Content Moderation for Sale: Pricing Attention through Steering and Certification

Heski Bar-Isaac, Rahul Deb and Matt Mitchell

TSE Digital Economics Conference
January 2025

The allocation of scarce attention resources on the Internet

- Content wants attention
- This requires being seen *and* being trusted
- Platforms control *both* aspects through steering views and through the way things are presented (explicit certification, ancillary information, display choices)

Model elements

- A monopoly "social media" platform
  - No content-to-consumer pricing
- Platform can distinguish good content from bad content (which is all that consumers care about) and can choose to be perfect quality
- But platform cannot tell how much good content values being seen
- Take a mechanism design approach: platform offers combinations of views, certification, and associated price and different types choose what suits them best
  - discrimination among types where use *both* quantity (views) and quality (certification)

Specific Questions

- If the platform controls what is seen, what role does certification play?
  - Imperfect certification increases the value to the platform of content providers with low willingness to pay for views (through opportunities to raise revenue from bad content)
- How might certification impact platform diversity?
  - Imperfect certification can improve content diversity and even consumer welfare relative to perfect certification
- What kinds of platforms are most likely to certify perfectly?
  - Platforms with higher opportunity costs of views
  - Platforms where consumer attention in perceived quality is convex

Model: Players

- Many content providers (pieces of content): differ in quality (good or bad) and value of attention
- One platform: observes content quality but not value of attention; can direct content and messages to different consumers in exchange for money from content providers.
- Many consumers: decide what to pay attention to based on certification messages: If believe content is good with probability μ, read it with probability $A(\mu)$ *"the attention function"*

Content Providers

- Good or bad
- Unit mass of good with private value $\theta \in [0, \overline{\theta}]$ distributed according to $F()$ with positive density
  - Value engagement at $\theta A v_g$ where v_g corresponds to *interested* views and a to attention
- Unlimited bad content
  - Value attentive views at $A v_b$ (either don't care about interest or with so many bad bots allocating interested views is trivial)

Platform costs

- Platform can distinguish good and bad providers
- Directing v_b views for a bad content provider costs the platform γv_b where $\gamma \in (0, 1)$.
- Directing v_g *interested* views for a good content provider costs the platform $\gamma v_b + c(V_g)$
  - Increasing, strictly convex, differentiable cost of finding interested users with $c(0) = c'(0) = 0$
  - $c(.)$ incurred at the level of individual content provider (rather than aggregate)

Platform Mechanism

- $M : \Theta \to \mathbb{R}_+$: Represents the message or certification assigned to an entity.
- $V_g : \Theta \to \mathbb{R}_+$: Denotes the number of good views assigned.
- $V_b : \Theta \to \mathbb{R}_+$: Denotes the number of bad views assigned.
- $P : \Theta \to \mathbb{R}_+$: Represents the price assigned to the entity.

Leading to quality of certificate

$$\mu(m) = \frac{\mathbb{E}\left[V_g(\theta) \mid M(\theta) = m\right]}{\mathbb{E}\left[V_g(\theta) + V_b(\theta) \mid M(\theta) = m\right]},$$

Platform's Problem

In principle, this is complicated, but

- there is always an equivalent mechanism where type has its own certificate (if pool then same mix of V_b to ensure the same μ and so can think about $\mu(\theta)$ as the quality provided to type θ)
- Pointwise solutions will be solutions to the overall problem as long as $A(\mu(\theta))V_g(\theta)$ is non-decreasing,

$$(V_g^p(\theta), \mu^p(\theta)) \in \arg\max_{v_g, \hat{\mu}} \left[\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}) v_g - c(v_g) - \gamma \frac{v_g}{\hat{\mu}} \right]$$

  where $\phi(\theta) = \theta - \frac{1-F(\theta)}{f(\theta)}$ is the virtual value of type θ and assumed to be increasing
- Compare to Mussa-Rosen: additional revenue and costs associated with bad content, and implications for revenue through attention

