# 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

# 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:

# 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

# This Study II

- Can clicks based coverage hurt readers? the newspaper?

# 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.

# 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 ($\approx 1\%$) new articles crowded out.
  - Only low power outages: Upto 660 ($\approx 0.7\%$) new articles written.

# A Stylized Model I

## The Newspaper

- 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 $f(.)$ is increasing in $c_i$, $\alpha_i$, $\frac{\partial^2 f(.)}{\partial c_i \partial \alpha_i} > 0$ and $\delta_{iq} \sim U[0,1]$.

# A Structural Model

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

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

# A Structural Model

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

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

- 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 ($f(.)=\sigma(\alpha_i c_i)^{\frac{1}{\sigma}}$):

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

- 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:

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

# 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

| Headline | Time Published |
|---|---|
| Japans regains nuclear power after reactor restarts | $5^{th}$ July, 2012 at 1:08 pm |
| Fukushima was 'man-made' disaster: Japanese probe | $5^{th}$ July, 2012 at 6:39 pm |
| Comission calls Fukushima n-crisis man-made disaster | $6^{th}$ July, 2012 at 1:37 am |
| 'Man-made' | $7^{th}$ July, 2012 at 12:33 am |
| Fukushima lessons | $7^{th}$ July, 2012 at 12:55 am |

- 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

| VARIABLES | (1)<br>log(views) | (2)<br>log(views) | (3)<br>log(views) |
|---|---|---|---|
| Rain | 0.0586*** | 0.0536*** | 0.0976*** |
| | (0.0141) | (0.0137) | (0.0190) |
| log(outage) | -0.0172*** | -0.0470*** | -0.0237** |
| | (0.00646) | (0.00631) | (0.00989) |
| Section f.e. | N | Y | Y |
| Month f.e | N | N | Y |
| F- Statistic | 15.94 | 53.88 | 14.80 |
| Observations | 60,671 | 60,671 | 60,671 |
| R-squared | 0.167 | 0.224 | 0.230 |

# Placebo Checks

- 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<br>ln(length) | 2sls<br>ln(length) | 2sls<br>ln(length) | 2st. Tobit<br>ln(length) |
|---|---|---|---|---|
| log(views) | 0.300*** | 3.017*** | 1.865** | 4.15*** |
| | (0.0183) | (0.920) | (0.893) | (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

| VARIABLES | Ols ln(articles) | 2sls ln(articles) | 2sls ln(articles) | 2st. Tobit ln(articles) |
|---|---|---|---|---|
| log(views) | 0.022*** | 0.311*** | 0.211*** | 0.460*** |
| | (0.0013) | (0.075) | (0.070) | (0.110) |
| $\sigma = \frac{1+\gamma_1}{\gamma_1}$ | | 4.33 | 5.76 | 3.17 |
| Month f.e. | N | N | Y | N |
| Section controls | N | N | Y | N |
| Over-id ($p$ value) | | 0.58 | 0.66 | - |
| Observations | 60,671 | 60,671 | 60,671 | 60,671 |

# Crowding Out and In of Articles I

- Simulate how many articles an average story would have recieved 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

| | %$\Delta$ in Coverage | $\Delta$ in Coverage | %All Articles |
|---|---|---|---|
| Rain | -3.5% | -928 | 1% |
| Power Outage | 3% | 660 | 0.67% |

# 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)).

## 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?