Curtailing False News, Amplifying Truth (Guriev, Henry, Marquis, and Zhuravskaya)

Comments

Giacomo Lemoli

IAST & TSE

A timely study

A timely study

Are concerns about ending fact-checking justified? *It depends*

A timely study

Are concerns about ending fact-checking justified? *It depends*

(Disclaimer: I will probably say "Twitter")

Overview

Truly impressive and relevant work

Overview

Truly impressive and relevant work

- High-stake setting: 2022 US mid-term elections
- Realistic setting: true/false tweets, $\sim$ sharing decision
- Different policies of interest: extra-click, priming, fact-check offer, assessment
- Formal model of sharing: incorporates reputation, persuasion, and signaling
- Experiment + structural estimation
- "Incentive-compatibility" of policies (quality of sharing + engagement)
- Very well-written and clear

Overview

Truly impressive and relevant work

- High-stake setting: 2022 US mid-term elections
- Realistic setting: true/false tweets, $\sim$ sharing decision
- Different policies of interest: extra-click, priming, fact-check offer, assessment
- Formal model of sharing: incorporates reputation, persuasion, and signaling
- Experiment + structural estimation
- "Incentive-compatibility" of policies (quality of sharing + engagement)
- Very well-written and clear

Broad comments:

- Experiment and external validity
- Theoretical mechanisms
- Policy implications

Experiment

- Sample not representative of Twitter users (educated, Democratic...)
- More discussion of the pre-treatment variables (esp. political views and information, Twitter activity)

Experiment

- Sample not representative of Twitter users (educated, Democratic...)
- More discussion of the pre-treatment variables (esp. political views and information, Twitter activity)
- For ATE: compare to effects on different samples in the literature
- Or: characterize the sub-population these results apply to (platform can tailor policies to profiles)
- Heterogeneity dimensions underexplored: profile of people more likely to share false vs true news under each regime (e.g. age Guess et al 2019)

Experiment

- Sample not representative of Twitter users (educated, Democratic...)
- More discussion of the pre-treatment variables (esp. political views and information, Twitter activity)
- For ATE: compare to effects on different samples in the literature
- Or: characterize the sub-population these results apply to (platform can tailor policies to profiles)
- Heterogeneity dimensions underexplored: profile of people more likely to share false vs true news under each regime (e.g. age Guess et al 2019)
- Why fact check offered on just 2 tweets? What do they infer about the other two tweets?
- Can people seek **external information**? E.g. can you check the time spent on the tweets page after each treatment?

Theoretical mechanisms

- Surprising that the asymmetry of priming effects does *not* reflect similar asymmetry in updating (as if people became *generally* more skeptic)
- Maybe some people decide to post tweets that they thought true, but would not post. But then we should see cost channel effect
- Alternatively, priming makes someone "shift" from false to true tweet, while "preventing" people from not sharing anymore
- Would be nice to understand better: what true tweets are shared more? Those with highest prior?
- How does priming help update on the false news? Do people use heuristics? (here the false economic tweet has strong cues and is the one with flatter prior). Would be the same for more "ambiguous" tweets?
- Correlation between veracity and partisan alignment: implications for structural estimates?

Theoretical mechanisms

- Impressed by relatively low relevance of partisan motives **in high-stake setting**
- Fact-check offer increases the (cognitive?) cost of sharing rather than increasing the quality of sharing
- A bit surprising that assessment treatment does not lead to updating (mechanism should be $\approx$ priming). Suggests that "framing" matters?

Implications for policy

- Results help making sense of dismissal of "fact-checking" nudge on Twitter (reading article before RT): possibly decreasing engagement
- Community notes may actually be better: clearly increase salience of reputation (assuming you are a good faith/real person)

Implications for policy

- Results help making sense of dismissal of "fact-checking" nudge on Twitter (reading article before RT): possibly decreasing engagement
- Community notes may actually be better: clearly increase salience of reputation (assuming you are a good faith/real person)

On simulations:

- What if you allow for more informed people? Currently 4% but the threshold may be too high. Digital literacy is more about detecting fake news with good chance rather than being on top of all news
- Heterogeneous profiles of people posting fake news vs true news after each treatment may suggest more targeted policies
- Implications considering also the *production* of fake news and how they would react to these policies