Benchmark 1: The engagement maximizing planner

- Consumers don't value bots

  $\implies$ pure certification $\bar{\mu}(\theta) = 1$ for all θ

- Consumers don't care how much providers want to be seen

  $\implies$ egalitarian content i.e. $\bar{V}_g(\theta)$ is constant for all θ

  $\implies$ generate views to the point that marginal cost is equal to the marginal benefit

$$\gamma + c'(\bar{V}_g) = A(1) = 1$$

Perfect certification: A natural benchmark to consider

European Commission

EN

Search

Search

Home > Press corner > Commission sends preliminary findings to X for breach of DSA

Available languages: English

PRESS RELEASE | Jul 11, 2024 | Brussels | 3 min read

Commission sends preliminary findings to X for breach of the Digital Services Act

PAGE CONTENTS

Top

Quote(s)

Today, the Commission has informed X of its preliminary view that it is in breach of the Digital Services Act (DSA) in areas linked to dark patterns, advertising transparency and data access for researchers.

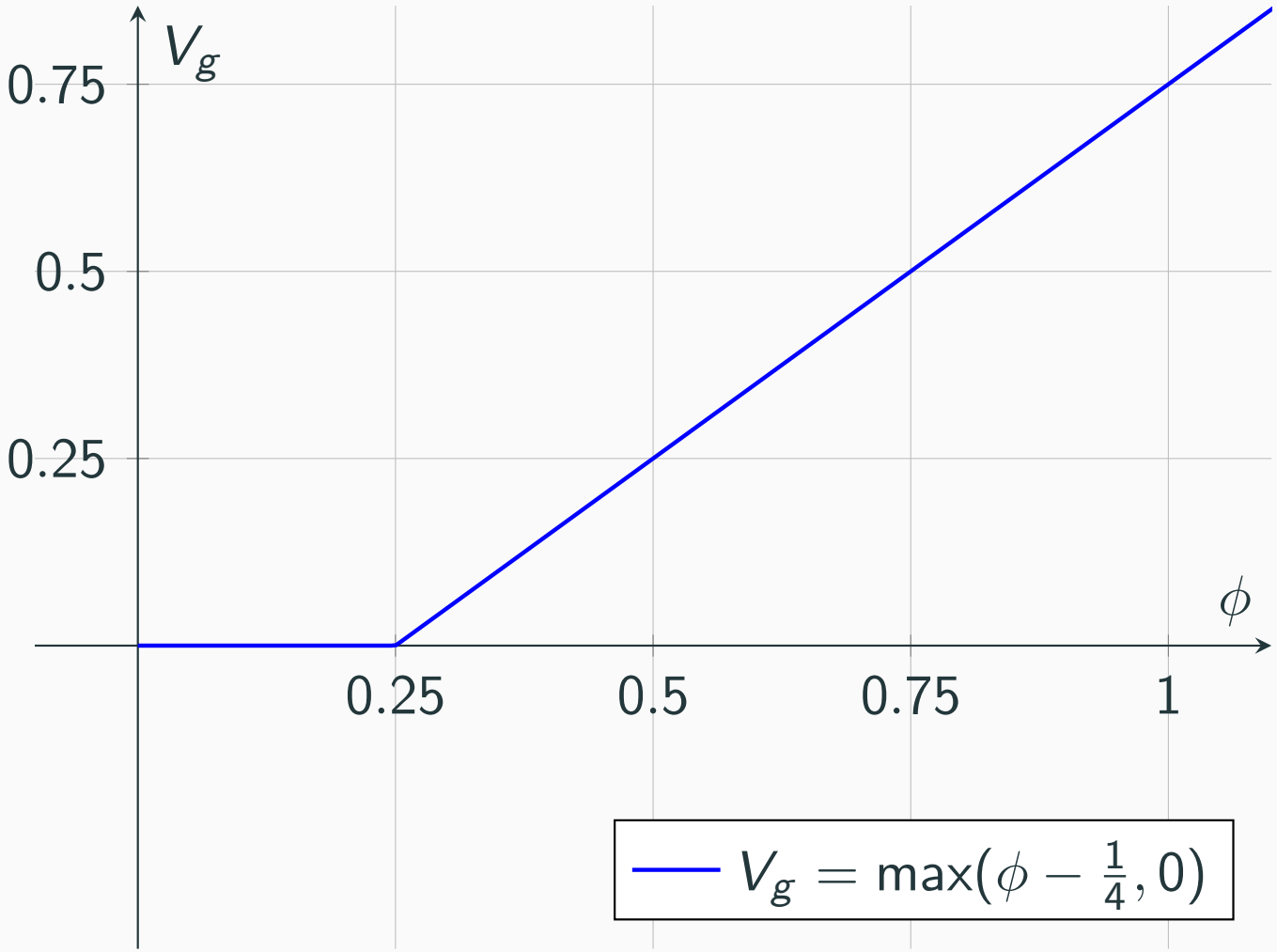
Perfect certification

- Consider perfect certification $\mu(\theta)) = 1$ then the only thing for the platform to choose is the number of views.
- The FOC with respect to views writes as

$$\gamma + c'(V_g(\theta)) = \phi(\theta) = A(1)\phi(\theta)$$

- Just like Mussa-Rosen: price discrimination brings distortion from planner problem since benefit is virtual value rather than social benefit (which is 1).
- Here that means a shift away from egalitarian content

