Crowding out the truth? A simple model of misinformation, polarization and meaningful social interactions

Fabrizio Germano U Pompeu Fabra

U Pompeu Fabra

Vicenç Gómez Francesco Sobbrio Tor Vergata U. Rome

Computational Politics e-symposium

#### Ranking matters for users!

9 Search ess Caused By Flu. Videos Next is tacking at whether or not the fla vacaine in at for the first time will receive it in two separate shots a Change location the flux shat to become effective Pages from Cana to the shafts importance, prompting many to Any time Past hour much the needle every year, but a new at does not work all the time. Its about 70 to 90 per cent effective which. The successor rate is somewhat leave in thicken and Post year More search tools our hands, cover your mouth. It's cald, the tree - DL Ordre WANKE: Boo to the Bu - Evening Neve (Flu Shot) Fects and Side Effects by MedicineNet ... Infuenza veccine - Wikipedia, the free encyclopedia Jump to Directiveness of veccing. Yell immunization against fits provided about a 75 immunit of first/senses tab in proverting baselabations from ... Effectiveness of Flu Shots Wildy Overestmated - Infuenza Veccine ... In a flaveocine celosis published in the Canadian Medical Association Journal about the effectiveness of the mass influence resciration program in Defants, ... Obesity May Hinder Flu Shots Effectiveness - Yahool News Flu shot: Your best bet for evolding influenze - MayoClinic.com Fig shots are the most effective way to prevent influenza and its complications. The Are flu shots effective? - Flu treatments naturally Fig shots are ineffective against new strains of fig, and only moderately effective Fig anota are instructive against new strains of ful, and only modelabley after analised common house. They also come all's series side affects. These are Get the flu veccine very fightly ca Here you getter your the shed this year? Are Flu Shots Effective? The Truth About File Shell Effective receiption The File Shell Proport Today. Goooooooogle > No Advanced search Search Help Sive us feedback Go to Google.com Marker Mary Description of the region reaction 11-111 07/84 which here Types 00 00 01 0000-10 000 10 000 Google Home Advertisies Programs Business Solutions Privacy About Google

▲ロ ▶ ▲周 ▶ ▲ 国 ▶ ▲ 国 ▶ ● の Q @

Pan et al. (2007); Novarese & Wilson (2013) Yom-Tov et al. (2013); Glick et al. (2014) Epstein & Robertson (2015)

# Engagement matters for platforms!



#### Weight Decision 12/15/2017

| *<br>Component   | Final Weight<br>for 2018Q1 |  |
|--|----------------------------|--|
| Like   | 1                          |  |
| Reaction,<br>Reshare without Text                                | 5                          |  |
| Non-sig Comment, Non-sig Reshare<br>Non-sig Message, Rsvp        | 15                         |  |
| Significant Comment,<br>Significant Reshare, Significant Message | 30                         |  |
| Groups Multiplier<br>(Non-friends)                               | 0.5                        |  |
| Strangers Multiplier<br>(non-friend-of-friend, small pages)      | 0.3                        |  |

 $\Downarrow$ 

*Facebook* whistleblowers (*WSJ*, 2021): MSI allegedly led to adverse effects in terms of misinformation and polarization (among others)



# This paper

1) Theoretical framework

Interactions of behavioural individuals with algorithmic weights

- $\Rightarrow$  Assess impact of an increase in MSI & personalization on:
  - Platform Engagement
  - Misinformation
  - Polarization

Main insights:

MSI: ↑ Engagement; ↑ Misinformation; ↑ Polarization

2) Direct empirical evidence on impact of MSI on polarization

### Model

**State** of the world  $\theta \in \mathbb{R}$  (e.g., net benefits of vaccines/emission reduction)

M news items (e.g., Facebook's post, Tweet, etc).

• Each carries an informative signal  $y_m \sim N(\theta, \sigma_y^2)$ .

