our model builds on the setup used by Gentzkow and Shapiro (2010) and Latham (2014). Consider a single newspaper which needs to decide how much coverage c_i to give to a particular story i [17]. The newspaper cares about the revenue generated by covering a story, which we assume is proportional to the number of readers with the total mass of readers normalized to 1[18]. There is a (marginal) disutility $\lambda_i \in R_+$ associated with covering any story i. Thus, $\lambda_i c_i$ captures the costs from covering a story. Formally, we express the payoff to the newspaper from allocating coverage c_i to story i as

$$\pi(i) = R(.) - \lambda_i c_i,$$

where $R(.)$ is the readership of the story i on which we elaborate below.

3.2 Readers

We consider a market with a unit mass of readers. The appeal of a story i to the readers is given by a parameter α_i, which is assumed to be common to all readers. The readers' enjoyment of reading a story depends on this preference parameter α_i, and the amount of coverage allocated to a story, c_i. We represent this utility by a function $f(.)$ and assume that it is monotonically increasing and strictly concave in c_i. This implies that individuals' utility increases in coverage but additional coverage has diminishing marginal utility. Further, we assume that $f(c_i, \alpha_i)$ is increasing in α_i and $\frac{\partial^2 f(c_i,\alpha_i)}{\partial c_i \partial \alpha_i} > 0$. Put simply, the marginal utility of reading additional material on story i is greater for preferred stories, i.e., stories with higher α_i. An individual reader q has the following utility from reading story i:

$$U_q(i) = f(c_i, \alpha_i) - \delta_{iq}.$$

[17] As will be detailed in Section 4, coverage can be thought of as the number of articles published as part of a story or the length of time a story is covered for. Moreover, as in Gentzkow and Shapiro (2010) and Latham (2014), we abstract away from competition between newspapers and focus on the independent decisions made by a single newspaper.

[18] This implies that the newspaper values all audiences the same.

The term δ_{iq} captures story characteristics which affect each reader idiosyncratically. This could be thought of as an idiosyncratic taste shock or a composite term which includes idiosyncratic reading costs. For tractability, we assume that δ_{iq} is uniformly distributed on $[0, 1]$.

The timing of the game in this simple model are as follows. First the news editor or the newspaper decides how much coverage to give to each individual story i. Next, the readers observe the coverage and characteristics of a story and decide whether to read story or not with the reservation utility being normalized to 0. Consumers read the stories and receive utility and the newspaper gains readership and hence revenue.

3.3 Analysis

Using backward induction, we first analyze the decision of the readers. Readers observe the coverage of story i by the newspaper, and their preference for that story α_i. Individual q will read/click on the story if:

$$f(c_i, \alpha_i) \geq \delta_{iq}. \tag{1}$$

This implies that the aggregate readership for story i equals:

$$R(c_i, \alpha_i) = \begin{cases} f(c_i, \alpha_i) & \text{if} f(c_i, \alpha_i) \in (0, 1), \\ 1 & \text{if} f(c_i, \alpha_i) \geq 1, \\ 0 & \text{if} f(c_i, \alpha_i) \leq 0. \end{cases} \tag{2}$$

We focus on the case where $f(c_i, \alpha_i) \in (0, 1)$. Next, taking the readership of the story into account, the newspaper chooses c_i to maximize its utility:

$$f(c_i, \alpha_i) - \lambda_i c_i. \tag{3}$$

The first order condition is given by

$$f'(c_i, \alpha_i) - \lambda_i = 0, \tag{4}$$

and the second order condition for a maximum is satisfied since $f(.)$ is strictly concave in c_i implying $f''(c_i, \alpha_i) < 0$.

In order to estimate the first order condition directly, we impose some structure on the model to find closed form solutions. We assume a functional form for $f(c_i, \alpha_i)$ given by $\sigma(\alpha_i c_i)^{\frac{1}{\sigma}}$ with $\sigma > 1$. This gives us the first order condition for the utility maximizing coverage for the newspaper as:

$$c_i = \alpha_i^{\frac{1}{\sigma-1}} \lambda_i^{\frac{\sigma}{1-\sigma}}. \tag{5}$$

From Equation (5) one can see that $\frac{\partial c_i}{\partial \alpha_i} = \frac{1}{\sigma-1} \alpha_i^{\frac{2-\sigma}{\sigma-1}} \lambda_i^{\frac{\sigma}{1-\sigma}} > 0$. The higher the appeal of a story, the greater the coverage it receives from the newspaper.

3.4 A Structural Model for Estimation

The predictions derived from the model form the basis of our empirical analysis. By imposing additional structure on the theoretical model, we can next estimate the equation $c_i = \alpha_i^{\frac{1}{\sigma-1}} \lambda_i^{\frac{\sigma}{1-\sigma}}$ directly. Taking the natural logarithm of c_i, yields a log-log specification:

$$log(c_i) = \frac{1}{\sigma - 1} log(\alpha_i) + \frac{\sigma}{1-\sigma} log(\lambda_i). \tag{6}$$

Assume

$$log(views_i) = \beta_1^{\alpha} + \beta_2^{\alpha} log(\alpha_i) + \psi_i^{\alpha}, \tag{7}$$

which implies that the logarithm of the views or clicks of story i is a function of the appeal of the story and an error term. As the theoretical model demonstrates, we expect the coverage of a story to increase in the views received.

Similarly, let

$$log(\lambda_i) = \beta_1^{\lambda} + \mathbf{X}_i' \beta_2^{\lambda} + \zeta_i^{\lambda}, \tag{8}$$

implying that the preference of the newspaper for any story (captured by λ_i) be a linear function of a set of covariates and an error term.

Substituting for the expressions from (7) and (8) into Equation (6) yields

$$log(c_i) = \frac{\sigma}{1-\sigma}\beta_1^{\lambda} - \frac{\beta_1^{\alpha}}{(\sigma-1)\beta_1^{\alpha}} + \frac{1}{(\sigma-1)\beta_2^{\alpha}} log(views_i) + \frac{\sigma}{1-\sigma}\mathbf{X}_i'\beta_2^{\lambda} - \frac{1}{(\sigma-1)\beta_2^{\alpha}}\psi_i^{\alpha} + \frac{\sigma}{1-\sigma}\zeta_i^{\lambda}, \tag{9}$$

which can be more succinctly written as

$$log(c_i) = \gamma_0 + \gamma_1 log(views_i) + \mathbf{X}_i'\gamma_2 + \epsilon_i,$$

where $\gamma_0 = \frac{\sigma}{1-\sigma}\beta_1^{\lambda} - \frac{\beta_1^{\alpha}}{(\sigma-1)\beta_1^{\alpha}}$, $\gamma_1 = \frac{1}{(\sigma-1)\beta_2^{\alpha}}$, $\gamma_2 = \frac{\sigma}{1-\sigma}\beta_2^{\lambda}$ and $\epsilon_i = -\frac{1}{(\sigma-1)\beta_2^{\alpha}}\psi_i^{\alpha} + \frac{\sigma}{1-\sigma}\zeta_i^{\lambda}$. We are interested in the coefficient $\gamma_1 = \frac{1}{(\sigma-1)\beta_2^{\alpha}}$ to evaluate the impact of views on coverage. We expect γ_1 to be positive since $\sigma - 1 > 0$.

We now move to describing the data in greater detail before highlighting our empirical strategy to identify our parameters of interest.

4 Data and Methodology

4.1 The Data: An Overview

The empirical analysis is performed by creating a unique dataset based on the articles published by a leading Indian national English daily on its online website.[19] We have access to data on all articles that were read online for a period of one year, from the 1^{st} of January 2012 to the 31^{st} of December 2012. For each article (or, equivalently URL) we observe the number of page views as well as the unique page views. We provide further details on these measures below.

To account for article level heterogeneity which may contribute to a story's popularity, we combined the data obtained from the newspaper with data collected by crawling the news website. We use an automated script to scrape the news site to collect data on the text of each article along with its headline, whether it had a picture on the page, the time and date the article was published, as well as the source of the article (whether the story was written by a journalist of the newspaper or was appropriated from a news wire). We also scrape the tagging information

[19] Under the Non-Disclosure Agreement signed with the newspaper, it will never be referred to by its name but as a "leading Indian national daily". The national daily is amongst the largest English dailies in terms of circulation rates. Moreover, it is a highly respected outlet in terms of the quality of content.

at the bottom of the article page (when available) to identify the section the article belonged to. We matched proprietary data from the newspaper with data scraped on each article to obtain the dataset which forms the basis of our analysis.

4.2 Identifying News Stories

Our unit of observation is a 'news story' which we define as a cluster of successive news articles published on a common issue or topic. To define a cluster we first mine the text of each article using natural language processing techniques, and identify the proper nouns appearing in each article. The proper nouns, which refer to entities such as people, places, institutions and events, provide us with a list of keywords which summarize the topic of the article. We sort the articles in the order of their publication date, starting from the oldest to the most recent. We then implement a clustering method which employs a word frequency algorithm borrowed from the computer science literature (see Radev et al. (1999)) to identify similarities between articles, also used by Franceschelli (2011). The algorithm compares an article to another article based on a proper noun list and frequency of these proper nouns. If proper nouns of the second article show sufficient similarity (i.e., exceeds a set threshold) to that of the first article, the algorithm then combines the two articles into a single cluster. We update the list of proper nouns for the cluster by including the proper nouns from both articles. We continue the search process by looking for another article published within a 24 hour window of the latest article of the cluster, and if this article contains a similarity score which is higher than the specified threshold, we add this article into the cluster as well. We continue this process until no other article can be assigned to this cluster within the designated 24 hour window.

As in Franceschelli (2011), the similarity coefficient between article A and a cluster of articles S is measured with

$$similarity(S, A) = \frac{\sum_{k=1}^{p} w_{Sk} \times w_{Ak}}{\{\sum_{k=1}^{p}(w_{Sk})^2 + \sum_{k=1}^{p}(w_{Ak})^2\}^{\frac{1}{2}}},$$

and w_{Sk} is defined as:

$$w_{Ak} = \frac{t_{fk} \times log(\frac{N}{n_k})}{\left\{\sum_{k=1}^{p}(t_{fk} \times log(\frac{N}{n_k}))^2\right\}^{\frac{1}{2}}}$$

where article A is represented as a vector $A = (tf_1, tf_2, ...tf_p)$ of proper noun frequencies. The term tf_j denotes the frequency of proper noun j in article A, and is recorded for all proper nouns that appear in any article in the news cycle. N is the total number of articles in the database

and n_j is the total number of articles in the database in which proper noun j appears. The factor $log(\frac{N}{n_j})$ weighs the term tf_j in a way to allocate greater weight to the proper nouns which appear relatively less frequently.

The similarity coefficient between article A and cluster S is normalized to take a value in the $[0, 1]$ interval.

An article could be placed in a cluster to which it actually does not belong due a threshold that is too low or it could be excluded from a cluster that it should belong to due to a threshold too high. Hence, the threshold value for similarity (S, A) has to find a balance between these two opposite effects.[20]

Notice that the clustering algorithm allows a story to continue as long as there is an article published on the same topic every 24 hours. If an article is not sufficiently similar to any other article within the 24 hour window, then the article itself becomes a story. Due to the large number of articles, we follow Franceschelli (2011) in dividing the year days into 24-hour news cycles and assign each article to exactly one story to limit the number of comparisons that need to be made between articles so that it is computationally manageable. After the process, we end up with 128,699 articles clustered into a total of 60,689 stories.

An example of a cluster is the articles on the Fukushima disaster in Japan, as listed in Table 1. The cluster began on the 5^{th} of July 2012 at 1:08 pm and went on till 7^{th} of July with the last article published at 12:55 am. The cluster ends with five articles, with at least one article written about the incident every 24 hours, and no articles published on the 8^{th} of July. Any articles published relevant to the Fukushima disaster on or after the 9^{th} of July are grouped into another cluster.[21]

4.3 The Dependent and Explanatory Variables

We now describe our main dependent and independent variables of interest with the descriptive statistics provided in Table 2.

[20] Our results are based on threshold values of 0.3 as in Franceschelli (2011) which seem to provide the best clustering results. To ensure that our results are not driven by a specific threshold value, we also re-run our results with threshold values of 0.4 and 0.5. Regression results with a threshold value of 0.4 are provided in the Data Appendix. Further results based on a threshold value of 0.5 is available on request.

[21] More examples for stories are in Table 21 in the Data Appendix.

4.3.1 Dependent Variables of Interest

We focus on two distinct yet related measures of editorial response to reader demand: the number of articles within a story and the duration of the story.

The Number of Articles in a Story: It is possible that a popular story may motivate the newspaper to publish more articles on the same topic in response to the demand from the consumers. This way, the editor may choose to give the readers updates on the issue through additional articles. The mean number of articles written under an issue is 1.70, and the median issue has only one article. The distribution is skewed, with a standard deviation of 7.34 articles.

The Length of a Story: In addition, the editor can choose to publish articles about an issue for a longer period of time. The length of time a story is covered is measured by the time elapsed between the first and the last article of story i. If a story involves a single article, then by definition, the duration is zero. The mean length of a story is 4.79 hours and our median story is part of a cluster which contains only one article and hence is of length zero.[22] The longest story in our cluster lasts for 85 days.

Since our dependent variables have a skewed distribution, we use the inverse hyperbolic sine transformation commonly utilized in studies of income and wealth.[23] Moreover, since the median story has one article or is of length zero, we use Tobit models (besides a two stage least squares approach) which explicitly takes this mass point in the data into account.

4.3.2 Explanatory Variables of Interest

Measures of Popularity: For each article within a story, we have the number of page views that it accumulated over a thirty day period. The mean number of views received by an article is just over 217.9 with a standard deviation of 695.38. To account for the skewed distribution we take the logarithmic transformation of the total number of views received by the article.

One must keep in mind that clicks of an online news article, in general, peak quickly after its publication. As an example, consider the number of views received by the most popular article of

[22] One can think of zero length to be measured with some error since we do not observe for how long the article stayed on the online page. This interpretation will also alleviate concerns about a 2sls model not being appropriate since there is a mass point at zero. Foster and Kalenkoski (2013) show that OLS performs better than Tobit when the mass point is due to measurement error. To account for the sizable number of zero durations and the resulting skewed distribution, we use the inverse hyperbolic sine transformation to the story length.

[23] See Pence (2006) for more on this. We tested a number of robustness checks in Section 9 to ensure that our results are not driven by outliers.

August 2012 about Indian wrestler Sushil Kumar making it to the wrestling final of the London Olympics first published on the 12th of August at 8:26 pm. It received 16,757 views on the same day out of a total of 16,953 received over a thirty day period which amounts to 99% of the total views.[24] To reduce the burden of data extraction for the newspaper, we acquire and use data on the thirty-day page views as an approximation to the first day page views received by an article. [25] [26]

Other Control Variables:

We also use article level characteristics as controls since these could influence the clicking decision. To control for the salience of an article's link, first we control for the length (i.e., the number of characters) of its headline. An article headline, on average, has 44.1 characters. Second, we control for whether an article has an image or not, which can help draw attention to the article. Third, we create a variable which aims to control for the section of the newspaper in which the article appeared. We use the story 'tags' listed at the bottom of each article to determine the section to which an article belongs to. For example, a national news article is generally tagged with 'nation-news', and articles with such tags are classified as being part of the national news section. In a similar fashion, we create section classifications on world-news, business, sports and entertainment news where we have information on tags. Unfortunately section tags exist only for a subset of articles in the dataset. To determine the section the unlabeled articles belong to, we use another machine learning method called the Naive Bayes Classifier. The method uses the tags and the text of each article for which we have the tagging information and gives the most likely section classification for each article that does not have the tags information in the dataset.[27]

Fourth, the word count of an article is taken into account as it may indirectly capture the resources put into writing it by the newspaper. Finally, we use an indicator for whether an article was sourced from news agencies such as the Press Trust of India (PTI), which do not

[24] The headline of the article read "Sushil Kumar scripts history at Olympics for India, wins Silver."

[25] In a different example, an article on the Presidential elections in India was first published on June 18^{th} at 7:14 pm. It received 10,654 views on the first day. Compared to the 10,829 total clicks received over a thirty day period, this amounts to about 99% of the total views. The headline of the article was "Presidential poll: Kalam rebuffs Mamata, Advani says will not contest against Mukherjee". Moreover, we randomly sampled a subset of the articles published and find that in every month, on average about 95% of clicks come within the first few hours of being published.

[26] As part of the data collection process, a 6-8 week time period was spent in the news organization. During this period, it was observed that the online team focused solely on the articles published on the same day and the number of views that they receive.

[27] For more on the methodology behind the Naive Bayes Classifier see Mitchell (2015).

have profit concerns in their reporting decisions. The source of a story can influence its cost and relevance to the newspaper.