Views, Perfect Certification, $\gamma = 1/4$, $c(x) = x^2/2$



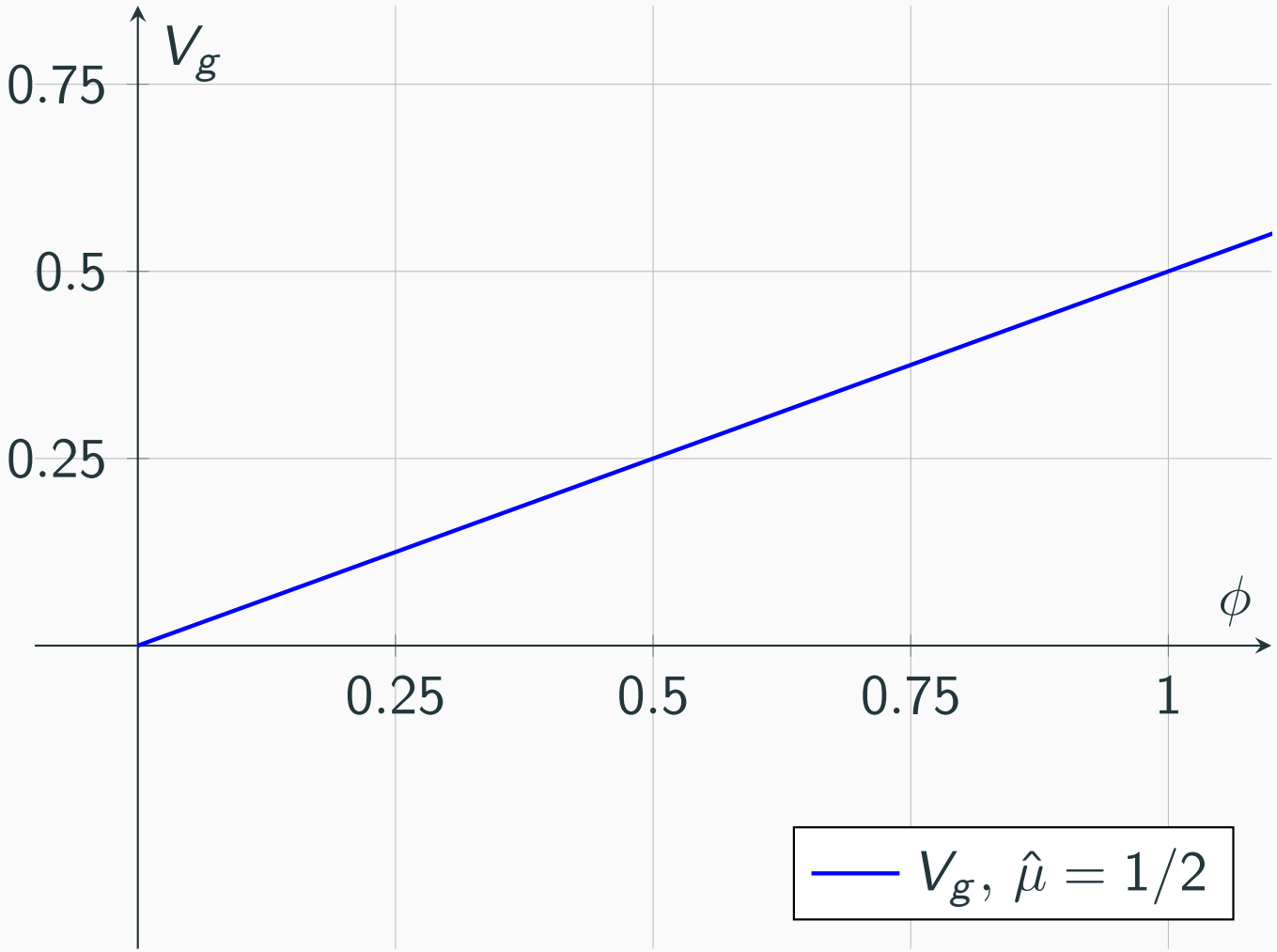
Single imperfect certificate

- Consider exogenously imperfect certification $\mu(\theta) = \hat{\mu}$, again the only for the platform to choose is the number of good views
- but each additional good view comes with $\frac{1-\hat{\mu}}{\hat{\mu}}$ bot views (and their associated revenue).
- Now the FOC with respect to views writes as

