"Content moderation for sale: pricing attention through steering and certification"

Heski Bar-Isaac, Rahul Deb, and Matthew Mitchell

Discussant: Yossi Spiegel

17th Digital Economics Conference, Toulouse, January 9-10, 2024

The main idea

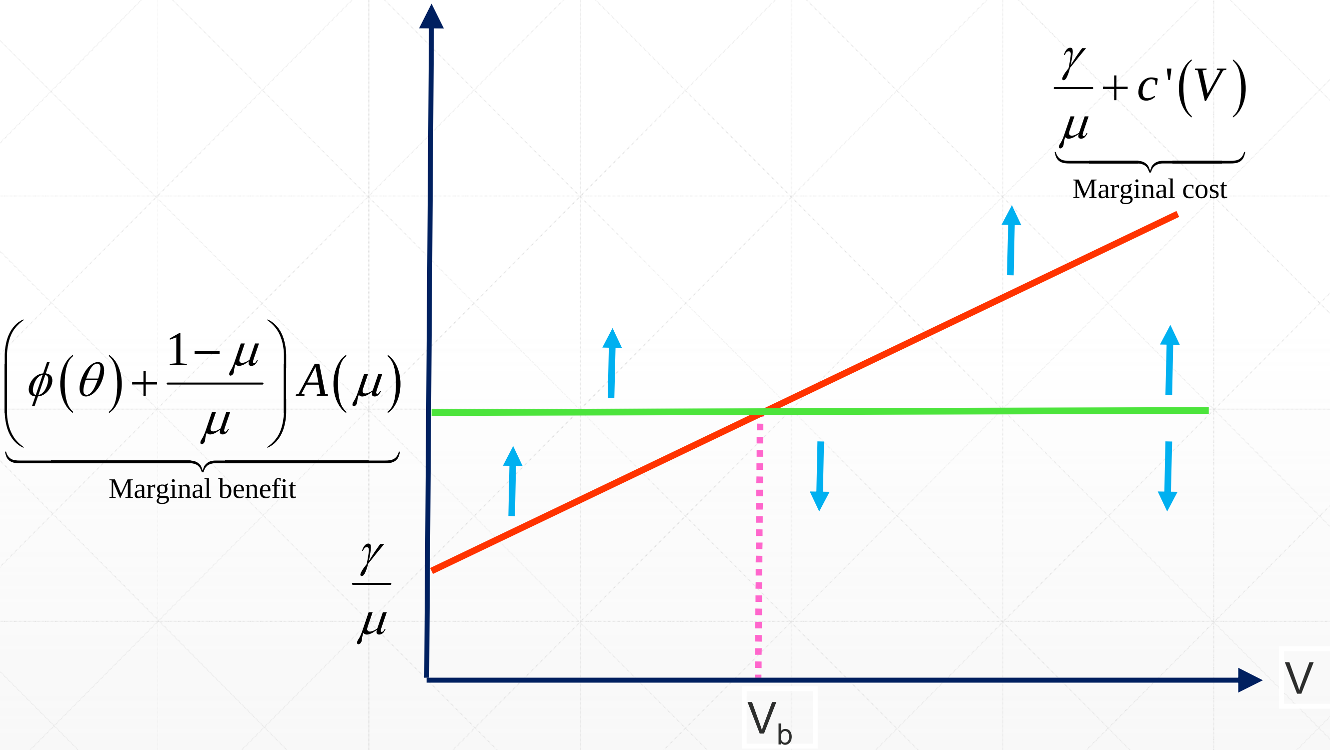
- A platform steers consumers to content and charges content providers for steering
- Content can be good of bad; consumers care only about good content
- The platform can certify that content is good. Does it have the right incentive to do so?
  - No: the platform wants to also certify bad content which generates extra revenue from bad content providers
- What are the welfare implications of this behavior?
  - Surprisingly better than we may think

