Clicks and Editorial Decisions: Does Popularity Shape Coverage?

Ananya Sen and Pinar Yildirim

Toulouse School of Economics; Wharton School, U. Pennsylvania

October 6, 2014
Introduction

What drives the decision of editors to cover one story vs. another?
Introduction

- What drives the decision of editors to cover one story vs. another?
- Supply side (Ansolabehere et al. (2006), Fridkin et al. (2002), Larcinese et al. (2011)) vs Demand side (Gentzkow and Shapiro (2010)).
Introduction

- What drives the decision of editors to cover one story vs. another?
- Supply side (Ansolabehere et al. (2006), Fridkin et al. (2002), Larcinese et al. (2011)) vs Demand side (Gentzkow and Shapiro (2010)).
- "Digital Drive": Aggregate circulation rates to real time URL level info.
Introduction

- What drives the decision of editors to cover one story vs. another?
- Supply side (Ansolabehere et al. (2006), Fridkin et al. (2002), Larcinese et al. (2011)) vs Demand side (Gentzkow and Shapiro (2010)).
- "Digital Drive": Aggregate circulation rates to real time URL level info.
- Debate on how to utilize real time clicks: Eg. The Verge, Vox.com.
This Study I

- To what extent, if at all, does coverage of stories (duration, frequency of articles) respond to the clicks received?
This Study I

- To what extent, if at all, does coverage of stories (duration, frequency of articles) respond to the clicks received?

- Lack of disaggregated data:
  - We use a unique dataset at the level of a URL from a big Indian national daily.
This Study I

- To what extent, if at all, does coverage of stories (duration, frequency of articles) respond to the clicks received?

- Lack of disaggregated data:
  - We use a unique dataset at the level of a URL from a big Indian national daily.

- Need to define a ‘story’:
  - Use text analysis to link articles
To what extent, if at all, does coverage of stories (duration, frequency of articles) respond to the clicks received?

Lack of disaggregated data:
- We use a unique dataset at the level of a URL from a big Indian national daily.

Need to define a ‘story’:
- Use text analysis to link articles

Endogeneity of clicks due to unobserved heterogeneity and reverse causality:
This Study I

- To what extent, if at all, does coverage of stories (duration, frequency of articles) respond to the clicks received?

- Lack of disaggregated data:
  - We use a unique dataset at the level of a URL from a big Indian national daily.

- Need to define a ‘story’:
  - Use text analysis to link articles

- Endogeneity of clicks due to unobserved heterogeneity and reverse causality:
  - Rainy days
  - Electricity shortages
Can clicks based coverage hurt readers? the newspaper?
Can clicks based coverage hurt readers? the newspaper?

Page views are noisy and don’t always signal the newsworthiness of a topic.

Coverage could often be driven by events like rain and power outages.
This Study II

- Can clicks based coverage hurt readers? the newspaper?
- Page views are noisy and don’t always signal the newsworthiness of a topic.
- Coverage could often be driven by events like rain and power outages.
- Could be detrimental to information provision and newspaper’s profits.
- Simulate counterfactuals to quantify potential crowding out.
Overview of the Results

- Stories first published on rainy days receive a larger number of clicks.
- Power outages are negatively correlated with clicks.
- One standard deviation increase in the views of a story increases its duration by 1.25-3 days with 1.5-3 additional articles.
Overview of the Results

- Stories first published on rainy days receive a larger number of clicks.
- Power outages are negatively correlated with clicks.
- One standard deviation increase in the views of a story increases its duration by 1.25-3 days with 1.5-3 additional articles.

- Two counterfactual situations to quantify the potential crowding out or in of new stories.
  - No rain: Upto 928 (≈ 1%) new articles crowded out.
  - Only low power outages: Upto 660 (≈ 0.7%) new articles written.
A single newspaper decides how much coverage $c_i$ to give story $i$.

The newspaper cares about its readership $R(c_i)$.

It has a disutility $\lambda_i \in R_+$ associated with story $i$.