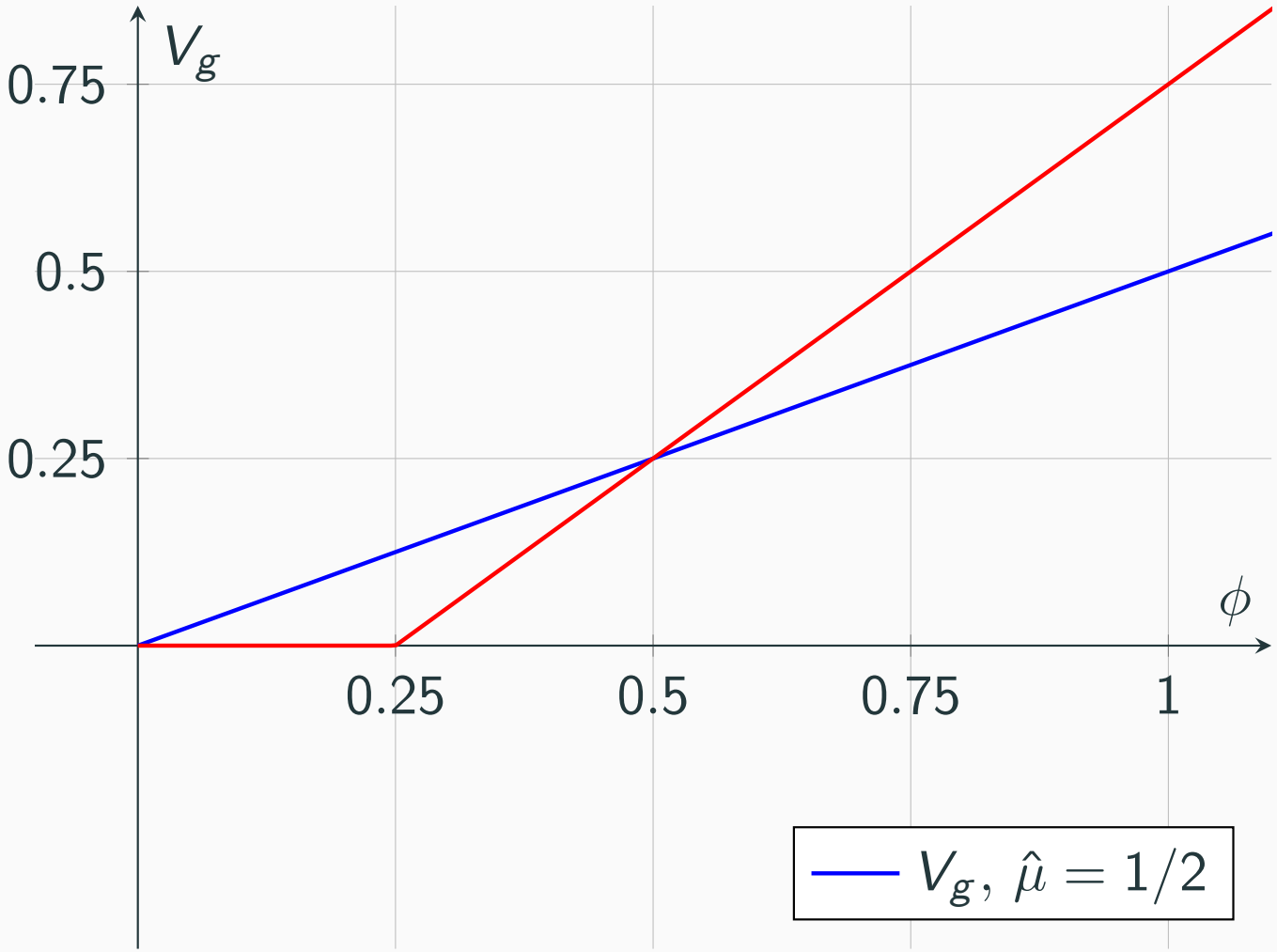
$$\frac{\gamma}{\hat{\mu}} + c'(V_g(\theta)) = (\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}})A(\hat{\mu})$$

More egalitarian than perfect certification for those served

V_g, Imperfect Certification, $\gamma = 1/4$, $\hat{\mu} = 1/2$, $c(x) = x^2/2$



V_g, Imperfect Certification, $\gamma = 1/4$, $\hat{\mu} = 1/2$, $c(x) = x^2/2$



Moving beyond a single certificate

- Single imperfect certificate allows platform to monetize bad content
- But might be sacrificing a lot from high-value genuine content
- Varying certificate quality means that platform can use polluted certificates to low-value content to monetize bots, without sacrificing as much revenue from high-value good content
- And so want to use *both* instruments to help with price discrimination
    - bad certificates less appealing so don't have to curtail views as drastically