#### N individuals:

- Each receives a private informative signal  $x_n \sim N(\theta, \sigma_x^2)$ .
- Sequentially access (in random order) a social media platform to read and, possibly, "highlight" (e.g., share) a news item m
- Are able to see whether m is "like-minded" or not. Yet they need to click on the news item in order to see y<sub>m</sub>.

・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・

# Model – Clicking (absent ranking)

 $\gamma_n =$  individual *n*'s propensity to click on "like-minded" news, *absent ranking* 

Individuals can be of three clicking types:

- confirmatory (τ<sub>C</sub>): more likely to click on "like-minded" news (γ<sub>C</sub> > 1/2)
- exploratory ( $\tau_E$ ): less likely to click on "like-minded" news ( $\gamma_E < 1/2$ )
- indifferent (ranking-driven)  $(\tau_I)$ :  $\gamma_I = 1/2$

The three types occur with probabilities  $p_C \ge 0$ ,  $p_E \ge 0$ , &  $p_I = 1 - p_C - p_E$ .

・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・

# Model – Highlighting

After clicking on m, individual sees the actual signal ym

Then highlight (e.g., share/comment) m with probability p<sub>a</sub>

Assumptions:

- ▶ Highlight only if sufficiently close to prior  $(|x_n y_m| < \frac{\sigma_x}{2})$ , An et al. 2014; Garz et al 2020)
- Individuals with more extreme priors are more likely to highlight. (Bakshy, Messing, Adamic, 2015, for "hard" news (i.e., political).

▶ Bakshy et. al (2015)

A D N A 目 N A E N A E N A B N A C N

#### Model – Attention Bias

Individuals have an attention bias calibrated by  $\beta \geq 1$ .

#### Interpretation:

If news items  $m_a$  and  $m_b$  have the same sign and  $m_a$  is one position up in the ranking  $\Rightarrow m_a$  will be  $\beta$  times more likely to be clicked wrt to  $m_b$ 

▲□▶▲□▶▲≡▶▲≡▶ ≡ めぬぐ

#### Model – Attention Bias

Individuals have an attention bias calibrated by  $\beta \geq 1$ .

#### Interpretation:

If news items  $m_a$  and  $m_b$  have the same sign and  $m_a$  is one position up in the ranking  $\Rightarrow m_a$  will be  $\beta$  times more likely to be clicked wrt to  $m_b$ 

#### All in all, the higher:

- a) the ranking of news item m;
- b) the propensity (absent ranking) of individual n to click on m

 $\Rightarrow$  the more likely *m* is to be clicked  $\bigcirc$  Clicking Prob.

(Germano, Gómez, Le Mens 2019; Germano and Sobbrio, 2020)

### Model – Algorithm: Popularity Ranking

Ranking algorithm updates popularity of each news item such that:

- a click has a weight of 1
- ▶ a highlight has a weight of  $\eta \in \mathbb{R}_+$ .

*Popularity* of news item m,  $\kappa_{n,m}$  updated according to:

$$\kappa_{n,m} = \kappa_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ 1 & \text{if } m \text{ is clicked on and not highlighted by } n \\ 1 + \eta & \text{if } m \text{ is clicked on and highlighted by } n \end{cases}$$

*Ranking* observed by *n* inversely related to popularity before clicking:

$$r_{n,m} < r_{n,m'} \iff \kappa_{n-1,m} < \kappa_{n-1,m'}.$$

・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・

## Recap

At time t = n, (random) individual n:

- Gets private signal  $x_n$  on  $\theta$  (e.g., net benefits of vaccine)
- Access social media and observes ranking of news items  $r_{n,m}$
- Given ranking, attention bias β, and propensity to choose like-minded items γ<sub>n</sub>: decides which m to click
- After learning y<sub>m</sub>, highlights m with probability p<sub>a</sub> and only if sufficiently close to her prior
- Algorithm updates the popularity (and ranking) of items:

$$\kappa_{n,m} = \kappa_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ 1 & \text{if } m \text{ is clicked on and not highlighted by } n \\ 1 + \eta & \text{if } m \text{ is clicked on and highlighted by } n. \end{cases}$$

◆□▶ ◆□▶ ◆注▶ ◆注▶ 注 のへで

At time t + 1 = n + 1....