The main idea

- Good content providers value views but how much is private information
- The platform engages in 2nd degree PD: it offers a menu with the no. of views, V(θ), and a payment, P(θ) (and possibly also the quality of the certificate which asserts that content is good with prob. μ(θ))
  - Why PD and not simple monopoly problem? Do the results depend on PD? Why?
- The FOC for the optimal V(θ) when there's only one type of certificate:

$$\underbrace{\underbrace{\frac{\gamma}{\mu}}_{\text{Cost of targeting}} + \underbrace{c'(V)}_{\text{Cost of views}}}_{\text{Marginal cost}} = \underbrace{\left(\underbrace{\phi(\theta)}_{\text{Virtual value}} + \underbrace{\frac{1-\mu}{\mu}}_{\text{No. of bad views per each good view}}\right)\underbrace{A(\mu)}_{\text{Prob. of attention}}}_{\text{Marginal benefit}}$$

Illustrating the main idea



$\frac{\gamma}{\mu}+c'(V)$
Marginal cost
$\left(\phi(\theta)+\frac{1-\mu}{\mu}\right)A(\mu)$
Marginal benefit
$\frac{\gamma}{\mu}$
V_b
V

If μ↓ (more bad content is certified):

- MC↑ (more targeting of bad content) ⇒ V_b↓
- A(μ)↓ (lower WTP of good content) ⇒ V_b↓
- (1-μ)/μ↑ (more income from bad content) ⇒ V_b↑

V_b may ↑

The highest θ for which V_b = 0 may↓ ⇒ Diversity ↑

Caveat: consumers do not care about diversity per se (all good content is equally good for consumers)

2^{nd} degree PD with a continuum of certificates

- Here the platform offers a menu, $V(\theta)$, $P(\theta)$, and $\mu(\theta)$
- Imperfect certification intentionally damages the quality of good content
  - It lowers the WTP of good content providers to pay the platform ⇒ why do it?
  - Damaging quality means "sell a certificate to bad content and make consumers more hesitant to pay attention"
  - The literature on damaged goods (e.g., Deneckere and McAfee, JEMS 1996) shows that damaging a good can help screen consumers
  - Here, bad content providers pay for the fake certificate that damages good content

Comments

- The model can be used to study a general 2nd PD problem with two types of customers: bad customers impose a negative externality on good customers
- Selling to bad customers lowers the WTP of good customers but generates an additional revenue

Policy

- Regulators nowadays are increasingly more concerned that platforms abuse their market power and influence what users view
- The paper shows that platforms may intentionally certify bad content to boost their views but that actually has a bright side: more good content is channeled to consumers
- But the model does not account for a few important considerations:
  - Bad content here is not "bad": it's useless. In reality, though
    - Bad content can harm users (e.g., psychological damage, misleading information)
    - Bad content imposes negative externalities (e.g., promoting violence, affects elections)
  - It's true that in the FB here the platform certifies only good content, but the consequences of bad content that consumers view are not that "bad"

Policy

- In reality, a platform cannot perfectly verify that content is good/bad
  - How imperfect verification of content matter?
- In reality there's cognitive overload: attention is limited and more views diminish the value of each piece of content
  - In the model, consumers can observe unlimited amount of content and all good content is equally good
  - If attention is limited, more is not necessarily better
  - If consumers care more about high θ content, then more diversity implies a lower expected quality drops so consumers are worse off
- The supply of content and its quality (good/bad) are exogenous; how's does the platform behavior affect the supply of content if it is endogenous?

Concluding remarks

- The paper deals with a topical and important problem and offers a clever model to study it
- There are many open question that can and should be addressed but this is a good starting point

Comments

- The theory is pretty convincing: why do we need lab experiments?
- Suppose the lab experiments were inconsistent with the theory: what does it mean?
  - The theory is wrong?
  - The lab does not replicate the model well?
- Experiments are best when they tell us how people behave or think
  - Example: lab experiments (e.g., Copper and Kühn, AEJ: Micro 2014) show that absent communication, it's hard to agree on collusion
- Here what is tested is the prediction from a model; not how people behave and think
- "Wind tunnel" works best when the lab replicates the model closely (auction formats); the lab cannot replicate a real-life cartel