5 Identification and Estimation

It is hard to establish the causal impact of reader demand on editorial behavior due to the concerns of unobserved heterogeneity, reverse causality, and measurement error. We are concerned about reverse causality since extended coverage of a story might itself generate greater interest and hence page views, which could bias our estimates. To deal with this issue, we use the number of clicks associated with the first article of a story as the key predictor of editorial coverage. Unobserved heterogeneity is also a concern because some characteristics of stories (such as 'newsworthiness') or preferences of editors are unobservable to the econometrician which would bias the estimates. Consider a story about the Ebola virus outbreak in 2014. It may receive greater coverage and number of clicks because it is considered 'important' by both the editor and the readers. It is exactly this story level 'importance' (or other characteristics), perceived either on the side of the readers or editors, which is unobservable to the econometrician. Moreover, since we use an aggregate measure to proxy for the real time views received by an article, there could be issues of measurement error. To avoid unobserved heterogeneity and measurement error artifacts, we instrument for the clicks received by the first article of each story. We use two instruments which provide an exogenous shock to the number of views received by the first article of each story.

Rainfall: The first instrument we use is rainfall, relying on the simple observation that when it is raining, people are less likely to spend time outdoors. We expect that spending more time indoors is associated with being more likely to go online and visit the news website.

We use daily data on rainfall obtained from the CEIC database. We use the days of the year when it rained in Delhi and Mumbai. Although we do not have a geographical breakdown of where the readers are coming from for each article, we have summary statistics on the geographical distribution of clicks for 4% of the sample. These two cities consistently rank in the top two in terms of total views on the website[28]. Our instrument takes the value 1 if a particular day saw

[28]These two cities account for up to 50-60% of the daily clicks coming from the top fifteen cities for the newspaper.

positive rainfall in either Delhi or Mumbai and zero otherwise. There are a total of 132 days for which we see either of these cities experiencing rainfall.

Power Outages: To construct our second instrument, we consider power outages during a day. Since power outages restrict access to the internet and reduce the clicks on any websites, we expect them to be negatively correlated with views that an article receives. We use publicly available data on daily power outages from the Northern Regional Load Dispatch Centre and Western Regional Load Dispatch Centre websites which are subsidiaries of the Ministry of Power of the Government of India[29]. Our power outage instrument is equal to the sum of the daily power shortage recorded in Delhi and Maharashtra (the state which Mumbai is part of)[30]. During the year 2012, on average there was a daily power shortage of 24.61 million kWh across Delhi and Maharashtra which amounts to 5% of the daily power consumption[31].

For power outage and rainfall to be valid instruments we need a sufficiently strong first stage relationship, which we estimate using OLS:

$$log(views_{it_1}) = \mu_1 + \mu_2 log(outages_{it_1}) + \mu_3 rain_{it_1} + \mathbf{X}'_{it_1}\mu_4 + \vartheta_{it_1} \tag{10}$$

where $\vartheta_{i1} \backsim N(\bar{\vartheta}, \sigma^2_{\vartheta})$. In particular, this specification looks at the impact of power outages and rainfall on the page views received by the first article (t_1) of story i on the day the article was published online. We expect that $\mu_2 < 0$ and $\mu_3 > 0$. Since rainfall is seasonal, we add month fixed effects. We use the predicted values of the views of the first article for each story i from Equation (10) and estimate its impact on the total coverage given to story i (c_i) by estimating a structural equation:

$$log(c_i) = \gamma_0 + \gamma_1 predicted(log(views_{it_1})) + \mathbf{X}'_{it_1}\gamma_2 + \epsilon_{it_1}.$$

If the newspaper does respond to the clicks by giving more coverage to a particular story, we would expect $\gamma_1 > 0$. The vector $\mathbf{X}$ consists of controls which include the section the article was

[29] See http://nrldc.in/ and http://wrldc.in/

[30] Daily power outage data at the level of the city, such as Mumbai, would have been ideal but was not available. Consulting the Ministry of Power and power generating stations made it clear that such information is not available.

[31] After combining all the data, we have 128,699 articles clustered into a total of 60,689 stories, spanning over 325 days. We lose data for some of the days because of missing power outage data for Maharashtra. To ensure the robustness of our results on the whole sample, we re-run our model using power outage data for Delhi alone, which gives us a total of 365 days to find similar results. Results are available upon request.

slotted in, the word count, the length of the headline of the article, the source of the article, whether the article had an image or not, as well as weekend and month fixed effects.

Apart from the first stage relationship strength, we need the exclusion restriction to be satisfied, i.e., $corr(outages, \epsilon_{it_1})$ and $corr(rain, \epsilon_{it_1}) = 0$. The fact that we have two instruments and one endogenous variable allows us to employ overidentification tests to test the combined exogeneity of the instruments.

Additionally, we carry out a set of placebo checks to see if editorial policy adapts to changes in readership due to rainfall or power shortages. We examine whether the number of articles published on these days is significantly different from that on any other day. We also look at the correlation between the observable characteristics of articles such as the word count, whether it was sourced from the PTI and check if they systematically differ due to rainfall or power outages.

6 Results

In this section, we present our baseline results starting with OLS estimates. We then move to the IV regressions focusing first on the performance of the instruments and then on the impact of the instrumented popularity variable on the length of the story. All specifications include the word count, whether it was sourced from the PTI, whether it had an image, the length of the headline and weekend fixed effects as controls. Huber-White standard errors robust to heteroskedasticity are used and reported in the parenthesis.

6.1 OLS Estimates

We first estimate a basic OLS specification by regressing the outcome variable of interest (the duration of the story, the number of articles written as part of the story) on the views received by the first article of the story:

$$log(c_i) = \gamma_0 + \gamma_1 log(views_{it_1}) + \mathbf{X}'_{it_1}\gamma_2 + \epsilon_i.$$

The results (reported in Table 3) show that the number of views are positively correlated with the length of a story. The relationship is statistically significant at the 1% level, with the coefficient

0.30. This implies that one standard deviation increase in the views received by the article would lead to a four hour increase in the duration of the story. The magnitude of this coefficient is quite stable over different specifications: adding section controls (given in column (2)) and month fixed effects (given in column (3)) do not impact the magnitude.

We find similar results when the outcome variable is the number of articles written by the newspaper in a story (columns (4)-(6)). An OLS regression shows a positive and statistically significant impact of views on the number of articles on a story. This relationship is stable across different specifications even with the addition of section controls (columns (5) and (6)) and month fixed effects (columns (6)). These estimates give preliminary descriptive evidence which is in line with our intuition.

6.2 Instrumental Variables Estimates

6.2.1 First Stage: How Power Outages and Rainy Days Affect Clicks

We carry out the first stage estimations to evaluate the validity of our instruments: rainfall and power outages. The coefficients on both instruments are statistically significant at 1% level with the expected signs as seen in Table 4. We can see that on a rainy day, the first article of a story receives a 5%-8% increase in the number of clicks on average. The results are in line with our intuition that on days with rainfall, people are more likely to visit the online news page and a greater number of page views would be recorded. Similarly, power outages are negatively related to the daily views implying that when people experience power shortages then they are unable to check the news as often. The instruments remain robust to addition of controls such as section fixed effects (columns (2) and (3)) and month fixed effects (column (3)). The F-statistic measuring the strength of the instruments is greater than 10 for every specification implying that they satisfy the rule of thumb for the instruments to not be weak[32].

Identification requires that rainfall and power shortages should be as if randomly assigned. This means that in the absence of these events, the stories published on these days would fare the same as stories published on any other day. This can be challenged on the grounds that even though electricity shortages and rainfall are determined orthogonally to editorial policy, they

[32] An F-statistic above 10 can be treated like a test with approximately a 5 percent significance level of the hypothesis that the maximum relative bias is at least 10 percent. For more on weak instruments Stock and Yogo (2002).

can be anticipated and hence editorial policy might adapt to it. We now check whether editorial policy responds to these outages or rainfall by carrying a set of placebo or falsification checks.

6.2.2 Placebo Checks

We regress a set of observables, which capture editorial decisions or newspaper effort, on rainy days and power outages. Table 5 shows that there is no statistically significant relationship between them and the two instruments. In columns (1) and (2) we investigate whether an article is more likely to be sourced from the PTI or whether it will be more verbose due to rainfall or power outages. We find no statistically or economically significant relationship between them. We also find no difference in the number of hard and soft news stories written due to rainfall or power outages (column (3)). Moreover, as seen in column (4), the number of articles published during a day, controlling for the number of words written, is unaffected by rain or power outages. On weekends, however, the likelihood of sourcing an article from the PTI, the number of words per article, the number of hard news stories and the number of articles published are statistically different with the relationship significant at the 1% level. The fact that we find significant differences between weekdays and weekends imply that editors anticipate and alter their editorial strategy to the (potentially positive) shock to reader attention which happen at regular intervals (weekends) as opposed to rainfall and power outages which are harder to systematically respond to since they occur irregularly. Moreover, these falsification tests provide suggestive evidence that the newspaper is unaware of these shocks to reader attention.

6.2.3 Second Stage: How Clicks Affect Length of the Story and the Number of Articles

We now turn to the impact of clicks or views on the length of a story. As can be seen in columns (1)-(3) of Table 6, views have a positive and statistically significant impact on the duration of the story. We find that a one standard deviation increase in clicks will lead to an increase in the length of the story by 1-2.5 days (given that the mean story length is about 5 hours). In other words, a story which has its first article receive clicks which is one standard deviation above the mean will have at least one additional article on the same topic within the next 24 hours. This relationship is robust to inclusion of section dummies (columns (2) and (3)), as well as month fixed effects (column (3)) though the magnitude does decrease with the addition of these controls. This provides evidence in favor of the implications coming from our model of editorial

decisions. News editors have the incentive to give greater coverage to stories which are popular to maximize revenue and these numbers point in that direction.

The overidentification test (Basmann's F-test) does not reject the null hypothesis of the collective exogeneity of the instruments with a p value which is consistently above the 10% level. There are seasonalities in both rainfall and power outages since rain is more likely to occur in the monsoon while electricity shortages are more prevalent in the summer due to high demand. High power demand, though, is only one of the reasons for power outages. In India, power outages are often due to power thefts and inefficiencies in the supply chain.[33] Hence, the processes underlying power outages and rainfall are sufficiently differentiated to take the results of the overidentification tests seriously.

While the duration of a story is an indicator of how much coverage is given to a story, there may be other editorial responses, such as the number of articles written within a story. The results of the second stage regression using the logarithm of the total number of articles written in the story as the dependent variable are in line with the results using the logarithm of duration. The 2sls estimates in columns (4)-(6) indicate that the views received by the first article of the story leads to a statistically significant increase in the number of articles written under it. A one standard deviation increase in the views leads to a 70% - 100% increase in the number of articles written, which amounts to 1.3- 2 additional articles as part of the story.

We next estimate a two stage Tobit model. The fact that the median story has a length of zero, and hence has no follow-up articles, implies that there is a mass point at zero. The Tobit model explicitly takes this into account. Instrumenting for the views, the results in Table 7 are qualitatively and quantitatively similar to the 2sls estimates for both duration (columns (1)-(3)) and number of articles (columns (4)-(6). The average marginal effects Table 8 indicate that a unit standard deviation increase in views leads to about a 1-3.5 day increase in the length of the story (columns (1)-(3). The relationship is statistically significant at conventional levels and stable across specifications which allow for the inclusion of section controls (in columns (2) and (3))[34]. The average marginal effects in Table 8 show that a unit standard deviation increase in views leads to an increase in the number of articles by 1.7-3 written as part of the story (columns (4)-(6)) .

[33] See http://www.livemint.com/Industry/tnV2NUSAK8PbFs7pSzoL0I/India-faces-daily-power-outageof-30000-MW.html

[34] Overidentification tests are extremely complicated in non-linear models though often the inference can be carried over from the tests on linear models. For more on this see Cameron and Trivedi (2013)

The OLS estimates are smaller than the 2sls estimates across different specifications. If the instruments have a heterogeneous impact on the clicks then OLS measures the average effect of clicks across all stories while 2sls captures the effect of stories which are marginal in the clicking decision. The stories which are marginal in the clicking decision are also likely to be marginal in the coverage decision which are going to be given additional coverage, for example, only if it rains on that particular day. In essence, we are measuring the local average treatment effect (LATE) using the 2sls which would be greater than the average treatment effect (ATE).

Following the logic of Eisensee and Stromberg (2007), if the impact of rain on follow up articles is higher for stories which are marginal in the clicking decision, then the correlation between clicks and coverage would be higher for stories which get closer to the average number of clicks. Therefore, we include the interaction between views and the absolute distance of the predicted probability of the first article of the story getting clicks greater than the mean from 0.5 in the regression of views on news coverage.[35] The results in Table 9 show that coefficient of the correlation between views and articles becomes twice as large (columns (2)-(4)) as the OLS estimate (column (1)). This is extremely coarse measure of being marginal and these estimates could be substantially higher for stories which are truly on the margin. This would especially be true if on rainy days, for example, new or irregular readers visit the website more and click on articles which are relatively more salient.[36]

Interpreting the popularity estimates

To get a sense of what the magnitudes of the popularity estimates imply, we relate them to two different kinds of estimates in the existing media economics literature.

First, we relate our estimates to studies which analyze the impact of news coverage on different outcome variables. In essence, we attach another piece to this sequence by analyzing how many clicks would be needed to generate a certain amount of coverage to have a significant impact on different outcomes. In particular, we relate our estimates to those in Lim et al. (2015). Lim et al. (2015) analyze the causal impact of coverage by newspapers on criminal sentencing by U.S. state court judges. Overall, they find that 12 newspaper articles, across all newspapers, on a particular crime story or judge increase prison sentences by 3 months on average. Our preferred estimates show that a unit standard deviation increase in clicks (approximately 900

[35] Thre predicted probability is derived from a regression on rain, outages and other controls.
[36] These arguments and magnitudes are similar to those in Eisensee and Stromberg (2007).

clicks) leads to a total of 4 articles on the same story on average. Hence, in this context, a unit standard deviation increase in clicks would lead to a one month increase in prison sentence which is non-negligible[37].

Next, we compare our estimates to existing estimates analysing the impact of political preferences on coverage decisions. Puglisi and Snyder (2010) find evidence in favour of agenda setting where pro-Democratic newspapers systematically provide less coverage to corruption scandals involving Democratic politicians. In particular, they find that a one standard deviation increase in preference for a Democratic leader, measured by news endorsements, reduces negative coverage by about 30%. The magnitude of the effects, as the authors state, is small though statistically significant. Larcinese, Puglisi and Snyder (2011) find similar quantitatively small effects of political preferences on coverage of economic news[38]. In contrast, we find significantly larger effects of clicks, with a one standard deviation increase in clicks leading to a 65% increase in coverage being our most conservative estimate.

7 Clicks based Coverage and the 'Quality' of Information

After having established the causal impact of clicks on coverage, we now move to analyzing the potential consequences of this editorial strategy on the quality of information. Providing additional coverage to stories which receive a higher number of clicks implies that more popular stories might crowd out stories which do not receive similar amount of reader attention.

We would like to investigate whether editorial preferences to extend coverage of popular stories holds irrespective of the content of a story, but measuring the informational content of stories is not a trivial task. To this end, we divide the news articles into 'hard' and 'soft' news based on the section of the newspaper the story is slotted in. Patterson (2000) argues that soft news are "typically more sensational, more personality-centered, less time-bound, more practical, and more incident-based than other news". To guide story classification, we use the 'section' references of articles. Hard news consists of stories in the National, International, Business and Opinion sections while soft stories are those which are slotted in the Entertainment, Sports, Technology

[37] Another example is Fang and Peress (2009). Fang and Peress (2009) find that firms which have high media coverage earn a 0.25% - 1% lower monthly return on the stock market relative to low media coverage firms. A firm is said to have high media coverage if 4 or more articles are written on it (which is the median number) within a month. According to our estimates, a one standard deviation increase in clicks on a story particular to a firm can reduce its return by upto 1% during a calendar month, which is an economically significant effect.

[38] They find insignificant effects of reader preferences.

and Lifestyle sections of the newspaper. According to Boczkowski (2010), National, International and Business news are more informative for an individual to participate in 'public' affairs while soft news such as Sports is considered to be related to 'non public' affairs. If soft news stories are given the additional coverage based on clicks, it may potentially crowd out hard news, at least in the short run. If editorial preferences to allocate higher resources to more popular stories hold irrespective of content, the 'quality' of information provided to readers will change.

We re-estimate the model after splitting the data under the hard and soft news labels. In the data, 64.3% of all articles are slotted under a hard news section while the rest 35.7% are labeled as soft news. The first stage results in Table 10 demonstrate how the instruments perform separately for hard and soft news. These results highlight some interesting characteristics about online news reading behaviour. The rainfall instrument has a quantitatively and statistically stronger impact on hard news than soft news. On the other hand, the power outage instrument in the case of hard news is statistically and economically insignificant. Power outage has a negative and statistically significant impact on soft news. This suggests that the demand for hard news is relatively less downwardly elastic as compared to soft news. To ensure that the first stage of the instrumental variables regression is strong, we instrument views received by hard news using only rainfall (column(2)). Similarly, for soft news, we primarily focus on variants of power outages such as outages normalized by weekly power outages to ensure that our instruments are strong. One can see in column (4) that normalized outages have a statistically significant (F-statistic of 11.40) negative impact on views. In column (5), we use both normalized rainfall and power outages with rainfall having a positive effect and outages having a negative effect on views. Note that all our baseline results are robust to using these alternative measures of the instruments[39].