Algorithm personalizes the ranking according to whether  $x_n$  on the left/right wrt  $\theta$  (group L/R)

Two rankings based on two separate measures of popularity.

 $\lambda \in [0,1]$  is a parameter calibrating the degree of personalization.

- $\triangleright$   $\lambda = 0$ : clicks and highlights from the other group do not count at all
- >  $\lambda = 1$ : no personalization: popularity for both groups always coincides.

#### **Evaluation indices**

Effects of  $\eta$  and  $\lambda$  on engagement and users' welfare?

Engagement: ENG= sum of clicks and highlights

• Misinformation: 
$$MIS = \frac{1}{N} \sum_{n \in N} |y(n) - \theta|;$$

• Polarization: 
$$POL = \frac{1}{N} |\sum_{n \in R} y(n) - \sum_{n' \in L} y(n')|;$$

where:

- ▶  $y(n) \in M$  denotes the signal of the news item clicked on by individual n.
- ▶ L(R) denotes the individuals with signals  $x_n$  with sign $(x_n) = -1$  (= +1).

▲□▶ ▲□▶ ▲□▶ ▲□▶ ▲□ ● ● ●

### Main results - Crowding out the truth



Figure: Users' clicking behavior (top) and highlighting behavior (bottom) for small  $\eta$  (left) and for large  $\eta$  (right) under non-flat highlighting.

▲ロ ▶ ▲周 ▶ ▲ 国 ▶ ▲ 国 ▶ ● の Q @



#### Main results: Intuition

An increase in  $\eta$ :

- More individuals willing to highlight items (more extremist) will be clicking on items they are actually interested in highlighting → Higher engagement
- Individuals less likely to click on news near the truth (y's ≈ θ)
  & more likely to click on items further away from the truth (y's ≈ -x\*, x\*). → More misinformation and polarization

・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・

 $\Rightarrow$  Crowding out the truth.

# MSI and Polarization: Empirical analysis

Theoretical prediction: an increase in weight of "highlights"  $(\eta)$ 

∜

Individuals more exposed to extremists contents
 Higher level of political polarization.

Empirical test: exploit Facebook's MSI update

Jan 2018: boost in the weight given to comments and shares

▲□▶ ▲□▶ ▲ 三▶ ▲ 三▶ 三 のへぐ

# Empirical analysis (2)

Data. Focus on Italy (IPSOS *Polimetro*):

- Weekly interviews on representative sample of Italian voting pop.

Info on whether internet primary source to form pol. opinion

 Italy 2017-2018: FB by far the first social media: 60% penetration rate (Twitter 23%), ~ 80% among internet users

・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・

- Ideological self-position: Dummy for moderate/non-moderate
- Probability of voting for each party: affective polarization.

### Empirical strategy

Difference-in-Differences:

 $Y_{i,m,t} = \alpha + \beta_1$ Opinion via internet<sub>i.m,t</sub> × Post MSI

 $+\beta_2$ Opinion via internet<sub>*i*,*m*,*t*</sub>  $+\beta_3$ Post MSI  $+\alpha_m + X_{i,t} + \varepsilon_{i,m,t}$  (1)

- Y<sub>i,m,t</sub> represents the outcome of interest relative to individual *i*, leaving in municipality *m* interviewed in the survey wave *t* (i.e., probability of declaring a non-moderate political ideology or weighted affective polarization).
- $\triangleright \alpha_m$  municipality fixed effect
- X<sub>i,t</sub>: socio-demographic control (age, gender, n. of resident family members, level of education, type of occupation, religiosity).
- Observations weighted according to Ipsos sampling weights