The payoff to the newspaper by giving coverage $c_i$ to story $i$:

$$ \pi(c_i) = R(c_i) - \lambda_i c_i $$
A Stylized Model II
The Readers

- There is a fixed set of potential readers of unit mass.
- An individual reader \( q \) has the following utility from reading story \( i \):

\[
U^q(i) = f(c_i, \alpha_i) - \delta_{iq}
\]

- \( \alpha_i \) is the appeal of/preference for story \( i \).
- The function \( f(\cdot) \) is increasing in \( c_i, \alpha_i, \frac{\partial^2 f(\cdot)}{\partial c_i \partial \alpha_i} > 0 \) and \( \delta_{iq} \sim U[0, 1] \).
A Structural Model

- The newspaper’s FOC ($f(.)=\sigma(\alpha_i c_i)^{\frac{1}{\sigma}}$):

$$c_i = \alpha_i^{1-\sigma} \lambda_i^{\sigma}$$
A Structural Model

The newspaper’s FOC \( f(\cdot) = \sigma(\alpha_i c_i)^{\frac{1}{\sigma}} \):

\[
c_i = \alpha_i^{\frac{1}{\sigma-1}} \lambda_i^{\frac{\sigma}{1-\sigma}}
\]

Taking the natural logarithm, we get a log-log specification:

\[
\log(c_i) = \frac{1}{\sigma - 1} \log(\alpha_i) + \frac{\sigma}{1 - \sigma} \log(\lambda_i)
\]
A Structural Model

- The newspaper’s FOC \( f(\cdot) = \sigma(\alpha_i c_i)^{\frac{1}{\sigma}} \): 
  \[
  c_i = \alpha_i^{\frac{1}{\sigma-1}} \lambda_i^{\frac{\sigma}{1-\sigma}}
  \]

- Taking the natural logarithm, we get a log-log specification: 
  \[
  \log(c_i) = \frac{1}{\sigma - 1} \log(\alpha_i) + \frac{\sigma}{1 - \sigma} \log(\lambda_i)
  \]

- Functional form assumptions and a bit of algebra gives: 
  \[
  \log(c_i) = \gamma_0 + \gamma_1 \log(\text{views}_i) + X_i'\gamma_2 + \epsilon_i
  \]
Data Description

- Data for the online edition of an Indian national daily for 2012.
- Data on all articles read during this period which includes:
  - The number of page views
  - The number of unique page views
Data Description

- Data for the online edition of an Indian national daily for 2012.
- Data on all articles read during this period which includes:
  - The number of page views
  - The number of unique page views

- Used a web crawler to combine it with publicly available information on:
  - The text of the article and the time it was first published.
  - The source of the story, whether it had an image, headline, tags.
A News Story

- A ‘news-story’ is defined as a cluster of articles based on a common underlying issue or topic.
- We use a word frequency algorithm to identify the similarity between articles.
- We follow Franceschelli (2011) by dividing the 365 days into 24-hour news cycles and assign each article to exactly one story.
News Story Example: Fukushima

<table>
<thead>
<tr>
<th>Headline</th>
<th>Time Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japans regains nuclear power after reactor restarts</td>
<td>5&lt;sup&gt;th&lt;/sup&gt; July, 2012 at 1:08 pm</td>
</tr>
<tr>
<td>Fukushima was ‘man-made’ disaster: Japanese probe</td>
<td>5&lt;sup&gt;th&lt;/sup&gt; July, 2012 at 6:39 pm</td>
</tr>
<tr>
<td>Comission calls Fukushima n-crisis man-made disaster</td>
<td>6&lt;sup&gt;th&lt;/sup&gt; July, 2012 at 1:37 am</td>
</tr>
<tr>
<td>‘Man-made’</td>
<td>7&lt;sup&gt;th&lt;/sup&gt; July, 2012 at 12:33 am</td>
</tr>
<tr>
<td>Fukushima lessons</td>
<td>7&lt;sup&gt;th&lt;/sup&gt; July, 2012 at 12:55 am</td>
</tr>
</tbody>
</table>

- The cluster consisted of five articles with an article every 24 hours related to the Fukushima incident.
Identification and Estimation

- Reverse causality: Greater coverage leads to greater reader interest.
  - Solution: Use the characteristics of only the first article of every story.
Identification and Estimation

- Reverse causality: Greater coverage leads to greater reader interest.
  - Solution: Use the characteristics of only the first article of every story.