The second stage results displayed in Table 11 are clear: Clicks have a positive and statistically significant impact (at 5% level) on coverage of hard news, but statistically and economically insignificant impact on the coverage of soft news stories. The baseline estimate for the whole sample, with the number of articles as the dependent variable, is 0.23 while for hard news it is 0.35 and for soft news it is 0.05. This evidence shows that conditional on the first article being published, the editor publishes additional articles in response to the clicks received by hard news stories only. [40]

A concern with this finding might be that soft news are fundamentally different in their char-

[39]We also use power outages normalized by weekly outages along with small regional holidays (for a positive shock to clicks) which point to same qualitative results. Results available upon request.

[40]The results are similar for story duration.

acteristics relative to hard news and therefore, are treated differently by the newspaper. For example, if news agencies do not provide enough soft news stories then the cost incurred by the newspaper might be much higher for soft news. We look at some summary statistics to get a sense of the difference in observable characteristics between hard and soft news in Table 12. The mean (and standard deviation) across hard and soft news articles of observables such as the proportion of articles sourced from a news agency (40% for hard news as opposed to 47% for soft), the word count per article (375 for hard and 400 for soft) and the length of the headline (46 characters for hard and 42 for soft) do not differ much. The mean number of articles in a story indicate that soft stories, with 1.55 articles per story, can and do have follow up articles when compared to 1.81 articles as part of an average hard news story. The median number of articles in a story is 1 for both types of stories. These numbers seem to indicate that hard and soft news stories are not fundamentally different in their observable characteristics[41].

To get a sense of the implications of this asymmetry of editorial strategy for the 'quality' or type of news, we look at a particular counterfactual situation. We estimate the coverage that would be given to an average soft news story if the editor was using the hard news strategy for soft news as well. In particular, we use $\hat{\gamma}_1$ from the 2sls regression for hard news and simulate the coverage received by an average soft news story under that strategy. We find that the length of the average soft story, in terms of the number of articles, would increase by 11%. Hence, if the newspaper followed a symmetric strategy of following up on soft and hard news there would be a definite increase in the amount of soft stories we see in the news. This evidence seems to indicate a clear and conscious editorial strategy such that hard news dominates the amount of total news[42].

These results look at how clicks-based editorial strategy differs across different types of stories conditional on having published the first article on a particular topic. This does not quite rule out the newspaper shifting its coverage to a larger amount of soft news on account of anticipating a higher number of clicks on soft news relative to hard news. In other words, the above result does not preclude soft news crowding out hard news by reducing the overall number of hard news articles or stories rather than just follow-ups on stories. We perform an additional check

[41] One dimension across which hard and soft news stories differ is the unconditional mean of the number of clicks. The unconditional mean of the number of clicks for soft articles is much higher at 311.44 compared to 233.55 for hard news.

[42] This echoes the sentiments in the American T.V. series "The Newsroom", based on real world behind the scenes editorial decisions and events, in which an actor once remarked: "I understand how market forces work in the news, but journalists have always been the people pushing back against them... If those reporters were being paid per person reading their story, the front page of the New York Times would look a lot different".

to test for such a strategy of the newspaper. We focus on big soft and hard news events, whose occurrence is exogenous to editorial decisions, during 2012 which would have the potential to crowd out either type of news. In particular, we create a list of the biggest celebrity scandals as well as a list of all the big corruption scandals which broke out during the year. We compile a list of corruption scandals based on information on Wikipedia and for each of the seven scandals, we look at the day the scandal broke out plus one day which gives 14 scandal days. Similarly, we create a list of the biggest celebrity scandals and look at the day each of the seven scandals broke out plus one day which gives us 14 scandal days as well. To test our hypothesis that hard news crowds out soft (and not vice-versa), we first regress the number of clicks received by hard news stories as a proportion of the total number of clicks in a day on whether there was a celebrity or a corruption scandal breaking out on that particular day. As can be seen from column (1) in Table 13, on scandal days there is a decline in the proportion of clicks received by hard news stories by about 7.5% and this relationship is statistically significant at the 10% level. On the other hand, on corruption scandal days clicks received by hard news articles increases by about 4.5% relative to soft news as seen in column (2). Even though there is a decrease in hard news clicks on celebrity scandal days, we do not see a corresponding crowding out of hard news by soft news articles as seen in column (2) of Table 13. We regress the proportion of hard news articles published on a day on whether there was a celebrity scandal on that day to find a statistically and economically insignificant impact. Contrastingly, there is a 2.2% increase (significant at the 5% level) in the proportion of hard news articles written on corruption scandal days as seen in column (4).

These results further strengthen the intuition behind the seemingly asymmetric treatment of different types of news stories[43]. They imply that editorial decisions to extend coverage based on clicks vary for soft and hard news. Contrary to extant concerns about click-bait journalism, when editorial policy responds to real time reader demands, hard news are more likely to get follow ups than soft news, and not vice-versa. While we cannot highlight a concrete mechanism for why editors treat different types of stories differently, the first stage regressions do highlight a potential channel. The fact that hard news clicks are not as downwardly elastic (since power outages do not have an impact on them) as the the clicks on soft news implies that the demand for hard news is more stable than soft news. This would imply, from a long term profit maxmizing

[43]We also find similar results when we analyze more 'expected' events such as cricket matches played by the Indian team and the presentation of the annual budget by the Finance minister of the country. Results available upon request.

perspective, that the editors should cater to the preferences of their stable readership which are those who click on the hard news stories. An alternative explanation could be that hard and soft news stories are priced differently in terms of advertisements. Conversations with the editors and business heads of different newspapers (including the one which we analyse) and examining their database showed that different news pages were not priced differently.[44]

8 Crowding Out of News

In this section we discuss the impact of page views on coverage driven by rainfall and electricity shortages, and the implications it might have for information provision and the newspaper's profits. We established earlier that on average stories published on a rainy day, for example, would not have received the additional coverage had it not rained. Thus, one might claim that the newspaper is misinterpreting the additional clicks due to rain as 'true' reader interest which in turn, is driving editorial coverage.

Counter to this argument, editors may be publishing more articles on the same story because they believe that readers want to read more on stories that they are familiar with. If this holds, in other cases where there is an exogenous increase in readers' attention and clicks on news stories, editors should extend coverage as well. We look at a particular context to test this prediction: national holidays. In columns (1)-(3) of Table 14, for national holidays, there is a statistically significant increase in the number of views received by the first article of the story published on that day which corresponds to the first stage estimation. There is a 10%-13% increase in clicks on national holidays which is comparable to the increase of 8.2% due to rainfall. Statistically, the F statistic for the first stage is well beyond the required threshold of 10. For the second stage results, as seen in columns (4)-(6), we find that instrumenting clicks by national holidays does not lead to a robust positive and significant impact on coverage. In column (5), we find a positive and significant effect but this disappears as soon as we include month fixed effects in column (6).

If the newspaper was in fact exploiting readers' preference to hear more on stories that they already know, then we would expect a positive and significant impact of clicks on coverage when

[44] There were two types of ads which are used by news websites. One is a drop down banner ad generally displayed when a reader lands on the homepage. The second one is based plainly on the total number of views/unique views. We did not find any difference in the way hard and soft pages were generating revenue based on the number of views. This was the situation till the end of 2014 when we last accessed their database.

instrumented for by national holidays. The fact that we do not find evidence of this suggests that the editor, anticipating the increase in attention on these days, does not respond to clicks with additional issue coverage.

We provide a second piece of suggestive evidence in line with this hypothesis. If the clicks coming from rain do actually represent true reader preferences then we should observe an increase in the number of clicks received by the follow-up article relative to a follow up article on a story published on a non-rainy day. In Table 15, we regress whether the first article of a story was published on a rainy day on the views received by the second article of the story. In column (1) we find that, if the first article was published on a rainy day then the second article gets a significantly lower number of views (significant at the 10% level) with a magnitude of 6%. This relationship, though, is not robust as the statistical significance goes away when we add section and month fixed effects in columns (2) and (3), possibly due to the small sample size. The magnitude of the coefficients in columns (2) and (3), between 4%-6%, still remain similar to the one found in column (1). These results provide evidence that the follow up articles do not see an increase in clicks relative to those published on non-rainy days, in line with the hypothesis that the editor might be misinterpreting clicks due to rainfall.

We believe to have identified a channel in which editorial focus on clicks leads to excessive crowding out of new information which might hamper the overall information provision to the readers. Moreover, clicks, unrelated to 'true' reader demand might be detrimental to the newspaper's revenue. In particular, the fact that in the absence of rain the editor would have made different coverage decisions might imply that the newspaper is leaving some money on the table.

9 Robustness Checks

In this section, we carry out a series of checks to ensure that our results are invariant to plausible changes in the baseline specifications. First, we look at the robustness of the results to using a different though connected measure of popularity: Unique views. Second, we look at whether using power shortages as a proportion of power consumption, as opposed absolute power shortages, changes the core results. Next, we estimate duration and poisson models and compare the estimates to the 2sls and Tobit specifications. Finally, we address the issue of outliers by re-estimating our model after dropping outliers along different dimensions.

9.1 Unique Views as Popularity Measure

In our baseline analysis, we have used the total number of views received by a story as the measure of popularity. While the total number of page views is used by newspapers to price advertisements, often this measure is complemented by the number of unique views[45]. Hence, we re-estimate our model using unique views as the indicator of story appeal.

In Table 16, in the first stage regressions, we find that the instruments are still strong with both rain and power outages being correlated with the F statistic for rainfall and the power outage instruments is over 10 for all specifications. In the second stage in columns (4) and (5), we find that the impact of unique views is extremely similar to that of total views. For example, when the dependent variable is the length of the story and the coefficient on unique views is 2.04 which is very close to the estimate of 2.06 found in the specification using total views. When the dependent variable is the number of articles in a story, the coefficients on unique views is 0.21 (column 4) , which are comparable in magnitude to our baseline results, economically and statistically. Thus, our benchmark results are robust to using unique views as an alternate measure of popularity.

9.2 Using Within-Month Variation in Rainfall and Outages

In the main model, total power outage in Delhi and Maharashtra are used as an instrument along with rainfall across Delhi and Mumbai. Controlling for month fixed effects accounts for seasonal factors which effect both rainfall and power outages, but does not take into account the variation in the severity of power shortages or rainfall within a month. To explicitly take into account within-month heterogeneity in power outages, we use the total daily power shortages in Delhi and Maharashtra as a proportion of the total daily consumption across the two cities[46]. To account for within-month variation in rainfall, we re-estimate our baseline specifications using the total daily rainfall across Delhi and Mumbai as a proportion of the total monthly rainfall across the two cities.

The results, displayed in Table 17, are remarkably similar to those in our baseline setting. The direction and magnitudes of the effect of views on the length of the story (column (1)) and the number of articles as part of the story (column (2)), when using power outages as a share of

[45] Conversations with the owners as well as the editor of the newspaper pointed in this direction.

[46] On average, the daily power shortage across Delhi and Maharashtra is about 5% of the total daily consumption.

consumption, are close to our basic specifications given in Tables 6 and 7. For example, the 2sls estimate (when the coverage variable is story length) is 2.09 which is almost exactly the same as the baseline estimate. Similarly when the outcome variable is number of articles, the estimate is 0.238 which is close to the value of 0.237 from our baseline setting. In columns (3) and (4), we use the total daily rainfall as a proportion of the total monthly rainfall as the instrument instead of using a rain dummy as in the baseline specifications. As in the case of power outages, the results are qualitatively and quantitatively similar to our baseline results. These robustness checks validate our findings from the IV strategy.

9.3 Duration and Poisson Model Estimates

As a robustness check, we first re-estimate our specifications using an Accelerated Failure Time (AFT) model. Even though such duration models are not easily amenable to endogeneity issues (Gowrisankaran and Town (1999), Bijwaard (2008)), we present the results here to check whether our estimation results stand the test of a different estimation technique.

We run the first stage regression to instrument for clicks and then take the predicted values and estimate a duration model in the second stage. Columns (1)-(3) of Table 18 shows the hazard rates of the story, in terms of the number of articles, with respect to the views received. The coefficient on the logarithm of predicted views is positive and statistically significant at the 1% level. Importantly, the magnitude of this coefficient is less than 1 which implies that stories with a higher number of clicks have a lower hazard rate, or, equivalently, conditional on getting a larger number of clicks, stories will last longer. The results from a hazard model with story length instead of the number of articles yields qualitatively similar results, as can be seen from columns (4)-(6) of Table 18.

As an additional check, we also estimate a poisson model where the number of articles is the dependent variable. This is appropriate considering that this is a count variable. Similar to earlier analyses, we use the predicted values of views from the first stage and estimate its impact on coverage in the second stage. Table 19 shows that the logarithm of predicted views has a positive and significant impact at the 1% level. Moreover, the estimates are stable across specifications, even when month fixed effects are introduced (column (3)).

These results give us further confidence in the baseline estimates. The AFT and Poisson models imply that the qualitative takeaway from these results is the same as what we saw in the 2sls or Tobit estimations.

9.4 Dealing with Outliers

An additional concern is that our results might be driven by certain outlier topics or issues which would receive a lot of media attention such as the Olympics or the Presidential elections. To ensure that these results are not driven by such outliers, we undertake two robustness checks.

First, we create a dummy variable to capture whether the story has more than one article or equivalently whether it was continued further or not. We use this as our dependent variable of interest as opposed to the total number of articles or the duration in the baseline specifications. This allows us to circumvent the impact of outliers which could have a significant effect when looking at the average number of articles in a story or its total duration. We estimate a linear probability model, instrumenting the clicks, to find that clicks still have a positive and statistically significant impact on the probability of being given additional coverage. The results displayed in column (1) of Table 20 provide greater confidence in our baseline results. Since the dependent variable is a 0-1 variable, all stories, irrespective of the number of articles or duration, are given the same weight.

As a further robustness check, we drop all stories which have more than 50 articles in their cluster and re-estimate our model. The results are displayed in column (2) of Table 20. Qualitatively and quantitatively, the estimates are very close to our baseline numbers.

9.5 Rain and Outages in the News

Finally, an issue with the estimations could be that rain and power outages could themselves be part of the news and hence might be driving our results. To address this, we undertake two robustness checks. We first drop any days which have a high amount of rainfall or power outages, which would make it more likely for it to be in the news. In column (3) of 20, we drop any days which have rainfall of more than 60 mm across Delhi and Mumbai while in column (4) we drop any days where the (log) power outages are above 4.3[47]. The results remain unchanged relative to our baseline estimates. These estimations corroborate the robustness of our results and show that they are not driven by a handful of atypically long stories or days of extremely high rainfall or power outages and the possibility of them being in the news.

[47] These correspond to the values over the 95th percentile. For example, floods typically take place if there is more than 75mm of rainfall within 24 hours (See http://www.marlborough.govt.nz/Services/Emergency-Management/Flooding.aspx).

Moreover, we run a keyword search and drop any articles which have the words 'rain', 'storm', 'outage' or 'power cut' as part of their headline. This would indicate whether the article is primarily about rainfall or power outages. These stories do not seem to drive our results as can be seen in columns (5) and (6) where the estimates are in line with our baseline numbers.

10 Conclusion

In this paper, we quantify the extent to which newspapers systematically respond to reader preferences. We provide the first evidence to show that editors respond to page views by giving more popular stories extended coverage in terms of duration and number of articles. Moreover, we find that, contrary to popular perception, if there is any crowding out of news due to clicks then it is hard news which crowds out soft news and not vice-versa. Finally, our identification strategy highlights the possibility that when editorial decisions are guided by noisy measures of demand such as page views, this can be detrimental to news provision as well as the profits of the newspaper.

We are able to address these issues using a unique dataset from a leading Indian daily which includes article level information such as the number of views and unique views received. We combine this information with data collected using a crawler such as the text of the article, when it was first published, the source of the article, and other available characteristics. Next, we employ an algorithm which clusters articles into 'stories' which become the unit of observation. We address the causal link between views and coverage by using exogenous variation in reader attention by focusing on rainfall and power shortages. We focus on rainy days since on these days people may remain indoors and may go online and read the news, whereas on days with power shortages readers may face interruptions in electricity supply and may have restricted access to the internet. Indeed, we find that stories whose first article is published on a rainy day (or a day with high power shortages) get significantly larger (smaller) number of views relative to other normal days. Stories which receive a higher number of clicks on their first article were given more coverage. Our IV strategy helps to establish the explanation that editors systematically allocate greater coverage to relatively more popular stories.

Apart from providing the first evidence to shed light on the factors affecting editorial decisions on a day to day basis at the level of a story, our results speak to the concern of the 'dumbing down' of news online. We find that the editors allocate a larger amount of resources to hard

news stories even if soft news get a larger number clicks. Assuming that hard news stories are more informative to readers in public life, we find that the concerns around clicks based editorial decisions might be misplaced. Our results also have implications for firm strategy. Our identification strategy shows that clicks and other online measures of popularity are noisy and are often driven by exogenous events such as rain and power outages which are divorced from real reader interest in the news content produced. It is, thus, imperative for newspapers to identify the potential sources of noise so that they can maximize their profits. Furthermore, this will help with the newspaper's function of information provision for its readers.