### Results: MSI and Non-moderate ideology

|                                       | (1)<br>Non-moderate | (2)<br>Non-moderate | (3)<br>Non-moderate | (4)<br>Non-moderate |
|---------------------------------------|---------------------|---------------------|---------------------|---------------------|
|                                       | Ideology            | Ideology            | Ideology            | Ideology            |
| Opinion via internet $	imes$ Post MSI | 0.062***            | 0.058***            | 0.051***            | 0.051***            |
|                                       | (0.016)             | (0.015)             | (0.014)             | (0.018)             |
| Opinion via internet                  | -0.012              | -0.006              | -0.012              | -0.012              |
|                                       | (0.020)             | (0.020)             | (0.024)             | (0.022)             |
| Post MSI                              | -0.017*             |                     |                     |                     |
|                                       | (0.009)             |                     |                     |                     |
| Observations                          | 25,690              | 25,690              | 25,690              | 25,690              |
| Mean outcome                          | 0.36                | 0.36                | 0.36                | 0.36                |
| SD outcome                            | 0.48                | 0.48                | 0.48                | 0.48                |
| Municipality FE                       | YES                 | YES                 | YES                 | YES                 |
| Date of interview FE                  | NO                  | YES                 | NO                  | NO                  |
| Province-Date of interview FE         | NO                  | NO                  | YES                 | YES                 |
| Cluster SE                            | Region              | Region              | Region              | Province            |

#### Table: MSI and non-moderate ideological position

Note: Time horizon: June 2017-June 2018 . Robust Standard Errors in parenthesis. \*\* p<0.01, \*\* p<0.05, \* p<0.1

### Results: MSI and Affective Polarization

#### Table: MSI and Affective Polarization

|  | (1)<br>Affective<br>Polarization | (2)<br>Affective<br>Polarization | (3)<br>Affective<br>Polarization | (4)<br>Affective<br>Polarization |  |
|--|----------------------------------|----------------------------------|----------------------------------|----------------------------------|--|
| Opinion via internet $\times$ Post MSI                                     | 0.054**                          | 0.055**                          | 0.073***                         | 0.073***                         |  |
| Opinion via internet   | -0.012                           | -0.011                           | -0.006                           | -0.006                           |  |
| Post MSI   | (0.023)<br>0.118***<br>(0.020)   | (0.022)                          | (0.023)                          | (0.025)                          |  |
| Observations<br>Mean outcome   | 14,499<br>1.29                   | 14,499<br>1.29                   | 14,499<br>1.29                   | 14,499<br>1.29                   |  |
| SD outcome   | 0.61                             | 0.61                             | 0.61                             | 0.61                             |  |
| Municipality FE<br>Date of interview FE<br>Province-Date of interview FE   | YES<br>NO<br>NO                  | YES<br>YES<br>NO                 | YES<br>NO<br>YES                 | YES<br>NO<br>YES                 |  |
| Cluster SE   | Region                           | Region                           | Region                           | Province                         |  |
| Nate: Time having 1, une 2017 June 2018 Debugt Standard Every in acceptage |                                  |                                  |                                  |                                  |  |

Note: Time horizon: June 2017-June 2018 . Robust Standard Errors in parenthesis. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

# Conclusions

A higher  $\eta$  ( $\uparrow$  weight on highlights in the ranking algorithm)

- Assuming bimodal propensity to highlight (Bakshy et al. 2015):
  - increases engagement
  - increases polarization
  - increases misinformation.
- ▶ Higher ideological extremism & affective polarization in Italy
- $\Rightarrow$  Theoretical & Empirical evidence on adverse effects of FB 2018 MSI update
  - A lower  $\lambda$  ( $\uparrow$  personalization) increases engagement & polarization.
- $\Rightarrow$  Theoretical support for "filter bubble" (Pariser, 2011)