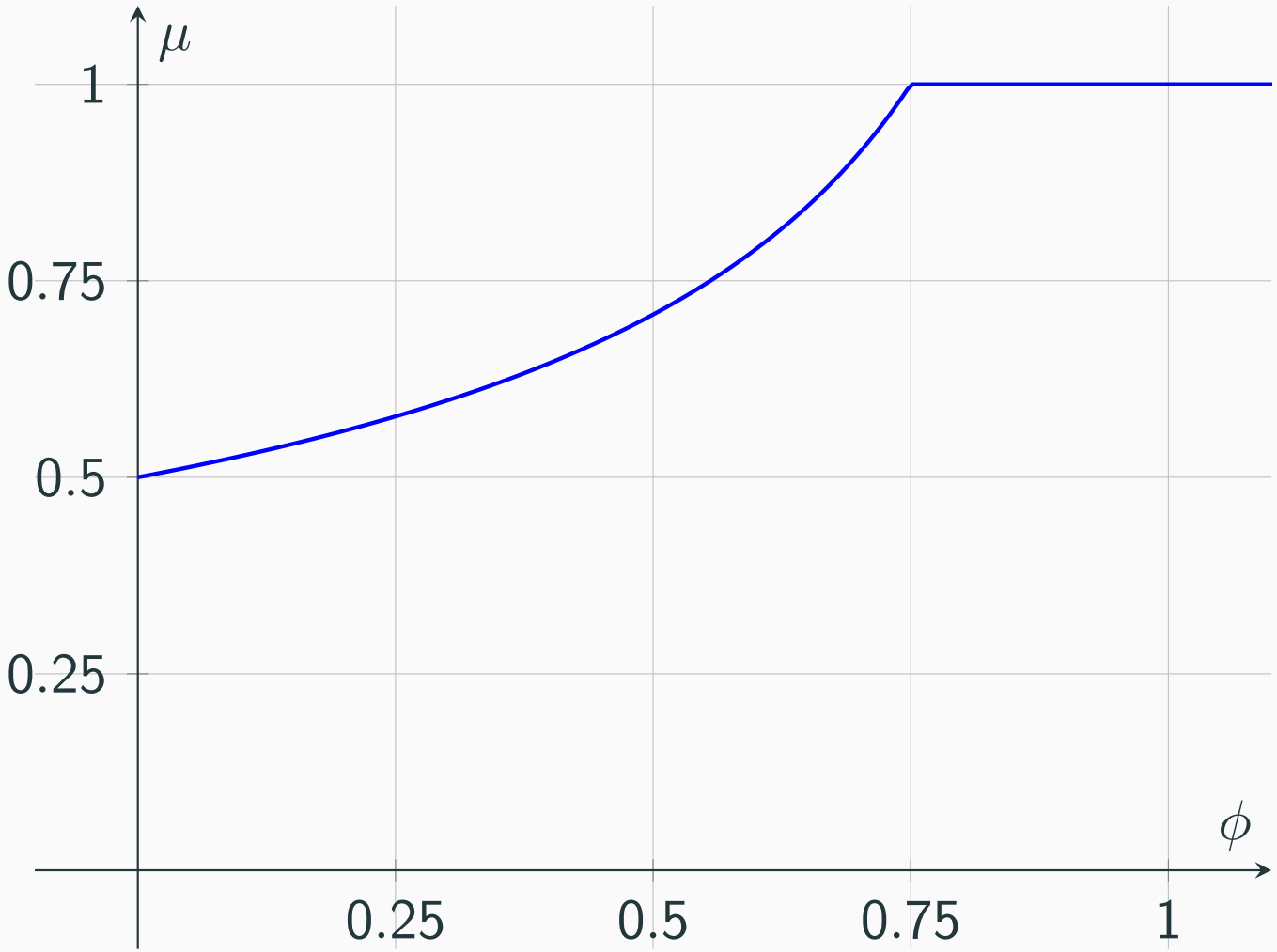
Characterization

- Can show that both $\mu^*(\theta)$ and $V_g^*(\theta)$ are non-decreasing
- FOC for views is

$$\frac{\gamma}{\mu^*(\theta)} + c'(V_g^*(\theta)) = (\phi(\theta) + \frac{1-\mu^*(\theta)}{\mu^*(\theta)})A(\mu^*(\theta))$$