Finally, the fact that choosing news stories and their follow-up articles based on popularity can bias the beliefs of readers has consequences for media policy. In essence, since newspapers would systematically follow up on more popular topics such as airplane crashes, crime, corruption scandals etc. then the rate of coverage will inevitably diverge from the true rates of occurrence of these events. Trying to adopt a more balanced approach through content self regulation by the media itself is advisable. Here, we point to self regulation rather than regulation through legislation so as not to infringe upon media freedom in modern democracies. Moreover, catering to popular demand implies that minority interests are sidelined which restricts viewpoint diversity, an issue which has been given serious attention by F.C.C. in the U.S. We hope this study will inform media practioners and policy makers who are concerned about these issues at large.

References

[1] Best, M. 2009. "If It Bleeds It Leads: Sensational Reporting, Imperfect Inference and Crime Policy". Mimeo.

[2] Bhaskar, V. 2009. "Rational Adversaries? Evidence from Randomised Trials in One Day Cricket". The Economic Journal (119): 1-23.

[3] Bijwaard, G. E. 2008. "Modeling migration dynamics of immigrants: the case of The Netherlands". Tinbergen Institute Discussion Paper.

[4] Boczkowski, P. J. 2010. "News at Work: Imitation in an Age of Information Abundance". University of Chicago Press.

[5] Cage, J. 2014. "Media competition, information provision and political participation". Mimeo, Harvard University.

[6] Cameron, A. C., and Trivedi, P. K. 2013. "Regression Analysis of Count Data". Cambridge University Press.

[7] Chan, J. and W, Suen. 2008. "A Spatial Theory of News Consumption and Electoral Competition". Review of Economic Studies (75): 699-728.

[8] Eisensee, T. and Strï¿œmberg, D. 2007. "News Droughts, News Floods, and U.S. Disaster Relief".The Quarterly Journal of Economics (122): 693-728.

[9] Fang, L., and Peress, J. 2009. "Media Coverage and the Cross-section of Stock Returns". The Journal of Finance (64): 2023-2052

[10] Foster, G. & Kalenkoski, C. M. 2013. "Tobit or OLS? An empirical evaluation under different diary window lengths". Applied Economics (45): 2994-3010.

[11] Franceschelli, I. 2011. "When the Ink is Gone The Impact of the Internet on News Coverage". Mimeo, Northwestern University.

[12] Gentzkow, M. and Shapiro, J. M. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers". Econometrica (78): 35-71.

[13] Gentzkow, M., Shapiro, J. M. and Sinkinson, M. 2011. "The Effect of Newspaper Entry and Exit on Electoral Politics". American Economic Review (101): 2980-3018.

[14] Gowrisankaran, G., and Town, R. J. 1999. "Estimating the quality of care in hospitals using instrumental variables". Journal of health economics (18): 747-767.

[15] Katona, Z., Knee, J. A. and Sarvary, M. 2013. "Agenda Chasing and Contests Among News Providers". Mimeo, Columbia Business School.

[16] Larcinese, V., Puglisi, R. and Snyder Jr., J. M. 2011. "Partisan bias in economic news: Evidence on the agenda-setting behavior of U.S. newspapers". Journal of Public Economics (95): 1178-1189.

[17] Latham, O. 2015. "Lame ducks and the media". Fortcoming, Economic Journal.

[18] Lim, C., Snyder, J. and Strï¿œmberg, D. 2015. "The Judge, The Politician, and the Press: Newspaper Coverage and Criminal Sentencing Across Electoral Systems". American Economic Journal: Applied Economics (7): 103-135.

[19] Mitchell, T. 2015. "Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression". Machine Learning (Second Edition). McGraw Hill.

[20] Mullainathan, S. and Shleifer, A. 2005. "The Market for News". American Economic Review (95): 1031-1053.

[21] Patterson, T. E. 2000. "Doing Well and Doing Good: How Soft News and Critical Journalism are Shrinking the News Audience and Weakening Democracy - and what News Outlets Can Do about it". Harvard University.

[22] Pence, K. M. 2006. "The role of wealth transformations: An application to estimating the effect of tax incentives on saving". Contributions in Economic Analysis & Policy (5):

[23] Prat, A. and Strï¿œmberg, D. 2013. "The Political Economy of Mass Media". Mimeo, CEPR Discussion Papers.

[24] Prior, M. 2005. "News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout". American Journal of Political Science (49): 577–592.

[25] Puglisi, R. and Snyder, J. 2010. "Newspaper Coverage of Political Scandals". Journal of Politics (73): 931-950.

[26] Radev, D. R., Hatzivassiloglou, V. and McKeown, K. R. 1999. "A description of the CIDR system as used for TDT-2". Broadcast News Workshop'99 Proceedings (205).

[27] Romer, D. 2006. "Do firms maximize? Evidence from professional football". Journal of Political Economy (114.2): 340-365.

[28] Salami, A., & Seamans, R. 2014. "The Effect of the Internet on Newspaper Readability". Mimeo, Net Institute (14-13).

[29] Singer, J. B., Domingo, D., Heinonen, A., Hermida, A., Paulussen, S., Quandt, T., & Vujnovic, M. 2011. "Participatory journalism: Guarding open gates at online newspapers". John Wiley & Sons.

[30] Stock, J. H. & Yogo, M. 2002. "Testing for Weak Instruments in Linear IV Regression". Mimeo. National Bureau of Economic Research.

Table 1: Example for Consecutive Articles Grouped into a Cluster

Headline	Date Published
Japan regains nuclear power after reactor restarts	5^{th} July, 2012
Fukushima was 'man-made' disaster: Japanese probe	5^{th} July, 2012
Commission calls Fukushima n-crisis man-made disaster	6^{th} July, 2012
'Man-made'	7^{th} July, 2012
Fukushima lessons	7^{th} July, 2012

Table 2: Summary Statistics

Variable	Observations	Mean	Std. Dev.	Min	Max
Story length (hours)	60,686	4.79	26	0	2030.23
Number of articles in a story	60,686	1.70	7.34	1	712
Views: First article	60,686	217.9	695.38	1	59460
Length of headline (characters)	60,686	44.17	14.99	3	244
Word count: First Article	60,686	382.6	260.6	17	8857
Daily power outage (mega units)	60,686	24.61	25.75	0.02	90.18

Table 3: OLS Estimates for Duration and Number of Articles

VARIABLES	(1) log(length)	(2) log(length)	(3) log(length)	(4) log(articles)	(5) log(articles)	(6) log(articles)
log(views)	0.300***	0.300***	0.297***	0.0225***	0.0225***	0.0222***
	(0.0183)	(0.0183)	(0.0184)	(0.00134)	(0.00134)	(0.00134)
Section fe	N	Y	Y	N	Y	Y
Month fe	N	N	Y	N	N	Y
Observations	60,671	60,671	60,671	60,671	60,671	60,671
R-squared	0.012	0.015	0.016	0.016	0.020	0.024

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether the article was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. Columns (1)-(3) has the length of the story as the dependent variable while columns(4)-(6) has the total number of articles.

Table 4: First Stage Estimates

VARIABLES	(1) log(views)	(2) log(views)	(3) log(views)
Rain	0.0586***	0.0583***	0.0842***
	(0.0141)	(0.0142)	(0.0197)
log(outage)	-0.0172***	-0.0174***	-0.0212**
	(0.00646)	(0.00651)	(0.0102)
Section fe	N	Y	Y
Month fe	N	N	Y
First Stage F-Stat	15.93	15.91	10.68
Observations	60,671	60,671	60,671
R-squared	0.167	0.167	0.171

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. The dependent variable is the logarithm of the views received by the first article of a story.

Table 5: Placebo Checks for Instruments

VARIABLES	(1) PTI	(2) Log(word count)	(3) Hard	(4) Total Articles
Rainfall	0.00298	0.00319	0.006	0.787
	(0.00190)	(0.00548)	(0.0055)	(2.608)
Log(outage)	-0.000795	-0.000785	0.001	0.456
	(0.00104)	(0.00290)	(0.002)	(1.318)
Weekend	0.00863***	0.0160***	-0.021***	-13.93***
	(0.00157)	(0.00453)	(0.0044)	(2.560)
Observations	100,770	100,770	60,688	324
R-squared	0.022	0.033	0.182	0.980

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include month fixed effects. Columns (1) and (2) include section controls while column (3) includes the total words written across all articles published in the day. Column (1) has whether the article is sourced from the PTI or not as the dependent variable while column (2) has words per article as the dependent variable. Column (3) has the total number of articles published on a day as the dependent variable.

Table 6: 2sls Results for Story Duration and Articles

VARIABLES	(1) log(length)	(2) log(length)	(3) log(length)	(4) log(articles)	(5) log(articles)	(6) log(articles)
log(views)	3.018***	2.800***	2.061**	0.312***	0.293***	0.237***
	(0.920)	(0.901)	(1.045)	(0.0755)	(0.0733)	(0.0847)
Section fe	N	Y	Y	N	N	Y
Month fe	N	N	Y	N	Y	Y
Over-id (p value)	0.995	0.128	0.133	0.588	0.587	0.622
Observations	60,671	60,671	60,671	60,671	60,671	60,671

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. The dependent variable is the duration of the story in columns (1)-(3) while it is the logarithm of the total number of articles as part of the story in columns (4)-(6).

Table 7: 2 Stage Tobit Results for Story Duration and Articles

VARIABLES	(1) log(length)	(2) log(length)	(3) log(length)	(4) log(articles)	(5) log(articles)	(6) log(articles)
log(views)	15.34***	15.31***	10.48*	4.15***	4.139***	1.018**
	(4.625)	(5.067)	(6.064))	(1.253)	(1.370)	(0.4048)
Section fe	N	Y	Y	N	N	Y
Month fe	N	N	Y	N	Y	Y
Observations	60,671	60,671	60,671	60,671	60,671	60,671

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not, as well as weekend fixed effects. The dependent variable is the logarithm (inverse hyperbolic sine transformation) of the length of the story in columns (1)-(3) while it is the logarithm of the number of articles in columns (4)-(6).

Table 8: Average Marginal Effects for Story Duration and Number of Articles

VARIABLES	(1) log(length)	(2) log(length)	(3) log(length)	(4) log(articles)	(5) log(articles)	(6) log(articles)
log(views)	1.537***	1.616***	1.024**	0.460***	0.493***	0.416**
	(0.368)	(0.429)	(0.408)	(0.110)	(0.131)	(0.167)
Section fe	N	Y	Y	N	N	Y
Month fe	N	N	Y	N	Y	Y
Observations	60,671	60,671	60,671	60,671	60,671	60,671

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. The dependent variable is the logarithm (inverse hyperbolic sine transformation) of the length of the story in columns (1)-(3) while it is the logarithm of the number of articles in columns (4)-(6).

Table 9: Heterogeneous Impact of Rain and Outages

VARIABLES	(1) log(articles)	(2) log(articles)	(3) log(articles)	(4) log(articles)
log(views)	0.0222***	0.0403***	0.0414***	0.0402***
	(0.00134)	(0.003)	(0.003)	(0.003)
log(views)x($\hat{prob}$-0.5)		-0.062***	-0.067***	-0.063***
		(0.009)	(0.009)	(0.009)
$\hat{prob}$-0.5		-0.569***	-0.527***	-0.029***
		(0.191)	(0.193)	(0.355)
Section f.e.	Y	N	Y	Y
Month f.e.	Y	N	N	Y
Observations	60,671	60,671	60,671	60,671
R-squared	0.185	0.017	0.021	0.025

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. The dependent variable is the logarithm of the number of articles in a story.

Table 14: National Holidays as the Intrument

VARIABLES	(1) log(views)	(2) log(views)	(3) log(views)	(4) log(articles)	(5) log(articles)	(6) log(articles)
Holidays	0.134***	0.135***	0.109***	0.460***	0.493***	0.416**
	(0.023)	(0.023)	(0.024)	(0.110)	(0.131)	(0.167)
log($\hat{views}$)				0.085	0.127**	0.055
				(0.057)	(0.058)	(0.068)
Section fe	N	Y	Y	N	N	Y
Month fe	N	N	Y	N	Y	Y
First Stage F-stat	32.34	32.5	20.62	-	-	-
R-squared	0.166	0.166	0.1707	-	-	-
Observations	60,671	60,671	60,671	60,671	60,671	60,671

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not. The dependent variable is the logarithm of the views received by the first article of a story in columns (1)-(3) which captures the first stage regression including holidays as the instrument. The logarithm of the total number of articles in the story is the dependent variable in columns (4)- (6) which captures the second stage estimation.

Table 10: Quality of Information: First Stage

VARIABLES	Hard log(views)	Hard log(views)	Soft log(views)	Soft log(views)	Soft log(views)
rain	0.0819***	0.0808***	0.0757**		
	(0.0241)	(0.0241)	(0.0336)		
log(outages)	-0.00861		-0.0410***		
	(0.0133)		(0.0159)		
log(normalized outages)				-0.0623***	-0.0678***
				(0.0184)	(0.0186)
log(normalized rain)					0.0131***
					(0.00476)
Month FE	Y	Y	Y	Y	Y
F-Stat	5.85	11.25	5.38	11.40	9.30
Observations	36,101	36,101	24,570	24,570	24,570
R-squared	0.140	0.140	0.223	0.223	0.223

Notes: The fist two columns display the first stage results for hard news while the last three are the results for soft news. The dependent variable is the logarithm of total views. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects.

Table 11: Quality of Information: Second Stage

VARIABLES	Hard ln(articles)	Hard ln(length)	Soft ln(articles)	Soft ln(articles)	Soft ln(length)	Soft ln(length)
log(views)	0.358**	4.642**	0.0114	0.0531	-1.178	-0.375
	(0.145)	(1.912)	(0.0851)	(0.0660)	(1.301)	(0.971)
Month FE	Y	Y	Y	Y	Y	Y
Observations	36,101	36,101	24,570	24,570	24,570	24,570
R-squared	-	-	0.018	0.004	-	-

Notes: The fist two columns display the second stage results for hard news while the last three are the results for soft news. The dependent variable in columns (1), (3)-(4) is the logarithm of the number of articles while it is logarithm of the story length in columns (2), (4)-(5). In columns (1)-(2), log(views) is instrumented by a rainfall dummy while in columns (3)-(5) the insttruments are normalized measure of power outages and rainfall. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects.

Table 12: Summary Statistics for Hard and Soft News

	Hard News		Soft News	
Variable	Mean	S.D	Mean	S.D.
Number of articles in a story	1.81	7.77	1.55	6.66
Length of headline	46.07	13.74	42.84	15.94
Word count	375.16	261.9	400.5	269.5
News Agency	0.40	0.49	0.47	0.49
Number of clicks per article	233.55	625.8	311.4	888.3

Table 13: Quality of Information: Scandals and Scams

	(1)	(2)	(3)	(4)
VARIABLES	$\frac{hardclicks}{totalclicks}$	$\frac{hardarticles}{totalarticles}$	$\frac{hardclicks}{totalclicks}$	$\frac{hardarticles}{totalarticles}$
Celebrity	-0.0781*	-0.0049		
	(0.0437)	(0.0120)		
Corruption			0.0441*	0.0204**
			(0.0236)	(0.0080)
Observations	366	366	366	366
R-squared	0.669	0.914	0.669	0.915

Notes: The dependent variable in column (1) and (3) is the proportion of clicks received by hard news articles in a day while in columns (2) and (4) it is the proportion of hard news articles published in a day. All columns include controls for the proportion of articles with images on the day, proportion of articles sourced from the PTI and the average word count per article on the day as well as weekend and month fixed effects. Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

Table 15: Views on Second Article

	(1)	(2)	(3)
VARIABLES	log(views#2)	log(views#2)	log(views#2)
I^{st}Rain	-0.0612*	-0.0397	-0.0625
	(0.0355)	(0.0356)	(0.0654)
Section fe	N	Y	Y
Month fe	N	N	Y
Observations	11,112	11,112	11,112
R-squared	0.058	0.061	0.068

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline. The dependent variable is the logarithm of the views received by the second article of a story in column.

Table 16: Unique Views as the popularity measure

VARIABLES	(1) log(unique views)	(2) log(unique views)	(3) log(unique views)	(4) log(length)	(5) log(articles)
Holidays	0.0568***	0.0572***	0.0812***	0.460***	0.493***
	(0.0141)	(0.0142)	(0.0197)	(0.110)	(0.131)
Log(outage)	-0.0231***	-0.0228***	-0.0228***		
	(0.00645)	(0.00650)	(0.0102)		
$\log(\hat{uniqueviews})$				2.041*	0.240***
				(1.064)	(0.086)
Section fe	N	Y	Y	Y	Y
Month fe	N	N	Y	Y	Y
First Stage F-stat	19.26	19.26	10.25	-	-
R-squared	0.166	0.166	0.17	-	-
Observations	60,671	60,671	60,671	60,671	60,671

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. The dependent variable is the logarithm of the unique views in columns (1)-(3) while column (4) has the length of the story as the dependent variable while in (5) we have the number of articles in the story as the dependent variable.