- Measurement Error, Unobserved Heterogeneity in Views:
  - Rainfall: Takes the value 1 if it rained on a particular day in either Delhi or Mumbai.
  - Electricity Shortages: Use a daily measure which is the total power outages in Delhi and Maharashtra.
## IV Estimation: First Stage

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) log(views)</th>
<th>(2) log(views)</th>
<th>(3) log(views)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rain</td>
<td>0.0586***</td>
<td>0.0536***</td>
<td>0.0976***</td>
</tr>
<tr>
<td></td>
<td>(0.0141)</td>
<td>(0.0137)</td>
<td>(0.0190)</td>
</tr>
<tr>
<td>log(outage)</td>
<td>-0.0172***</td>
<td>-0.0470***</td>
<td>-0.0237**</td>
</tr>
<tr>
<td></td>
<td>(0.00646)</td>
<td>(0.00631)</td>
<td>(0.00989)</td>
</tr>
<tr>
<td>Section f.e.</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Month f.e</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>F- Statistic</td>
<td>15.94</td>
<td>53.88</td>
<td>14.80</td>
</tr>
<tr>
<td>Observations</td>
<td>60,671</td>
<td>60,671</td>
<td>60,671</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.167</td>
<td>0.224</td>
<td>0.230</td>
</tr>
</tbody>
</table>
Falsification tests indicate that the newspaper is unaware of these shocks to reader attention.

No difference in the words per article, probability of sourcing from an agency or number of articles published.

There is a difference on weekends implying a different editorial policy.
## IV Estimation: Length of the Story

| VARIABLES | Ols  
| ln(length) | 2sls  
| ln(length) | 2sls  
| ln(length) | 2st. Tobit  
| ln(length) |
|---|---|---|---|---|
| log(views) | 0.300*** (0.0183) | 3.017*** (0.920) | 1.865** (0.893) | 4.15*** (1.253) |
| $\sigma = \frac{1+\gamma_1}{\gamma_1}$ | 1.33 | 1.5 | 1.25 |
| Section f.e. | N | N | Y | N |
| Month f.e | N | N | Y | N |
| Over-id ($p$ value) | 0.99 | 0.14 | - |
| Observations | 60,671 | 60,671 | 60,671 | 60,671 |
### IV Estimation: Number of Articles

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>Ols</th>
<th>2sls</th>
<th>2sls</th>
<th>2st. Tobit</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(views)</td>
<td>0.022*** (0.0013)</td>
<td>0.311*** (0.075)</td>
<td>0.211*** (0.070)</td>
<td>0.460*** (0.110)</td>
</tr>
<tr>
<td>(\sigma = \frac{1+\gamma_1}{\gamma_1})</td>
<td>4.33</td>
<td>5.76</td>
<td>3.17</td>
<td></td>
</tr>
<tr>
<td>Month f.e.</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Section controls</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Over-id ((p \text{ value}))</td>
<td>0.58</td>
<td>0.66</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>60,671</td>
<td>60,671</td>
<td>60,671</td>
<td>60,671</td>
</tr>
</tbody>
</table>
Crowding Out and In of Articles I

- Simulate how many articles an average story would have received if:

1. There was no rain.
2. There were only low power outages.

- Change the number of views an average story receives but have the same characteristics.
### Crowding Out and In of Articles II

<table>
<thead>
<tr>
<th>Event</th>
<th>%Δ in Coverage</th>
<th>Δ in Coverage</th>
<th>%All Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rain</td>
<td>-3.5%</td>
<td>-928</td>
<td>1%</td>
</tr>
<tr>
<td>Power Outage</td>
<td>3%</td>
<td>660</td>
<td>0.67%</td>
</tr>
</tbody>
</table>
Robustness Checks

- Power outage as a proportion of daily consumption.
- Excluding outliers.
- Daily rainfall normalized by monthly mean.
- Unique views.
- Duration models.
Contribution and Next steps

- First to quantify the impact of reader preferences (e.g. clicks) on online editorial coverage decisions.
- Related to the literature on media bias (Mullainathan and Shleifer (2005), Gentzkow and Shapiro (2006, 2010)).

- Next steps: Impact of clicks on distribution of story types?
Contribution and Next steps

- First to quantify the impact of reader preferences (e.g. clicks) on online editorial coverage decisions.
- Related to the literature on media bias (Mullainathan and Shleifer (2005), Gentzkow and Shapiro (2006, 2010)).
- First evidence to identify the possibility that focusing on page views may be detrimental to information provision and firm’s profits.
- Implications for firm strategy as well as media policy (FCC diversity, PCI code of ethics).
- Next steps: Impact of clicks on distribution of story types?