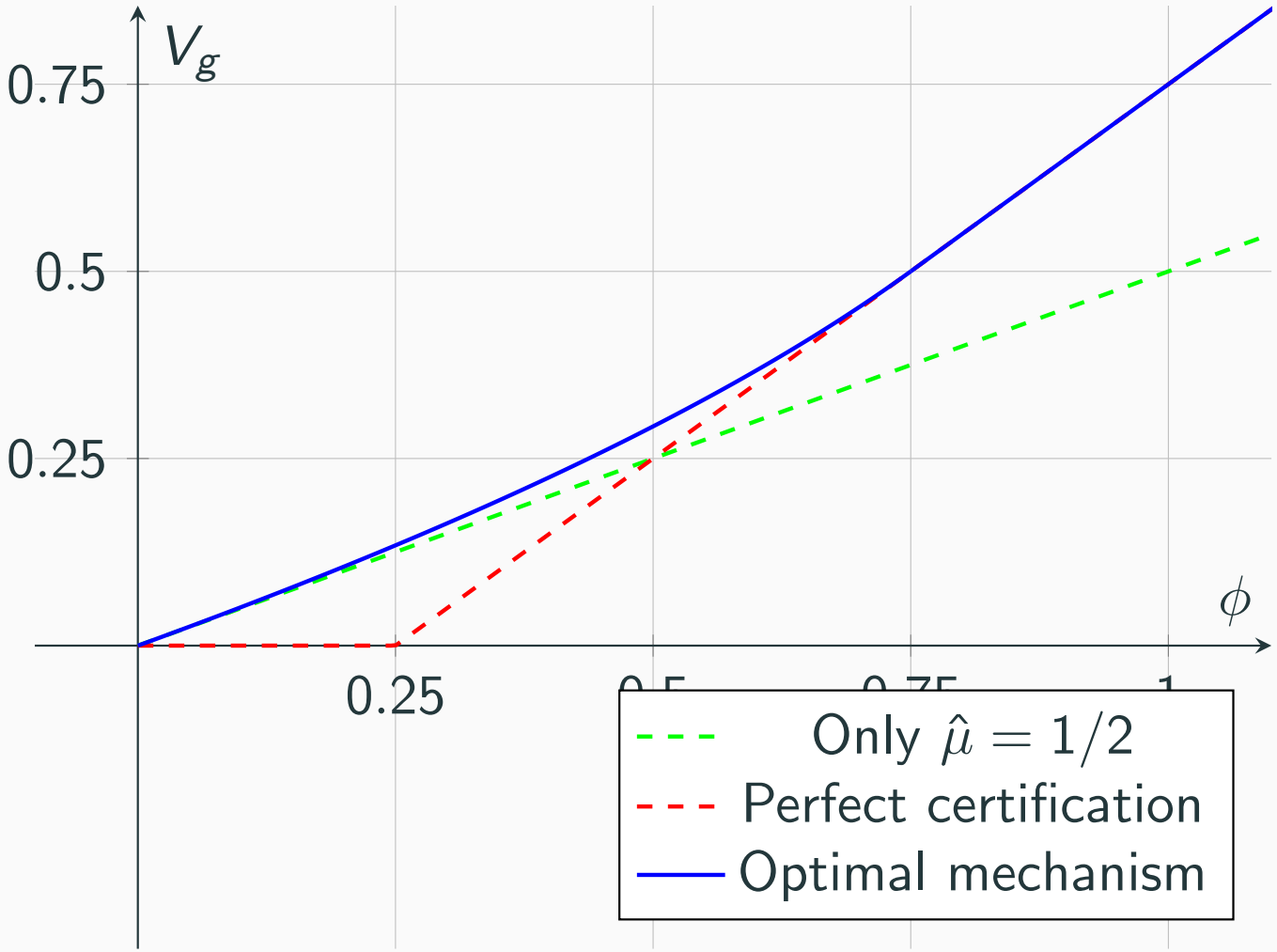
More content diversity than single checkmark