Table 17: Accounting for Within-Month Variation

VARIABLES	(1) log(length)	(2) log(articles)	(3) log(length)	(4) log(articles)
log(views)	2.098**	0.237***	2.357**	0.251***
	(1.061)	(0.0856)	(1.006)	(0.081)
Section f.e.	Y	Y	Y	Y
Month f.e.	Y	Y	Y	Y
Over-id (p value)	0.13	0.62	0.11	0.58
First stage F stat	10.42	10.42	12.44	12.44
Observations	59,895	59,895	60,671	60,671

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. Columns (1) and (3) have the length of the story as the dependent variable while in (2) and (4) we have the number of articles in the story as the dependent variable. Columns (1) and (2) are estimated using 2sls in which the instruments are rainfall and total daily power outages in Delhi and Maharashtra as a share of total daily consumption. Columns (3) and (4) are using total daily rainfall as a proportion of the total monthly rainfall across Delhi and Mumbai and total daily power outages in Delhi and Maharashtra as instruments.

Table 18: Duration model for Articles and Length

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	log(articles)	log(articles)	log(articles)	log(length)	log(length)	log(length)
log(views)	0.373***	0.254***	0.171***	0.284***	0.113***	0.073***
	(0.049)	(0.027)	(0.022)	(0.037)	(0.012)	(0.009)
Section f.e.	N	Y	Y	N	Y	Y
Month f.e.	N	N	Y	N	N	Y
Observations	60,688	60,688	60,688	60,688	60,688	60,688

Notes: Standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. The dependent variable is the story length in columns.

Table 19: Poisson Model

VARIABLES	storylength	storylength	storylength
log(views)	1.522***	1.474**	2.711***
	(0.538)	(0.630)	(1.001)
Section f.e.	N	Y	Y
Month f.e.	N	N	Y
Observations	60,688	60,688	60,688

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. All specifications are estimated using 2sls. Column (1) has the dependent variable which measures whether the story was continued or not (= 1 if it was continued and 0 if not). In columns (2)-(6), the dependent variable is the number of articles in the cluster.

Table 20: Outliers, and Articles about Rain&Outages

VARIABLES	(1) Continue Story	(2) log(articles)	(3) log(articles)	(4) log(articles)	(5) log(articles)	(6) log(articles)
log(views)	0.373***	0.254***	0.171***	0.284***	0.285***	0.230***
	(0.049)	(0.027)	(0.022)	(0.037)	(0.072)	(0.085)
Section f.e.	N	Y	Y	N	Y	Y
Month f.e.	N	N	Y	N	N	Y
Overid (*p* value)	0.92	0.60	0.74	0.56	0.02	0.58
First stage F stat	15.94	15.94	11	11.20	15.95	10.31
Observations	60,671	60,581	59,815	53,723	60,688	60,688

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. All specifications are estimated using 2sls. Column (1) has the dependent variable which measures whether the story was continued or not (= 1 if it was continued and 0 if not). In columns (2)-(6), the dependent variable is the number of articles in the cluster.

Data Appendix

Proprietary Data from the Newspaper

Page views: The number of times a particular web page was accessed during a session. If a user starts on page A and then moves on to page B and then moves back to page A, then page A gets two page views while page B gets one.

Unique page views: This measure does not take into account the multiple views that a web page might receive. Thus if a user starts on page A and then moves to page B after which she moves back to page A, then both page A and B have one unique page view each while page A still has two page views.

Table 21: Examples Stories

Story	First article	Last article	# of Articles
NDA protest against price hike	31 May, 4:01 am	1 Jun, 3:39 am	2
Junagadh temple stampede	20 Feb, 12:30 pm	21 Feb, 2:01 am	4
Bill Gates's visit to India	31 May, 5:51 am	2 Jun, 2:14 am	5
Vice Presidential Elections	14 Jul, 4:05 pm	15 Jul, 9:39 am	10
Mohamed Nasheed's resignation	10 Feb, 10:40 pm	15 Feb, 9:12 pm	14
India in the T20 Cricket World Cup	1 of Oct, 9:57 am	4 of Oct, 9:17 am	17
U.S Fed intervention and its impact	10 Sept, 7:47 pm	15 Sept, 3:34 am	22
Gurudwara Shootings in the U.S	6 Aug, 6:56 am	9 Aug, 4:17 pm	37
Hurricane Sandy	30 Oct, 1:48 am	10 Nov, 10:34 am	117
Hu Jintao and Sino-India relations	1 Nov, 3:16 pm	30 Nov, 1:20 am	253
The Budget and its controversies	9 Mar, 6:41 pm	2 Apr, 2:06 am	305
Presidential Elections and related issues	5 Jun, 3:11 am	2 Aug, 11:59 am	444
2012 London Olympics	13 Jul, 4:17 pm	19 Aug, 6:40 pm	510

Table 22: Variable Descriptions

Variable Name	Description
log(views)	Logarithm of the views received by an article
log(articles)	Logarithm of the number of articles part of a story
log(length)	Logarithm of the length/duration of a story
Rain	1 if it rained in either Delhi or Mumbai on that particular day
Weekend	1 if the day of the week was either Saturday or Sunday
Headline	Number of characters in the headline
log(word count)	Number of words in the article
PTI	1 if the article was sourced from the Press Trust of India
Image	1 if the article had a picture on the page
log(unique views)	Logarithm of the number of unique views received by the article
Time on page	Average amount of time spent on the article by readers
Sports	1 if the article had a tag which was 'sports'
National	1 if the article had a tag which was 'nation news'
World	1 if the article had a tag which was 'world news'
Entertainment	1 if the article had a tag which was 'entertainment'
Crime	1 if the article had a tag which was 'crime'
Delhi	1 if the article had a tag which was 'Delhi'
Cricket	1 if the article had a tag which was 'cricket'

Additional Robustness Checks

Robustness of the Results to Similarity Threshold

The results reported in the manuscript relies on a similarity threshold of 0.3 in clustering the articles. This threshold was determined with the goal of balancing type I and type II errors in the clustering process and to achieve meaningful clusters as much as possible. To make sure our results are robust to alternative threshold levels, we re-ran the baseline estimations using a tighter threshold of 0.4. The results are provided in Table 23. The first observation is, as expected, the number of stories increased to 63,952. Since 0.4 is a more stringent threshold for clustering, a higher number of articles stand as clusters of their own[48].

The results are qualitatively and quantitatively extremely similar to the ones in the baseline setting. The rain and power outage instruments are statistically significant and working in the right direction. The F statistic is always greater than 10, which satisfies the rule of thumb. Looking at the second stage results in Table 24, first when the coverage variable is the length of the story, we find that the results are in line with what we saw in the baseline model. In the 2sls, we find a positive and significant effect of clicks and the coefficients are between 1.5 and 2.5 in magnitude corresponding to baseline model's findings. Turning

[48]We only present the robustness results for the threshold value 0.4. Results when the threshold value is 0.5 are similar and available upon request.

to the number of articles written in the story, we also find the 2sls are in line with the results presented in Sections 5.2.3 and 5.2.4.

Robustness of the Dictionary of Proper Nouns

As mentioned in the section where we explain the clustering algorithm above, we use natural language processing (NLP) techniques to extract proper nouns from the text of each article. NLP techniques work by taking one sentence at a time and treating each sentence as a string of words. It then identifies proper nouns based on the position of nouns, verbs, and adjectives in a sentence. This method, even though at the frontier of the computer science literature, involves measurement errors. To get more confidence in our estimates, we re-run our model after modifying the dictionary of proper nouns extracted from the articles.

A worry might be that if the number of proper nouns extracted from an article may not be too large then it might attach itself to certain other articles in an almost random manner. We drop any articles which had less than five unique proper nouns (even if they occur multiple times) and then re-compute the duration of the story as well as the number of articles in the story. We then re-estimate our 2sls specifications which can be seen in Table 25. The results are in line with our main estimates.

In another robustness check, we drop all stories which start with an article which has less than five unique proper nouns. This, again, serves as a robustness check to see if our results are driven by articles which might be randomly attached to the start of a story. The results are reported in Table 26. The results are in line with our baseline estimates in terms of both direction and magnitude.

Table 23: First Stage for 0.4 Similarity Threshold

VARIABLES	(1) log(views)	(2) log(views)	(3) log(views)
Rain	0.0609***	0.0606***	0.0827***
	(0.0138)	(0.0139)	(0.0193)
log(outage)	-0.0210***	-0.0211***	-0.0201**
	(0.00637)	(0.00642)	(0.0101)
Section f.e.	N	Y	Y
Month f.e.	N	N	Y
F stat	20.11	20.10	10.54
Observations	63,952	63,952	63,952
R-squared	0.169	0.169	0.173

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects.

Table 24: Second Stage for 0.4 Similarity Threshold

VARIABLES	(1) log(length)	(2) log(length)	(3) log(articles)	(4) log(articles)
log(views)	2.197***	1.782*	0.309***	0.212***
	(0.767)	(1.041)	(0.0666)	(0.0817)
Section f.e.	N	Y	N	Y
Month f.e.	N	Y	N	Y
Observations	63,952	63,952	63,952	63,952

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. All specifications are estimated using 2sls. Column (1) has the dependent variable which measures whether the story was continued or not (= 1 if it was continued and 0 if not). In columns (2)-(6), the dependent variable is the number of articles in the cluster.

Table 25: Dictionary of Proper Nouns I

VARIABLES	(1) log(length)	(2) log(length)	(3) log(articles)	(4) log(articles)
log(views)	2.691***	2.300**	0.310***	0.279***
	(0.926)	(0.1064)	(0.0784)	(0.0890)
Section f.e.	N	Y	N	Y
Month f.e.	N	Y	N	Y
First stage F stat	14.65	10.64	14.65	10.64
Observations	54,088	54,088	54,088	54,088

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. All specifications are estimated using 2sls. Columns (1) and (2) have the duration of the story as the dependent variable while in columns (3) and (4), the dependent variable is the number of articles in the cluster.

Table 26: Dictionary of Proper Nouns II

VARIABLES	(1) log(length)	(2) log(length)	(3) log(articles)	(4) log(articles)
log(views)	2.953***	2.221**	0.279***	0.249***
	(0.999)	(1.801)	(0.0760)	(0.0846)
Section f.e.	N	Y	N	Y
Month f.e.	N	Y	N	Y
First stage F stat	13.83	10.62	13.83	10.62
Observations	55,120	55,120	55,120	55,120

Notes: Robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. All columns include controls for the *log*(word count) of the article, whether it was sourced from the PTI, the length of its headline, whether it had an image or not as well as weekend fixed effects. All specifications are estimated using 2sls. Columns (1) and (2) have the duration of the story as the dependent variable while in columns (3) and (4), the dependent variable is the number of articles in the cluster.

Chapter III: Statistical Externalities and Online Markets

with Paul Seabright*

1 Introduction

Can too much choice make decision makers worse off? In this paper, we address this question in the context of online markets. We show that a technology which expands the choice set of consumers or decision makers can actually make them worse off. We argue that this is extremely relevant with the rise of the internet in general and digitization in particular. On the one hand, the internet enables an infinite shelf space which reduces entry costs for new products that in turn allows consumers to access a 'long tail' of products. On the other hand, lower entry costs could reduce the average quality of the overall pool of products, offset the former effect and possibly even lower the expected payoff of the decision makers. The intuition is that as the cost of entry declines, the newer products may be of a lower quality than those which were already present on the market. This may make the consumers worse off if they do not evaluate the quality of the products at all, and may do so, if the negative quality effect outweighs the option value effect, even when they do evaluate the quality in equilibrium.

The phenomenon described by the model is relevant to various technological developments in the digital age. Digitization allows consumers access to a long tail of new products. Despite the potential benefits of digitization, it is widely believed that the long tail of products has created a new form of scarcity on the side of the consumers - the scarcity of attention. Many people feel inundated with much more information than they can effectively process (see Klingberg, 2010). But an even more fundamental problem is statistical - if our limited information processing

*Toulouse School of Economics, Institute for Advanced Studies at Toulouse and CEPR.

capacities were the only difficulty, an effective solution would consist of processing only a random subset of the information we receive. The reason this is not a real solution in practice is that as we receive more information, random subsets of that information might be less valuable to us than the more limited information used to be.

In the standard models of online markets, such as in Brynjolfsson et al. (2003) and Waldfogel and Aguiar (2015), an increase in the number of new products (books or music available online) increases surplus by expanding the choice set of the consumers. Such a phenomenon is also present in matching models of labour markets since an increase in the number of (online) job applications for a vacancy increases the pool of applicants available for the firm to choose from. Our model provides a simple story about how consumers choose a product (or firms choosing a candidate to fill a vacancy) and highlight a particular friction involved in the process to which attention has not been drawn till now. We consider a model with products of heterogeneous quality and where consumers can evaluate quality at a cost. We allow for an extremely flexible screening procedure which resembles the process of drawing balls from an urn without replacement.

In our setting, a fall in the cost of entry can induce a rise in the number of products available that reduces the payoff to the consumer. This can happen if the new products reduce the average quality of the existing pool of products available to a particular consumer. This can be true, most obviously, if the consumer does not screen for product quality, provided that the marginal product is of a lower quality than the average product. But it may also be true even if the consumer engages in costly screening for product quality in equilibrium. We also find that aggregate welfare can be higher in an equilibrium where entry costs are higher, thus restricting the choice set of the consumers.

Similarly, with the advent of the internet as a medium for job search, there has been a substantial reduction in the cost of applying for jobs. There seems to be very little physical search involved, with many jobs offering online applications. It is easy to conceive that lower quality workers are more likely as a result to apply for positions in pure hope, just because it is now much easier to do so[1]. Another obvious context is

[1]This sentiment is echoed by people in the industry who point out that the online process has led them to "build a bigger pipeline than they need, which ends up drawing applications from far too many unqualified candidates" and "these companies are paying the price of having to sift through them all". For more see http://knowledge.wharton.upenn.edu/article/why-the-job-search-

that of applications for colleges and universities. There is evidence that applications for college in the U.S have been on the increase over the past decade. Applications to only one institution declined as a proportion of the total from 22% in 1999 to 14% in 2009. This has led to some skepticism as Fred Hargadon, former Dean of admissions at Princeton and Stanford, doubts that more and more applicants make for a stronger class. He says "I couldn't pick a better class out of 30,000 applicants than out of 15,000" [2]. In this paper we suggest reasons for fearing that large applicant pools may even be harming these institutions.

This paper contributes to two different literatures by highlighting a simple mechanism related to congestion externalities in the context of online markets. Brynjolfsson et al. (2003) and Waldfogel and Aguiar (2015) analyse the welfare gains to consumers of the increase in the availability of products online. Brynjolfsson et al. (2003) argue that with the advent of internet retailers who have an almost infinite "virtual inventory" provide increased benefits of product variety to consumers. For example, Amazon.com provides consumers with millions of choices for books which can be ordered online as opposed to a few thousand options in a brick and mortar store. Waldfogel and Aguiar (2015) highlight that Brynjolfsson et al. (2003) might be underestimating the benefits from the increased product variety if there is an ex-ante uncertainty about the quality of the incremental products available. Neither of these studies accounts for the fundamental phenomenon underlying our model: the marginal product may be of a lower quality than the average quality of the existing pool of products which would further reduce the average quality of the pool. Moreover, both these studies ignore the cost of evaluating product quality on the part of consumers. Hence, the broader implication is that not accounting for this phenomenon, there might be an overestimation of the benefits from greater product variety due to a proliferation of the internet.

The mechanism by which a decision maker can be worse off with more choice than less is more broadly applicable in the case of matching markets as well. Burdett et al. (2001) highlight the urn-ball matching friction in a model of directed search where there are multiple sellers and buyers but the latter can approach only one seller. Thus, in a large labour market, due to lack of co-ordination between the buyers there might be a situation where some sellers have more buyers than they

is-like-throwing-paper-airplanes-into-the-galaxy/

[2]This point of view is echoed by Karl M. Furstenberg, dean of admissions and financial aid at Dartmouth College from 1992 to 2007, who says "I don't think these larger applicant pools are materially improving the quality of their classes" . For more see "Application Inflation: When Is Enough Enough?": New York Times (November, 2010).

can service while some may not receive enough. Albrecht et al. (2004) analyze a model of directed search with wage posting where workers can make multiple applications to firms. The fact that a worker can send multiple applications reduces the chances of a firm's not getting any applications at all. On the other hand, the same worker applies to different firms which reduces the chance that any given offer will be accepted. This is another co-ordination friction in addition to the urn-ball matching friction. Halaburda et al. (2016) highlight similar frictions in a two-sided market context. To the best of our knowledge, our study is the first to highlight a simple mechanism such that a reduction in the cost of applications might make a decision maker worse off through a decline in the average quality of the pool of applicants. We also analyse and highlight a potential need to increase the cost of entry to market (cost of applying) for new products (new applicants) and how that would lead to an increase in welfare.

There are a few empirical papers which analyse online matching in the context of labour markets. Chandler et al. (2015) analyse, in a field experiment in an online labour market, the impact of making the application process more costly by making applicants answer additional questions on match quality and worker performance. In line with our results, they find that with the additional application costs the workers are positively selected in terms of past earnings, experience and better feedback. Kuhn and Skuterud (2004) using internet usage data in U.S suggest that the internet was ineffective in reducing unemployment durations.

In section 2, we analyse a simple model with one firm and two workers. In section 3, we allow for a general model with a large number of firms and workers. Section 4 looks at extensions to the general model while section 5 concludes.

2 The Model

We now set up a simple model to demonstrate the fundamental friction we want to highlight in the context of online markets. We embed the model in a labour market setting with two potential applicants (unemployed workers) and a firm which has to fill a vacancy. The cost of making an application varies depending on whether applications can be made online or not. The model is not particular to this setting and is applicable to online markets in general, as discussed above. The firm is the decision maker in this setting and the applicants are the products that the decision maker has to choose from.

2.1 The Setup

2.1.1 Workers and Firms

There are two risk neutral unemployed workers indexed by $i \in \{1, 2\}$. Each worker can be one of two types $\theta_i \in \{a, b\}$. Nature chooses the type of each worker according to the unconditional probabilities:

$$prob(\theta_i = a) = p_i \in (0, 1) \tag{1}$$

with[3]:

$$p_1 > p_2$$

There is one risk neutral firm which has one vacancy and may hire either of the workers. The value of the match to the firm from hiring a particular type of worker is given by:

$$v(\theta_i = a) = V \tag{2}$$

$$v(\theta_i = b) = X \tag{3}$$

where $V > X \geq 0$. Since $p_1 > p_2$, worker 1 always has a higher chance of being a good match for the firm.

The workers choose whether or not to apply to the firm. Each worker can make an application at a cost $t > 0$. We denote the action set for the worker i by $\{g_{i1}, g_{i0}\}$, where g_{i1} corresponds to worker i applying and g_{i0} corresponds to worker i not applying. The firm, upon receiving an application, needs to decide whether to incur a cost $c > 0$ per application to screen and identify the type θ_i of the worker (with

[3]A slightly different modelling approach would be to allow each worker i to be ex-ante symmetric and then draw an $i.i.d$ noisy signal $s_i \in \{High, Low\}$ about her true ability (a or b). We can, for example, have each signal be drawn with probability $\frac{1}{2}$ with $p(a|High) > p(a|Low)$. With such a formulation, our results will still hold as long as both workers draw different signals and can be ranked by the firm. This should be seen as an explanation of the results rather than a shortcoming of the model. With (ex-ante) homogeneous agents a firm is always weakly better off with more applications. A statistical decline in the quality of the pool of applicants is a necessary condition for our results to go through.

probability 1) or not. Let $s_l^1, l \in \{0,1\}$ be the screening action of the firm if it receives one application and $s_r^2, r \in \{0,1,2\}$ be the screening action of the firm if it receives two applications.

If the firm gets two applications, it cannot distinguish which application corresponds to which worker. The firm can decide to evaluate the applications in detail to find out θ_i or just randomly pick one out of the pile without incurring any cost[4]. When the firm screens for the type of the worker, the evaluation procedure is such that the employer first randomly picks out one application from the two. It evaluates the application for θ_i. If it does not turn out to be a type a worker then it will evaluate the next applicant. Thus, we allow for the firm to choose the number of applications it would want to screen.

2.1.2 The Information Structure

Neither the firm nor the worker observes the intrinsic type of the worker, but the probability distribution over types for each worker is common knowledge. However, there is an important source of informational asymmetry which develops between the firm and a worker who makes an application: workers observe their index but firms do not (though they may be able to infer it in equilibrium), and workers, unlike firms, therefore know whether their probability of being of type a is equal to p_1 or p_2[5].

Let $\lambda(f|n)$ denote the belief of the firm about the identity of the workers who have applied, with $f \in \{1, 2, (1,2)\}$ when it gets $n \in \{0,1,2\}$ applications.

2.1.3 Payoffs

Let $\pi(.)$ and $U(.)$ denote the payoff function for the firm and workers respectively. To simplify notation for the sections below, we define the expected wage for each

[4]In this case, the firm commits to hiring the worker it randomly picks out.

[5]This information structure ensures that we can generate an equilibrium in which both workers apply when the firm screens. An equivalent formulation, with results being qualitatively similar, could allow a worker to have private information about her intrinsic type with the firm having an imperfect screening technology.

application and screening strategy as $u^i\left(g_{ik}|g_{-ih}, s_j^n\right)$with a slight abuse of notation such that $i = 1(2)$ and $-i = 2(1)$.

After a worker is hired, production takes place, and the firm and the worker then bargain over the realized surplus. We use the Nash bargaining solution to capture the outcome, on the assumption that if bargaining breaks down the workers and the firm receive zero payoffs.[6][7]

Let γ and $1-\gamma$ be the bargaining power of the firm and the worker respectively. Both workers have the same bargaining power. In the Nash bargaining stage, the screening costs incurred in the previous period do not matter since they are sunk. The wage $w(g_1, g_2, s_j^n)$ is set so that the logarithm of the joint Nash product is maximized.

2.1.4 The Timing of the Game

The sequence of moves in this game is given by:

Period 0: Nature chooses the type of the worker.

Period 1: Workers choose whether to apply for the job.

Period 2: The firm decides (a) whether to screen and (b) whether to hire.

Period 3: If the firm decides to hire, then production takes place and there is Nash bargaining over the division of the surplus- the size of which is observed by the firm and the hired worker.[8].

We look for perfect Bayesian equilibria of the game in pure strategies. To characterize the equilibrium, we also need to define the beliefs of the firm off the equilibrium

[6]Different reservation values for the workers or introducing limited liability for them will allow the results to remain qualitatively the same.

[7]Nash Bargaining is a standard assumption in the labour literature. One can modify this to fit other online markets better by having a flat pre-determined price being paid by the decision maker for the product. This would make the analysis essentially the same.

[8]Thus, in this setup there is bargaining over the surplus ex-post. We avoid considering bargaining before the size of the surplus is known, which would add complications without increasing insight.

path. This matters particularly when, in equilibrium, the firm gets applications from both workers (or from neither) and hence has to specify beliefs for the out of equilibrium event of getting one application[9]. This out of equilibrium event will have a non-singleton information set with $\lambda^{*'}(.)$ being the belief that the firm holds that it is worker 1 who has applied when it gets only one application.

We now consider in detail the beliefs and actions of the players, solving backwards from the end. All proofs are in the Appendix .

2.2 Solving the Model

2.2.1 Period 3: Nash Bargaining Solution

In the spirit of backward induction, we start from period 3 which is the Nash bargaining stage.

Since the firm and the hired worker bargain over the realized surplus there is no asymmetric information at this stage. The wage w is set so that the logarithm of the joint Nash product is maximized: $max_w\ log\left[(v(\theta_i) - w - 0)^{\gamma}(w - 0)^{1-\gamma}\right]$, which gives a first order condition:

$$\frac{1-\gamma}{w} - \frac{\gamma}{v(\theta_i) - w} = 0 \tag{4}$$

The second order condition is given by $\frac{\gamma-1}{w^2} - \frac{\gamma}{(v(\theta_i)-w)^2} < 0$. The ex-post equilibrium wage is given by

$$w = (1-\gamma)v(\theta_i) \tag{5}$$

We have $v(\theta_i) = V$ or X depending on whether the firm screens or not.

[9]When the firm gets only one application in equilibrium, off the equilibrium path the information set is a singletion and hence we will have $\lambda^{*'}(f = (1,2)|n = 2) = 1$ in that case.

2.2.2 Period 2: Firm's Screening Strategy

The gross value of the match that the firm gets when it does not screen worker i , in expectation, is $p_iV - (1-p_i)X$. The payoff to the firm, taking into account the expected wage paid out is:

$$\pi(s_0^1|g_{i1}, g_{-i0}) = p_i\gamma V + (1-p_i)\gamma X \tag{6}$$

If the firm gets only one application then it will never have any incentive to screen because the expected value of hiring either worker is positive.

We define $p \equiv \frac{p_1+p_2}{2}$. The payoff to different strategies when the firm gets two applications is given by:

$$\pi(s_0^2|g_{11}, g_{21}) = p\gamma V + (1-p)\,\gamma X$$

$$\pi(s_1^2|g_{11}, g_{21}) = p\gamma V + \frac{1}{2}\sum_{i=1}^{2}(1-p_i)(p_i\gamma V + (1-p_i)\gamma X) - c$$

In the case of the event$\{s_1^2, g_{11}, g_{21}\}$, there is a $\frac{p_i}{2}$ chance of picking the application of worker i and finding her to be type a in the first draw out of the pile of applications. Thus, there is a $(\frac{1-p_i}{2})$ chance of not finding $\theta_i = a$ which provides the continuation probability. The firm will have no incentive to screen the second application if the first draw does not lead to a type a worker since the expected value from the second draw would always be positive and the firm will always hire.

Conditional on getting two applications, it is efficient for the firm to screen both for the worker type by incurring c per application if $\pi(s_1^2|g_{11}, g_{21}) \geq \pi(s_0^2|g_{11}, g_{21})$ which means:

$$c \leq c^+ \equiv \gamma(p - p_1p_2)(V - X) \tag{7}$$

Thus, if the screening costs $c \leq c^+$are not too high, then it will be optimal for the firm to screen both applications.

2.2.3 Period 1: Workers' Application Strategy

In the first stage, the workers have to choose to either apply or not taking into account the screening strategy of the firm as well the wage they would get at the Nash

bargaining stage.

The payoff to the worker i from applying (conditional on worker $-i$ not applying) is $U^i\left(g_{i1}|(g_{-i0}, s_0^1)\right) = (1-\gamma)(p_iV + (1-p_i)X) - t$. If both workers apply and if the firm screen applications for θ_i then the payoff is:

$$U^i\left(g_{i1}|(g_{-i1}, s_1^2)\right) = (1-\gamma)(1-\frac{p_{-i}}{2})V + \frac{1}{2}(1-\gamma)(1-p_i)X - t \tag{8}$$

If both workers apply and the firm does not screen, we have $U^1\left(g_{i1}|(g_{-i1}, s_0^2)\right) = \frac{1}{2}(1-\gamma)(p_iV - (1-p_i)X) - t$. The intuition behind these is based on the evaluation procedure as highlighted in the previous section.

We can define different thresholds on t to identify efficient application actions by the workers. If t is low enough (below a certain threshold) then both workers will have the incentive to apply. If t is high enough but not too high then worker 1 will be the only applicant. This happens, for a large range of parameter values, because for any screening action of the firm worker 1 gets a higher expected payoff relative to worker 2. For a range of parameter values, it is the case that the firm could get one application from only either worker 1 or 2 since it might be in the interest of any one worker to apply but not both simultaneously.

2.3 The Equilibria of the Model

We will now characterize the different pure strategy perfect Bayesian equilibrium outcomes of the model.

Proposition 1. The Perfect Bayesian Equilibrium outcomes can be classified as follows[10]:

(I) The firm gets two applications with both the workers applying if t is low enough, and it screens if c is low enough.

(II) The firm gets one application with only worker 1 applying if t is high enough.

[10]The exact thresholds on t and c are given in the Appendix.

(III) There can be multiple equilibria where firm receives one application either from worker 1 or worker 2 for (intermediate values of) $t \in [u^1(g_{11}|g_{21}, s_k^2), u^2(g_{21}|g_{10}, s_0^1)]$.

The proposition follows from the thresholds computed above. In equilibrium, for different parameter values, a firm may have different number of applicants ranging from no applicants to both workers applying[11]. In (III) there are multiple equilibria at the application stage in which either worker 1 applies or worker 2. It is never the case that, for these parameter values, both workers have the incentive to apply together.

In the next section, we establish a result which will capture the crux of the paper.

2.4 The Externality of the Marginal Applicant

The main idea we want to highlight is that the firm might be worse off getting two applications rather than one. We first provide a definition for what we mean by the marginal applicant:

Definition 2. Consider any equilibrium in which both workers apply, and increase t while holding all other parameters constant. The marginal applicant is the first to switch from applying to not applying.

We can identify the marginal applicant by looking at the workers' application strategies as we increase t.

We now provide a formal definition of the externality imposed on the firm by an applicant:

Definition 3. The externality imposed by an applicant $i \in \{1, 2\}$ is defined by $\mathcal{E}(i) \equiv \pi(s_j^{2*}|g_{11}^*, g_{21}^*) - \pi(s_0^1|g_{-i1}, g_{i0})$, which is the difference in the payoff $\pi(s_j^{2*}|g_{11}^*, g_{21}^*)$, received by the firm in an equilibrium where it receives two applications, with$j \in \{0, 1\}$, and the payoff $\pi(s_0^1|g_{-i1}, g_{i0})$ it would have got if it had received only one from applicant $-i$.

[11]The no application equilibrium is economically uninteresting. Details of the characterization of such equilibria are available from the authors upon request.

For any any equilibrium outcome in which the firm gets two applications, there is a range of parameter values such that, if t increases sufficiently to lead the firm to receive one application instead of two, it will receive a higher payoff. Thus the marginal applicant, denoted by h, can impose a negative externality, $\mathcal{E}(h)< 0$, on the firm. More formally:

Proposition 2.

(I) In an equilibrium where the firm does not screen, we have $\mathcal{E}(h) < 0$ if the marginal applicant is worker 2.

(II) In an equilibrium where the firm screens, we have $\mathcal{E}(h) < 0$ if the marginal applicant is worker 2 and $c \in [\gamma (p_2 - p_1 p_2)(V - X), c^{+}]$.

The method to show this is straightforward. First, we need to compare the difference in the payoff that the firm receives when it gets two applications in equilibrium and with the payoff it would have got had it received only one application, keeping the screening strategy fixed. This provides us with a condition on the parameter c. Using these conditions, we then need to verify whether the initial constraints on the firm's equilibrium strategies still hold ex-post.

In (I), the firm will be trivially worse off if worker 2 is the marginal applicant because of a decline in the average quality of the pool of applicants. Earlier, when the firm received one application, it would have received a share of the expected surplus generated by hiring worker 1. Now, it receives a share of the surplus by hiring either worker 1 or worker 2 where the overall size of the surplus, in expectation, declines when worker 2 is the marginal applicant. If the marginal applicant is worker 1 then the firm will be better off because the statistical increase in the quality of applicants. Figure 3.1 shows how the firm's payoff evolves with t when there can be multiple equilibria with the marginal application coming from either worker 1 or worker 2.

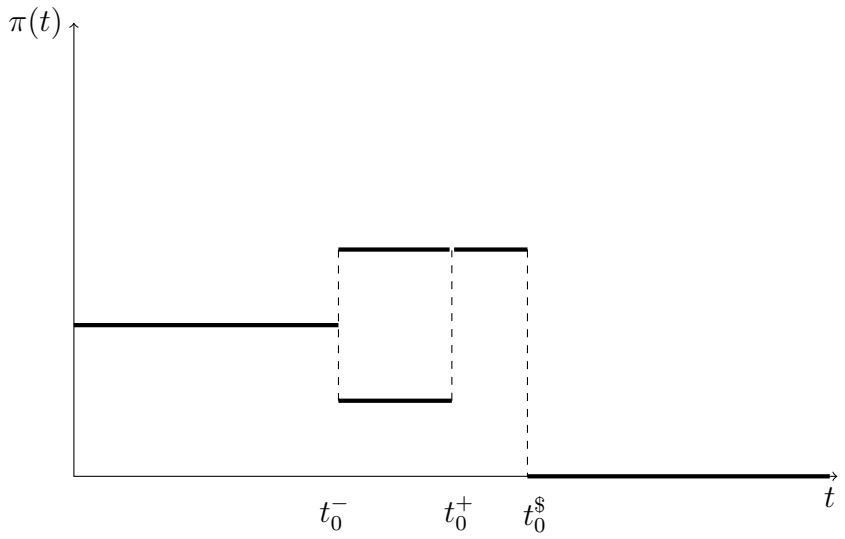

Figure 2.1: Firm's payoff evolution with multiple equilibria

If t is too high, that is $t \geq t_0^\$$, then no worker applies. If $t \in (t_0^-, t_0^+)$ then the firm gets one application from either worker 1 or worker 2. If $t \leq t_0^-$ there is a discontinuous rise or fall in the expected payoff of the firm depending on who is the marginal applicant.

A marginal application has two effects when the firm screens. Each additional application creates a positive option value for the firm. If the firm does not find a type a worker in the first draw, it has the option of looking at the second applicant. If the marginal applicant is a lower quality worker then there is a negative quality effect because of a statistical decline in average quality of the pool of applicants. Thus looking at (II), the firm will be worse off if the marginal applicant is worker 2 and if the quality effect outweighs the option value effect. If the firm screens applications and it gets the marginal application from worker 1 then it is always better off. The graphical intuition of this case is very similar to the figure above.

Corollary 1. If $p_1 = p_2$ (workers are homogeneous), then the firm can never have a negative externality from an additional application.

This follows straight from the result established above. If the workers are homogeneous then there is never a decline in the average quality of the pool of applicants with an additional application. Hence, if the firm does not screen then its expected

payoff remains the same from randomly picking an application. Moreover, if the firm screens then the positive option value of an additional candidate always dominates since there is no statistical decline in average quality.

2.5 Pricing and Welfare

In this section, we look to endogenize the cost of applying by allowing the firm to choose it, and we analyze its impact on aggregate welfare. We compare total welfare across equilibrium outcomes. As with the externality, we start with an equilibrium where both workers apply (and the firm does not control t) and compare the welfare to an equilibrium in which the firm prices away the marginal worker, hence getting only one application. In this case, we allow the firm to set a non-monetary cost e on top of t such that $t+e$, with $e \geq 0$, is the total cost of making an application. To keep things simple, we assume that a firm can impose this on the applicants without an additional cost to itself. Now, before the workers apply, the firm decides on e. Upon observing the total cost, workers decide to apply or not.[12] The aggregate welfare when the firm gets two applications (when t is low) is given by $\mathcal{W}^2$ as compared to when it gets only one (when t is set high),$\mathcal{W}^1$.