The optimal mechanism: Continuously imperfect certificates



$A(\mu) = \mu, \gamma = 1/4,\ c(x) = x^2/2$

The optimal mechanism: Content Diversity



Comp stats analyzed in the paper

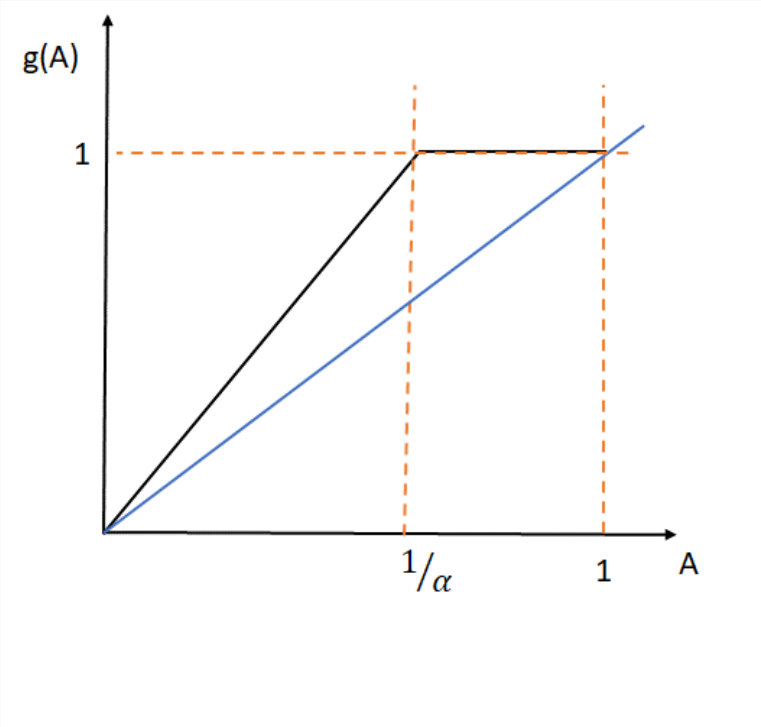
- Cost of ads (γ): Cheaper to run bot traffic $\implies$ more bad content
- improved targeting ($\kappa c(V_g)$) Cheaper targeting $\implies$ more good views; content skews to high-value content
- Transform attention $A(\mu)$ to be more concave then certificate quality is (weakly) lower

Shape of Attention

Proposition:
Suppose $\hat{A}(\mu) = g(A(\mu))$ for some increasing, differentiable, concave $g(\cdot)$ with $g(0) = 0$ and $g(1) = 1$. Then, for all θ, $\mu^*(\theta)$ is weakly lower under $\hat{A}(\mu)$ than under $A(\mu)$.

Intuition:

- Consider the concave transformation $g(A) = \min\{1, \alpha A\}$ with $\alpha > 1$.
- No reason exists to provide certification better than μ that sets $A(\mu) = \frac{1}{\alpha}$.



Intuition cont

- For lower α, the first-order condition (FOC) with respect to μ is:

$$A(\mu) - \mu^2 \left(\varphi + \frac{1-\mu}{\mu} \right) A'(\mu) = \gamma$$

- Scaling A has a linear effect on the left-hand side and is equivalent to lowering γ to $\frac{\gamma}{\alpha}$, which reduces quality.

Other things we try to speak to in the paper

- what happens as γ approaches 0?
- what if consumers suffer harm from bad content?
- What if consumers are addicted to social media?

Small γ

- Limiting case where γ goes to 0 and consider $A(\mu) = \mu^{\alpha}$
- concave platforms ($\alpha < 1$) always perform worse that perfect certification
- convex platforms ($\alpha > 1$) can perform better that perfect certification

Conclusions

- More money to be made from using *both* steering and imperfect certification
- imperfect certification brings bad content
- imperfect certification can benefit consumers through content diversity
- the extent of imperfect certification depends on costs of providing views, and, critically, on consumers sensitivity to bot traffic——convexity of attention
- lively policy discussion on consumer protection that might also worry about factors we ignore
    - naivete
    - externalities (a la Bursztyn, Hanel, Jimenez and Roth (2023)

Haiku Summary

Certify poorly

so you can sell more to bad?

Attention matters.

Content Moderation for Sale: Pricing Attention through Steering and Certificati

Bar-Isaac, Deb, and Mitchell

Comments/critiques gratefully received at heski.bar-isaac@rotman.utoronto.ca (and rahul.deb@bc.edu; and matthew.mitchell@rotman.utoronto.ca)