We focus on the cases where worker 2 is the marginal applicant such that the firm has incentive to price away the negative externality to get:

Proposition 3. For any pair of equilibrium outcomes, the firm can set $t + e^*$ to price away worker 2 which leads to $\mathcal{W}^1 \geq \mathcal{W}^2$ for a subset of the parameter space.

[12]This cost, in general, can be in the form of an application fee or can come from making the application a very long and time consuming process. A significantly longer and different application form is one of the main reasons why the University of Chicago, for a long time, got significantly less applications for its undergraduate class compared to other similarly ranked institutions. They had what was called the "Uncommon Application," in contrast to the Common Application, the standardized form that allows students to apply to any of hundreds of participating colleges (New York Times, November 2010). Its application forced applicants, apart from other things, to write long and creative essays on topics such as "If you could balance on a tightrope, over what landscape would you walk? (No net)". Such exercises were used to weed out candidates who would not be a good match for the institution.

In general, the point at which we switch from a single application to a two application equilibrium can reduce the aggregate welfare. Worker 2 has the same probability of being picked out by the firm as worker 1 but generates a surplus ≈ 0 as compared to the strictly positive surplus generated by worker 1. Even as t declines further, the welfare in a two application equilibrium might be outweighed by the welfare of the one application equilibrium. In particular, if the firm screens in the two application equilibrium, then this prevents it from incurring an additional c in expectation which can potentially increase welfare.

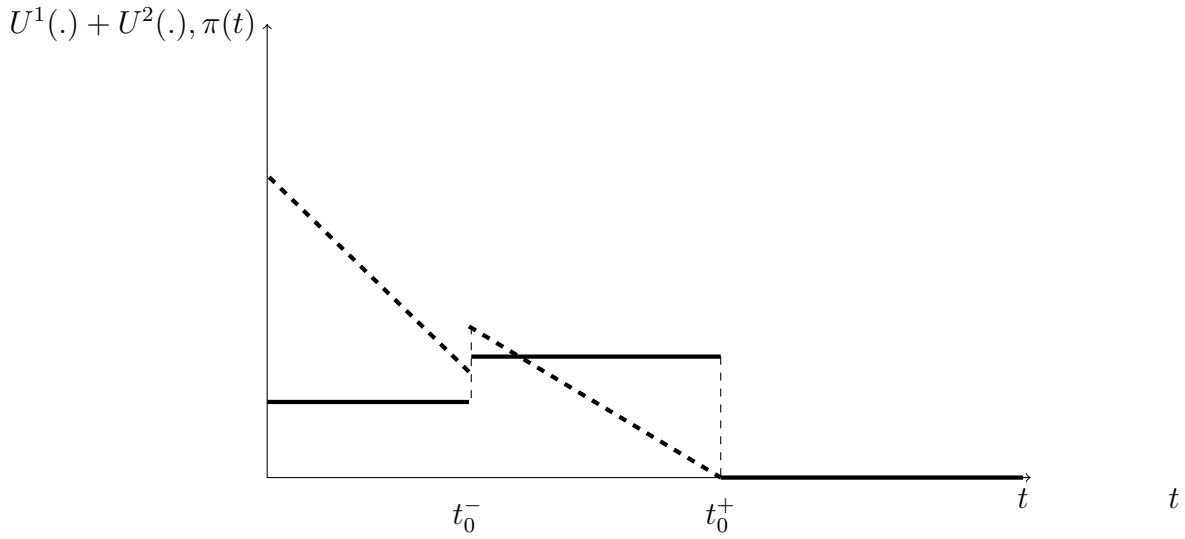

Figure 2.2: Firm's and workers' payoff evolution with t

Focussing on the pair of equilibrium outcomes where the firm does not screen, figure 3.2 shows how the payoff to both the workers (given by the thick dashed line) as well as the firm (given by the solid line) evolves with t . When t reaches below t_0^+, worker 1 is willing to apply. As t declines further, the payoff of worker 1 has a one-to-one increase since it is the residual claimant of any decline in the application costs. As $t \leq t_0^-$, then worker 2 also applies which first reduces the combined welfare of the workers but as t declines further leads to a gradual increase. One can see that for values of t not too low, the aggregate welfare when the firm receives one application can outweigh the aggregate welfare when the firm receives two.

There is a caveat which should be kept in mind. The outside option of each of the

workers is normalized to zero. If worker 1 has a high enough outside option relative to worker 2, then such a pricing policy could make worker 1 not apply and choose its outside option. A similar concern would be if e was a monetary cost imposed on the applicants. In the presence of budget constraints, such a pricing policy could result in counter productive outcomes for the firm.

3 Extention: Multiple Firms and Workers

We now sketch out an extension to the model with a large but finite number of heterogeneous workers and firms. We will highlight how the intuition of the externality in the previous section still holds, with the details being relegated to the appendix. As in the basic setup, each worker can potentially be a good match with any of the firms. We have a framework with discrete worker types with the intrinsic ability revealed if the firm screens applicants.

We denote workers by $i \in \{1, 2, ..., L\}$ and the firms by $m \in \{1, 2, ..., K\}$ with $L \geq K$. A worker i can be one of K possible types with $\theta_i \in \{1, 2, ...K\}$. As in the basic model, we allow each worker to have a positive probability to be a good match with any of the firms with:

$$p(\theta_i = m) = p_{im} \in (0, 1), \forall i, m$$

with $\sum_m p_{im} = 1, \forall i$.

The value of the match created by firm m hiring worker i is:

$$v(\theta_i = m) = V$$
$$v(\theta_i \neq m) = X$$

A worker, as before, can apply to a firm at a cost $t > 0$. We allow the workers to send an arbitrary number of applications with a maximum of one per firm. The cost of evaluating applications is again captured by $c > 0$.

If a worker recevies multiple offers, we allow firms to compete for a worker's services in the Bertrand sense by offering a higher expected wage. In this competition, what matters is the expected value of the surplus of its highest and second highest

offers. Consider firm $m(s)$ which provides the (second) highest value amongst the offers made to worker i. Since each firm knows the index of the worker being competed for, firm m will be able to hire by providing a payoff of $E(M_s^i) \equiv p_s V + (1 - p_s)X$ to the worker, with its own payoff being $E(M_m^i) - E(M_s^i)$.

We characterize the symmetric equilibrium of the model in which all workers apply to all firms where the firms choose the same screening strategy in equilibrium. The intuition behind the result relies on the application and screening cost thresholds as in Section 3.3.

We then establish the existence of a negative externality of the marginal applicant. Now, in addition to a statistical increase or decrease in quality of the pool of applicants, there is another effect because of the ex-post competition. As in the basic model, there is a 'pure' quality effect. A marginal applicant of greater (less) than average quality increases (decreases) the quality of the pool of applicants.

Now, however, there is a competition effect as well. If firm m does not rank 'high' or if the marginal applicant has outside options which will lead to a high expected payoff being provided as a result of the competition, then the firm might still be worse off even with higher than average applicant. Similarly, firm m might be better off with a lower ability marginal applicant if the decline in quality is outweighed by weak competition from other firms for this particular candidate.

As an illustration, consider an equilibrium outcome in which each firm gets L applications and no firm screens any application. For any such equilibrium there is a range of parameter values such that , if t increases sufficiently to lead firm m to receive $L - 1$ applications (with the marginal applicant switching from applying to not applying) instead of L, it will receive a higher payoff.

We define

$$\nabla_{L-1} \equiv \sum_{r=1}^{L-1} \frac{1}{L-1} \left(1 - \frac{1}{L}\right)^{k-1} \gamma E(M_m^r) - \sum_{i=1}^{L} [\frac{1}{L} \left(1 - \frac{1}{L}\right)^{k-1} \gamma E(M_m^i)]$$

which will capture what we will call the (pure) quality effect which measures the difference in the expected payoff for firm m when it hires an individual, not picked out by any other firm, randomly from a pool of L instead of a pool of $L-1$ applications.

Next, we define

$$\triangle_{L-1} \equiv \sum_{r=1}^{L-1}\sum_{j=1}^{k-1}[\frac{1}{L-1}\binom{k-1}{j}\frac{1}{L^j}\left(1-\frac{1}{L}\right)^{k-1-j}\mathbf{1}_{\{m\succ s\}_i}(E(M_m^r)-E(M_s^r))]$$

$$-\sum_{i=1}^{L}\sum_{j=1}^{k-1}[\frac{1}{L}\binom{k-1}{j}\frac{1}{L^j}\left(1-\frac{1}{L}\right)^{k-1-j}\mathbf{1}_{\{m\succ s\}_i}(E(M_m^i)-E(M_s^i))]$$

which will capture the competition effect which measures the difference in the expected payoff for firm m when it competes for a worker, picked out by atleast one other firm, randomly drawn from a pool of L instead of a pool of $L-1$ applications. In such a situation, we find that firm m will be worse off with L rather than $L-1$ applications if $\nabla_{L-1}+\triangle_{L-1}>0$.

4 Conclusion

In this paper we address the question of whether a potential employer can be worse off with more applications than less. We show that, contrary to conventional wisdom, this can happen in a variety of cases. The basic mechanism which drives the results is that, as application costs decline, if the additional applications are made by lower quality applicants then the firm can be worse off. We provide a framework with a flexible screening procedure in which a particular firm can be better or worse off with an additional application. We show that even when the firm has access to a screening technology and never hires a worker who is intrinsically 'bad', it can be worse off. If worker heterogeneity is high enough, then the additional application to the firm always comes from a worker who has a lower ex-ante quality which makes the possibility of a negative externality greater.

The only occasion when a firm gets a low quality worker as the only applicant is in a situation where there are multiple equilibria and either worker could have made the application to that firm. Thus, the occasions when a negative externality of an additional application can arise are far greater than when a positive externality occurs. We then look at a situation where the firm can price applications; such an increase could raise aggregate welfare and not only the firm's profits. We then extend the model to allow for a large number of firms and workers to show that the basic insights still go through. In this extended model, we also allow for competition for a

worker's service between the firms if she receives more than one offer.

This phenomenon has yet to be explored in the literature in spite of its evident pertinence in the context of the digital age. With communication costs declining rapidly, individuals are often swamped with much more information than they would want. This phenomenon has also been noticed in the context of firms and other institutions receiving many more applications than they would ideally want.

Bibliography

[1] 'Application Inflation: When is Enough Enough?', New York Times, 5th November 2010.

[2] Albrecht, J., P.A. Gautier, S. Tan and S. Vroman, (2004), Matching with multiple applications, Economic Letters, vol. 84(3), pp. 311-314.

[3] Albrecht, J., P.A. Gautier and S. Vroman, (2006), Equilibrium directed search with multiple applications, Review of Economic Studies, vol. 73(4), pp. 869- 891.

[4] Burdett, K., S. Shi, and R. Wright, (2001), Pricing and matching with frictions, Journal of Political Economy, vol. 109(5), pp. 1060-1085.

[5] Chandler, D., J. Hortoni, and R. Johari, (2015), Market Congestion and Application Costs, mimeo.

[6] Cho, I.,K., and D. Kreps (1987), Signaling Games and Stable Equilibria, Quarterly Journal of Economics, vol. 102, pp 179–221.

[7] Diamond, P (1982), Wage Determination and Efficiency in Search Equilibrium, Review of Economic Studies, pp 217-227.

[8] Eeckhout, J. and P. Kircher (2010), Sorting and decentralized price competition, Econometrica, vol. 78(2), pp. 539-574.

[9] Galenianos, M. and P. Kircher, (2009), Directed search with multiple job applications, Journal of Economic Theory, vol. 114(2), pp. 445-471.

[10] Montgomery, J D, (1991), Equilibrium Wage Dispersion and Interindustry Wage Differentials, The Quarterly Journal of Economics, vol. 106(1), pp. 163-79.

[11] Mortensen, D. and C. Pissarides (1994), Job Creation and Job Destruction in the Theory of Unemployment, Review of Economic Studies, vol (61), 397-415.

[12] Pissarides, C.A. (1985), Short-Run Dynamics of Unemployment, Vacancies, and Real wages, The American Economic Review, Vol. 75, Issue 4, 676-690.

[13] Shimer, R., (2005), The assignment of workers to jobs in an economy with coordination frictions, Journal of Political Economy, vol. 113(5), pp. 996-1025.

[14] Villena-Roldan, B. (2008), Aggregate implications of employer search and recruiting selection. mimeo.

[15] Wolthoff, R. (2011), Applications and interviews, mimeo university of Toronto.

Appendix I: Omitted Proofs

Proposition 1: We will establish the result by examining the different cases one at a time.
(I) We first define a threshold on the cost of applying t with $t_k^* \equiv u^2(g_{21}|g_{11}, s_k^2)$. If $t \leq t_k^*$, then it is in the interest of both workers to apply. If either worker deviates, it will get a payoff = 0. Conditional on getting two applications the equilibrium strategy for the firm is to screen for θ_i if $c \leq c^+$. The belief on the equilibrium path is $\lambda^*(f = (1,2)|n = 2) = 1$ since this information set is a singleton. Off the equilibrium path, there is a non singleton information set. If the worker deviates she receives zero, we can have $\lambda^{*'}(.) \in (0,1)$ which will sustain this equilibrium outcome for the given parameter values due to equilibrium domination .

(II) If $u^1(g_{11}|g_{21}, s_k^2) \leq u^2(g_{21}|g_{10}, s_j^1)$ with $u^2(g_{21}|g_{11}, s_k^2) \leq t \leq u^1(g_{11}|g_{21}, s_k^2)$ then worker 1 will apply since $U^1\left(g_{11}|(g_{10}, s_0^1)\right) \geq 0$. Worker 2 will still not apply since even if she was the only applicant since $U^2\left(g_{21}|(g_{11}, s_k^2)\right) \leq 0$. Using Bayes' rule along the equilibrium path implies that $\lambda^*(f = 1|n = 1) = 1$. The firm, conditional on getting one application, will not screen. Belief off the equilibrium path $\lambda^{*'}(k = (1,2)|n = 2) = 1$ is unique since the information set is a singleton.

(III) If $u^1(g_{11}|g_{21}, s_k^2) \leq t \leq u^2(g_{21}|g_{10}, s_j^1)$ then worker 1 would find it profitable to apply if worker 2 does not and vice-versa. Thus, there could be two equilibria at the application stage in pure strategies. One is in which worker 1 applies (since $t \leq u^2(g_{21}|g_{10}, s_j^1) \leq u^1(g_{11}|g_{20}, s_j^1)$) and worker 2 does not (since $u^1(g_{11}|g_{21}, s_k^2) \leq t$). The other one is in which worker 2 applies and worker 1 does not. The firm's screening strategy will be to screen in either case. The beliefs on and off the equilibrium path are specified in manner similar to what we saw above. QED

Proposition 2: We establish the result for each of the firm strategies.

(a) Consider the two application-equilibrium given by $\{(s_0^{2*}, s_0^{1*}), g_{11}^*, g_{21}^*\}$. There is a negative externality of the marginal applicant h if $\pi(s_0^{2*}|g_{11}^*, g_{21}^*) - \pi(s_0^1|g_{1e}, g_{2-e}) \leq 0$, which implies that:

$$p_i\gamma V - (1-p_i)\gamma X - p\gamma V + (1-p)\gamma X \geq 0 \tag{9}$$

This will hold if $e = 1$ which means $i = 1$. This holds by assumption since $p_1 > p_2$ and hence $p_1 > \frac{p_1+p_2}{2}$. This is the case for sure if $u^1(g_{11}|g_{21}, s_k^2) \geq u^2(g_{21}|g_{10}, s_j^1)$. If this inequality does not hold, then i could be either 1 or 2. If $i = 2$, then there would

be a positive externality since $p_2 < p$ with $\pi(s_0^{2*}|g_{11}^*, g_{21}^*) - \pi(s_0^1|g_{10}, g_{21}) \geq 0$.

(b) Next, consider the equilibrium given by $\{(s_2^{2*}, s_0^{1*}), g_{11}^*, g_{21}^*\}$. For the negative externality to exist, we need to show that $\pi(s_2^{2*}|g_{11}^*, g_{21}^*) - \pi(s_0^1|g_{11}, g_{20}) \leq 0$, which holds if the cost of screening is high enough:

$$c \geq \gamma(p_2 - p_1 p_2)(V - X)$$

This inequality will only hold if the marginal applicant is worker 2. QED

Proposition 3. We will proceed by looking at each possible pair of equilibrium outcomes with worker 2 being the marginal applicant:

First, consider $\{s_1^2, g_{11}, g_{21}\}$ and $\{s_0^1, g_{11}, g_{20}\}$. This implies that $\mathcal{W}^2 = V(p - p_1 p_2) + X(1 - p - \gamma(p + p_1 p_2)) - 2t - c$ while $\mathcal{W}^1 = p_1 V + (1 - p_1)X - t - e^*$. Pricing away worker 2 can lead to an increase in aggregate welfare if $\mathcal{W}^2 - \mathcal{W}^1 \leq 0$. For $\mathcal{W}^2 - \mathcal{W}^1 \leq 0$ we need $V(\frac{p_1 - p_2}{2} + p_1 p_2) + X(\frac{p_2 - p_1}{2} + \gamma(p + p_1 p_2) \geq e^* - t - c$. The left hand side of the inequality is positive even if $\gamma = 0$ and hence, for e^* small enough or t or c large enough we will have $\mathcal{W}^1 \geq \mathcal{W}^2$.

Next, consider the pair $\{s_0^2, g_{11}, g_{21}\}$ and $\{s_0^1, g_{11}, g_{20}\}$. The outcome $\{s_0^2, g_{11}, g_{21}\}$ leads to $\mathcal{W}^2 = \frac{1}{2}(p_1 V + (1 - p_1)X) + \frac{1}{2}(p_2 V + (1 - p_2)X) - 2t$ while the latter leads to $p_1 V + (1 - p_1)X - t - e^*$. Pricing away worker 2 can lead to an increase in aggregate welfare if $\mathcal{W}^2 - \mathcal{W}^1 \leq 0$. For $\mathcal{W}^2 \leq \mathcal{W}^1$ it must be the case that $\frac{1}{2}(p_1 - p_2)(V - X) \geq e^* - t$. This inequality holds for an e^* low enough even if $t = 0$. QED

Appendix II: Extension to Many Firms and Workers

Using the setup outlined in the body of the article, we first define the payoffs of the firms and the workers from different screening strategies. Define $E(M_m^i) \equiv p_{im}V + (1-p_{im})X$, which is the expected surplus generated when worker i matches with firm m when it does not screen[13].

Payoff to the Firms

We characterize the payoffs to the firms when the cost of applying is low enough such that each worker makes k applications.

The expected payoff to firm m if each firm gets l applications and they decide not to screen is: $\pi((0,l)_m|\mathbf{g_k},(\mathbf{0},\mathbf{l})_{-\mathbf{m}}) =$

$$\sum_{i=1}^{l}\{\frac{1}{l}\left(1-\frac{1}{l}\right)^{k-1}\gamma E(M_m^i)+$$

$$[\sum_{j=1}^{k-1}\frac{1}{l}\binom{k-1}{j}\frac{1}{l^j}\left(1-\frac{1}{l}\right)^{k-1-j}\mathbf{1}_{\{m\succ s\}_i}(E(M_m^i)-E(M_s^i))]\}$$

If no other firm picks out worker i, then firm m gets a share γ of the surplus created. This happens with probability $\frac{1}{l}\left(1-\frac{1}{l}\right)^{k-1}$. If worker i does get multiple offers, firm m will only be able to hire her if it provides the highest offer.

Now suppose each firm gets l applications and they decide to screen for θ_i. Thus, there can be $l-1!$ sequences in which applications are picked out which define the expected payoff for the firm: $\pi((l,l)_m|\mathbf{g_k},(\mathbf{l},\mathbf{l})_{-\mathbf{m}})=$

$$\sum_{i=1}^{K}\frac{1}{l-1!}\sum_{l-1!}\left\{\sum_{h=0}^{l-1}\frac{1}{l-h}(1-p_{jm})^h(\gamma p_{im}V-c)\right\}$$

[13]For the purpose of analytical simplicity we will normalize $X=0$ which will allow us to derive a clean expression for the payoffs when the firms screen.

In the first draw of applications, if worker i is picked out then the expected payoff is $\frac{1}{l}(\gamma p_{im} V - c)$. If the firm picks out worker j in the first draw, there is a probability $\frac{1}{l}(1 - p_{jm})$, that the firm does not find say worker j with $\theta_j = m$. The firms keeps evaluating applications at a cost c till it does not find a $\theta_i = m$. The above expression is with a slight abuse of notation because the product $(1 - p_{jm})^h$ as well as the worker picked out at any draw will depend on the exact identity of applicants picked out in previous draws.

Payoffs of the Workers

We now look at the payoffs that worker i would get from making k applications and the conditions under which it would be efficient to do so.

Suppose the firms were not screening for θ_i, then if all the other workers were applying to k firms, worker i's payoff would be: $U\left(g_k^i | \mathbf{g_k^{-i}}, (\mathbf{0}, \mathbf{l})_{\mathbf{N}}\right) =$

$$\frac{1}{l}\left(1 - \frac{1}{l}\right)^{k-1}(1-\gamma)\sum_{h=1}^{k} E(M_h^i) + \sum_{j=2}^{k}\binom{k}{j}\frac{1}{l^j}\left(1 - \frac{1}{l}\right)^{k-j} E(M_s^i) - kt$$

The payoffs follow the screening procedure. Next, we look at the payoff to the worker when the firm does screen for intrinsic ability. If all the other workers were applying to k firms, worker i's payoff would be:$U\left(g_k^i | \mathbf{g_k^{-i}}, (\mathbf{l}, \mathbf{l})_{\mathbf{N}}\right) =$

$$\sum_{m=1}^{k}\frac{1}{l-1!}\sum_{l-1!}\left(\sum_{h=0}^{l-1}\frac{(1-\gamma)}{l-h}(1-p_{jm})^h(p_{im}V)\right) - kt$$

This expression follows the evaluation procedure adopted by the firms. Worker i could, from an ex-ante perspective, be a good match for any of the k firms. The worker could be picked in the first draw or in the second draw and so on. This can happen for each of the firms that the worker has applied to which leads to the expression above.

The Equilibrium of the General Model

We focus on symmetric pure strategy perfect Bayesian equilibria where all workers apply to all the firms[14].

Proposition 4. The pure strategy equilibrium outcomes of the game are as follows:

(a) Each worker applies to all firms and the firms screen if $t \leq t_i^{++}, \forall i$ and $c \leq$min $\{c_m^+, \gamma p_{gm} V\}$ $\forall m$.

(b) Each worker applies to all firms and the firms do not screen if $t \leq \hat{t}_i, \forall i$ and $c \geq$min $\{\hat{c}_m, \gamma p_{gm} V\}$ $\forall m$.

(c) The rest of the parameter space admits no symmetric equilibria in which all workers make K applications.

The decisions to apply and screen depend on the different parameter values. If the cost of making an application is low enough then all workers apply to all firms, given their screening strategy. For the firm to screen all applications in expectation, it must be that the cost of screening is low enough ($c \leq \gamma p_{gm} V$). We can establish that $\hat{c}_m \geq c_m^+, \forall m$ because of the weak submodularity of the firms' payoffs in their screening actions. If $\hat{c}_m > c > c_m^+$ then there would not be any symmetric equilibrium since it violates the conditions for the firms to adopt symmetric screening strategies while asymmetric equilibria can be constructed[15].

The Externality in the General Model

We consider a symmetric equilibrium in which each firm m gets L applications. We define the marginal applicant in exactly the same way as in the basic model. We re-state the definition of what we mean by the externality imposed by an applicant:

[14] It is a reasonable conjecture that one can characterize the equilibria in the same way when each worker sends an arbitrary number of applications since the trade-offs involved are identical.

[15] Details of the computation on the various thresholds is given below in the proofs.

Definition 4. The externality imposed by an applicant i is defined by

$$\mathcal{E}(i) \equiv \pi((q,L)_m|\mathbf{g}^*_{\mathbf{k}}, \mathbf{s}^*_{-\mathbf{m}}(\mathbf{L})) - \pi((h,L-1)_m|(g_{i,k-1}, \mathbf{g}_{-\mathbf{i},\mathbf{k}}), \mathbf{s}_{-\mathbf{m}}(\mathbf{L}))$$

, which is the difference in the payoff $\pi((q,L)_m|\mathbf{g}^*_{\mathbf{k}}, \mathbf{s}^*_{-\mathbf{m}}(\mathbf{L}))$, received by the firm in the equilibrium $\{((q,L)^*_m, (h,L-1)^*_m), \mathbf{g}^*_{\mathbf{k}}, \mathbf{s}^*_{-\mathbf{m}}(\mathbf{L})\}$, with $h = 0(L-1)$ if $q = 0(L)$and the payoff $\pi((h,L-1)_m|(g_{i,k-1}, \mathbf{g}_{-\mathbf{i},\mathbf{k}}), \mathbf{s}_{-\mathbf{m}}(\mathbf{L})$ it would have got if it had received only $L-1$ applications[16]. We denote the marginal applicant by y and $\frac{l}{2}$ the index of the average applicant from the existing pool based on firm m's initial ranking.

Proposition 5.

(I) In the equilibrium where the firms do not screen , $\mathcal{E}(y) < 0$ if $\nabla_{L-1} + \triangle_{L-1} > 0$.

(II) In the equilibrium where the firms screen , $\mathcal{E}(y) < 0$ if $h \geq \frac{l}{2}$ and if $c \in [\frac{1}{\delta}BV, c^+_m]$ otherwise $\mathcal{E}(y) > 0$.

The method to prove this proposition is similar to what we saw in the benchmark model. The intuition behind (I) has been discussed in Section 3.4.

The trade-off in (II) is the same as in the basic model. If in equilibrium all the firms screen, firm m can still be worse off with the additional application if the option value provided by the marginal applicant is outweighed by the statistical decline in average quality.

Proofs

Proposition 4. We will characterize the equilibrium looking at one case at a time.

(a) Let us first look at the response of the firms. Given that all workers apply to all firms, firm m will screen all the L applications if $\pi((L,L)_m|\mathbf{g_K}, (\mathbf{L},\mathbf{L})_{-\mathbf{m}}) \geq \pi((0,L)_m|\mathbf{g_K}, (\mathbf{L},\mathbf{L})_{-\mathbf{m}})$ or if the cost of screening $c \leq c^+_m$ otherwise will screen none. Taking this into account, workers apply to all K firms if $t \leq t^{++}_i, \forall i \left(U\left(g^i_K|\mathbf{g^{-i}_K}, (\mathbf{L},\mathbf{L})_{\mathbf{N}}\right) \geq U\left(g^i_{K-1}|\mathbf{g^{-i}_K}, (\mathbf{L},\mathbf{L})_{\mathbf{N}}\right), \forall i\right)$. The Kth application would bring the lowest ex-ante payoff and hence would provide the tightest restriction on the cost of making an application. On the equilibrium path, we have $\lambda^{m*}(f = (1,2,...,L)|n = L) = 1$ for each

[16]Note that $\{((q,L)^*_m, (h,L-1)^*_m), \mathbf{g}^*_{\mathbf{k}}, \mathbf{s}^*_{-\mathbf{m}}(\mathbf{L})\}$ does not completely define the equilibrium since we would need to define how many applications each firm would screen for any number that it might receive.

of the m firms. As in a previous proof, there is equilibrium domination and hence this equilibrium can be sustained by any beliefs $\lambda^{m'*}(.) \in (0,1)$ off the equilibrium path.

(b) In a similar way, we can see that if $\pi((0,L)_m|\mathbf{g_K},(\mathbf{0},\mathbf{L})_{-\mathbf{m}}) \leq \pi((0,L)_m|\mathbf{g_K},(\mathbf{L},\mathbf{L})_{-\mathbf{m}})$ or equivalently if $c \geq \hat{c}_m, \forall m$, then no firm will have an incentive to screen. Additionally, if $t \leq \hat{t}_i, \forall i$ (that is when $U\left(g^i_K|\mathbf{g^{-i}_K},(\mathbf{0},\mathbf{L})_{\mathbf{N}}\right) \geq U\left(g^i_{K-1}|\mathbf{g^{-i}_K},(\mathbf{0},\mathbf{L})_{\mathbf{N}}\right)$), then in equilibrium all workers make K applications since no firm screens any of the applications. Since $t \leq \hat{t}_i, \forall i$, it provides the tightest restriction on the cost of applying. The beliefs can be characterized as in the previous case.

The case where workers apply to $K-k$ firms could be solved in a similar manner since the fundamental trade-offs involved are very similar even though each firm might end up receiving a different number of applications since the initial ranking of firms by workers can be different. The equilibrium in which no (or some) firm gets any applications can be characterized as in the benchmark model. We can isolate unique out of equilibrium beliefs to deal with issues of multiplicity of equilibria. QED

Proposition 5. We will prove this proposition by looking at firm m . We look at the different cases as follows:

(a) Let us look at the equilibrium outcome $\{((0,L)^*_m,(0,L-1)^*_m),\mathbf{g^*_k},(\mathbf{0},\mathbf{L})^*_{-\mathbf{m}}\}$. For there to be a negative externality from the marginal applicant y, it must be the case that:

$$\pi(0,L)_m|\mathbf{g^*_k},(\mathbf{0},\mathbf{L})^*_{-\mathbf{m}}) - \pi((0,L-1)_m|(g_{i,k-1},\mathbf{g_{-i,k}}),(\mathbf{0},\mathbf{L})_{-\mathbf{m}}) \leq 0$$

If the firm gets only $L-1$ applications, its payoff is $\sum_{r=1}^{L-1}\{\frac{1}{L-1}\left(1-\frac{1}{L}\right)^{k-1}\gamma E(M^r_m)+$ $\sum_{j=2}^{k}[\frac{1}{L-1}\binom{k-1}{j}\frac{1}{L^j}\left(1-\frac{1}{L}\right)^{k-1-j}\mathbf{1}_{\{m\succ s\}_r}(E(M^r_m)-(E(M^r_s)]\}$. An additional application reduces the probability of the initial $L-1$ being the only applicants to be picked. We know that

$$\nabla_{L-1} \equiv \sum_{r=1}^{L-1}\frac{1}{L-1}\left(1-\frac{1}{L}\right)^{k-1}\gamma E(M^r_m) - \sum_{i=1}^{L}[\frac{1}{L}\left(1-\frac{1}{L}\right)^{k-1}E(M^i_m)]$$

. $\nabla_{L-1} > 0 (< 0)$ if the marginal applicant y is below (above) the average quality of the pool of applicants. We can have $\triangle_{L-1} \lesseqgtr 0$ depending on the competition from the other firms for the candidate picked out. To see this, re-write $\triangle_{L-1}$ as:

$$\triangle_{L-1} \equiv \sum_{r=1}^{L-1}\sum_{j=1}^{k-1}[\frac{1}{L-1}\binom{k-1}{j}\frac{1}{L^j}\left(1-\frac{1}{L}\right)^{k-1-j}\mathbf{1}_{\{m\succ s\}_i}(E(M_m^r)-E(M_s^r))]$$

$$-\sum_{i=1}^{L-1}\sum_{j=1}^{k-1}[\frac{1}{L}\binom{k-1}{j}\frac{1}{L^j}\left(1-\frac{1}{L}\right)^{k-1-j}\mathbf{1}_{\{m\succ s\}_i}(E(M_m^i)-E(M_s^i))]$$

$$-\sum_{j=2}^{k-1}[\frac{1}{L}\binom{k-1}{j}\frac{1}{L^j}\left(1-\frac{1}{L}\right)^{k-1-j}\mathbf{1}_{\{m\succ s\}_y}(E(M_m^i)-E(M_s^y))]$$

We can see that the sum of the first two terms is always positive. Thus, whether $\triangle_{L-1}$ is positive or negative depends on the magnitude of the first two terms and how it compares with the third. For $\mathcal{E}(y)<0$, it must be that $\nabla_{L-1}+\triangle_{L-1}>0$.

(b) We next look at the equilibrium outcome $\{((L,L)_m^*,(L{-}1,L{-}1)_m^*),\mathbf{g_k^*},(\mathbf{L},\mathbf{L})_{-\mathbf{m}}^*\}$. For a negative externality from the marginal applicant y, it must be the case that:

$$\pi((L,L)_m|\mathbf{g_k^*},(\mathbf{L},\mathbf{L})_{-\mathbf{m}}^*)-\pi((L-1,L-1)_m|(g_{i,k-1},\mathbf{g_{-i,k}}),(\mathbf{L},\mathbf{L})_{-\mathbf{m}})\leq 0$$

This inequality can be written as
$\sum_{i=1}^{L}\sum_{L-1!}\frac{1}{L-1!}\left\{\sum_{h=0}^{L-1}\frac{1}{L-h}(1-p_{jm})^h(\gamma p_{im}V-c)\right\}\leq$
$\sum_{r=1}^{L/y}\sum_{L-2!}\frac{1}{L-2!}\left\{\sum_{h=0}^{L-1}\frac{1}{L-h}(1-p_{jm})^h(\gamma p_{rm}V-c)\right\}$
which gives the condition $c\geq\frac{1}{\delta}BV$ where $\delta\equiv\sum_{L-1!}\frac{L}{L-1!}\left(\sum_{h=0}^{L-1}\frac{1}{L-h}(1-p_{jm})^h\right)$
$-\sum_{L-2!}\frac{L-1}{L-1!}\left(\sum_{h=0}^{L-2}\frac{1}{L-h}(1-p_{km})\right)$ is the coefficent on c. If the inequality holds, the additional cost imposed on the firm from screening outweighs the option value provided by the marginal applicant. Also, it must be that $c\in[\frac{1}{\delta}BV,c_m^+]\neq\{\phi\},\forall m$. The above inequality holds if the rank of $y\geq\frac{L}{2}$ for firm m, i.e the marginal applicant is above the average quality otherwise $\mathcal{E}(y)>0$. QED