Content Moderation in Presence of Fringe Platforms

Iván Rendo (Toulouse School of Economics)

14th AEDE Conference *Barcelona, Spain, 27 June 2024*

Introduction

- **Online** hateful/**unsafe content** considered bad *per se*, and:
  - e.g. Jiménez-Durán (2022) links online hate to **offline violence**
  - e.g. **20%** of terrorists radicalized **exclusively** online

(Hamiz and Ariza, 2022)

Introduction

- **Online** hateful/**unsafe content** considered bad *per se*, and:
  - e.g. Jiménez-Durán (2022) links online hate to **offline violence**
  - e.g. **20%** of terrorists radicalized **exclusively** online

(Hamiz and Ariza, 2022)

➡ EU Response: **Digital Services Act**

Introduction

- **Online** hateful/**unsafe content** considered bad *per se*, and:
  - e.g. Jiménez-Durán (2022) links online hate to **offline violence**
  - e.g. **20%** of terrorists radicalized **exclusively** online

(Hamiz and Ariza, 2022)

➡ EU Response: **Digital Services Act**

But… **users could migrate to small (fringe) platforms!**

Introduction

- **Online** hateful/**unsafe content** considered bad *per se*, and:
  - e.g. Jiménez-Durán (2022) links online hate to **offline violence**
  - e.g. **20%** of terrorists radicalized **exclusively** online

(Hamiz and Ariza, 2022)

➡ EU Response: **Digital Services Act**

But… **users could migrate to small (fringe) platforms!**

- Rizzi (2023), Agarwal et al. (2022)

Introduction

- **Online** hateful/**unsafe content** considered bad *per se*, and:
  - e.g. Jiménez-Durán (2022) links online hate to **offline violence**
  - e.g. **20%** of terrorists radicalized **exclusively** online

(Hamiz and Ariza, 2022)

➡ EU Response: **Digital Services Act**

But… **users could migrate to small (fringe) platforms!**

- Rizzi (2023), Agarwal et al. (2022)
  - ➡ ↑ **moderation** on Twitter = ↑ **migration** to fringe platforms

Introduction

- **Online** hateful/**unsafe content** considered bad *per se*, and:
  - e.g. Jiménez-Durán (2022) links online hate to **offline violence**
  - e.g. **20%** of terrorists radicalized **exclusively** online

(Hamiz and Ariza, 2022)

➡ EU Response: **Digital Services Act**

But… **users could migrate to small (fringe) platforms!**

- Rizzi (2023), Agarwal et al. (2022)

  ➡ ↑ **moderation** on Twitter = ↑ **migration** to fringe platforms

~ 6% of the US citizens use fringe platforms: Parler, Truth…

(Stocking et al., 2022)

Today

Today

Platforms' competition model to analyze the **net effect** of

Content Moderation on the level of Content Unsafety

…while **allowing** for **Migration*** to a **fringe, unmoderated** platform

Today

Platforms' competition model to analyze the **net effect** of

Content Moderation on the level of Content Unsafety

…while **allowing** for **Migration*** to a **fringe, unmoderated** platform

Main Contribution: unsafety cannot be canceled out

Madio & Quinn (2024); Liu et al. (2021), it can be. They do welfare analysis

Today

Platforms' competition model to analyze the **net effect** of

Content Moderation on the level of Content Unsafety

…while **allowing** for **Migration*** to a **fringe, unmoderated** platform

Main Contribution: unsafety cannot be canceled out

Madio & Quinn (2024); Liu et al. (2021), it can be. They do welfare analysis

Research Questions:

➡ How **users choice** is determined by **content moderation policies**

➡ How the **level of unsafe content** is affected by **users choice**

➡ **Characterize the optimal regulation to minimize unsafe content**

(Technical version)

Model

(Technical version)

Model

Content unsafety is given by a metric $\theta \in [0,1]$ (the higher, the unsafer)

⛔ (Technical version)

Model

Content unsafety is given by a metric $\theta \in [0,1]$ (the higher, the unsafer)

Content moderation policy (K): any $\theta > K$ cannot be posted

(Technical version)

Model

Content unsafety is given by a metric $\theta \in [0,1]$ (the higher, the unsafer)

Content moderation policy (K): any $\theta > K$ cannot be posted

Users:

- Read and create content in the platform they join
- Different preferences for unsafety (some like hate, some not)
- Like network size, dislike reading content far from their unsafety

⇨ (Technical version)

Model

Content unsafety is given by a metric $\theta \in [0,1]$ (the higher, the unsafer)

Content moderation policy (K): any $\theta > K$ cannot be posted

Users:

- Read and create content in the platform they join
- Different preferences for unsafety (some like hate, some not)
- Like network size, dislike reading content far from their unsafety

2 Asymmetric **Platforms**:

- A **Moderated** one, higher quality platform: **moderates (bans) content**
  - Maximizes revenues from **advertisers** (averse to unsafe content)
- A **Fringe** one, lower quality platform: **no content moderation**

(Technical version)

Model

Content unsafety is given by a metric $\theta \in [0,1]$ (the higher, the unsafer)

Content moderation policy (K): any $\theta > K$ cannot be posted

Users:

- Read and create content in the platform they join
- Different preferences for unsafety (some like hate, some not)
- Like network size, dislike reading content far from their unsafety

2 Asymmetric **Platforms**:

Twitter, Instagram, Facebook

- A **Moderated** one, higher quality platform: **moderates (bans) content**
  - Maximizes revenues from **advertisers** (averse to unsafe content)
- A **Fringe** one, lower quality platform: **no content moderation**

8Chan, Truth, Parler

(Technical version)

Model

Content unsafety is given by a metric $\theta \in [0,1]$ (the higher, the unsafer)

Content moderation policy (K): any $\theta > K$ cannot be posted

Users:

- Read and create content in the platform they join
- Different preferences for unsafety (some like hate, some not)
- Like network size, dislike reading content far from their unsafety

2 Asymmetric **Platforms**:

Twitter, Instagram, Facebook

- A **Moderated** one, higher quality platform: **moderates (bans) content**
  - Maximizes revenues from **advertisers** (averse to unsafe content)
- A **Fringe** one, lower quality platform: **no content moderation**

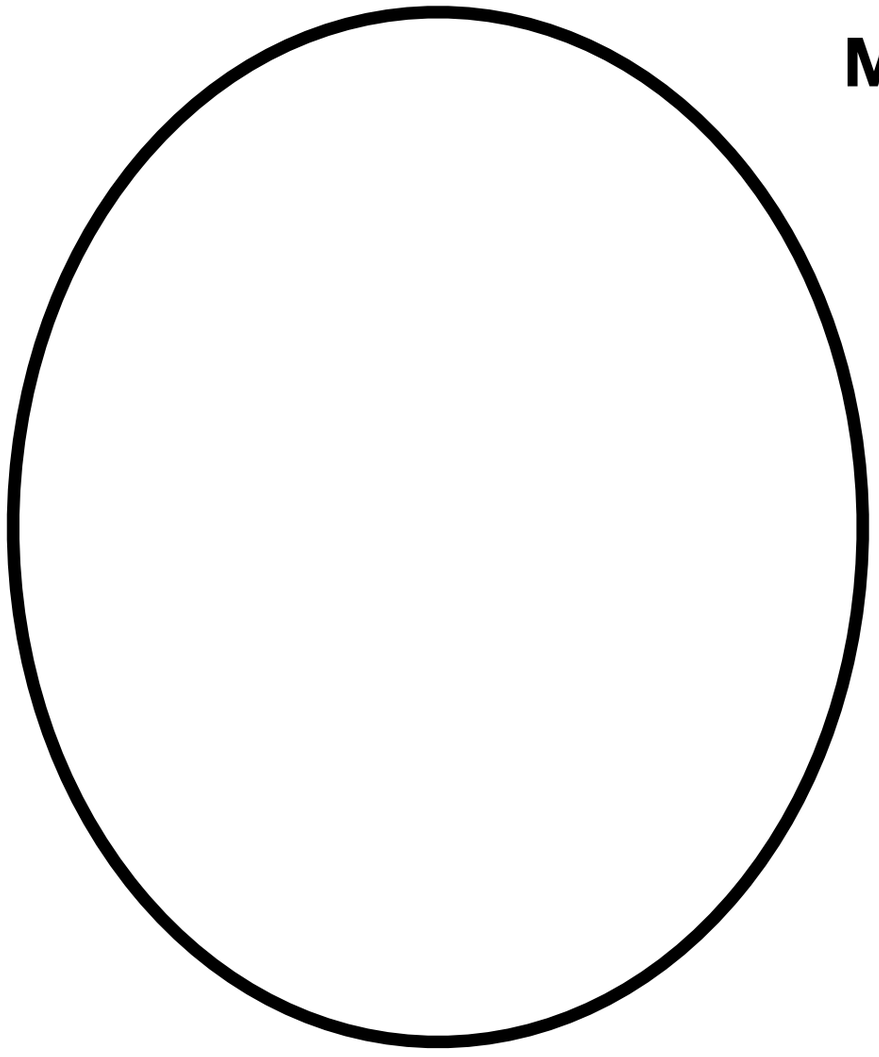
8Chan, Truth, Parler

The **moderated** chooses moderation policy to max size & min hate on it

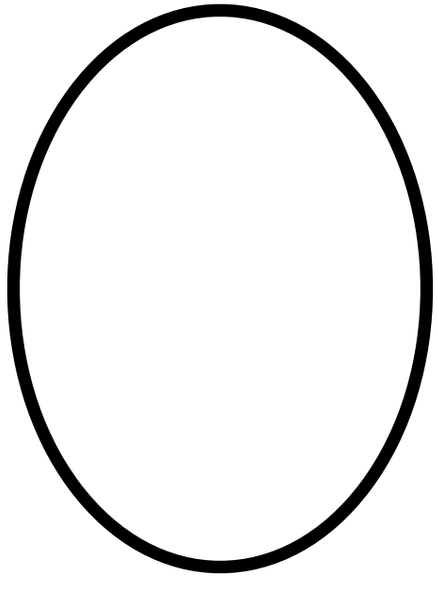
The **fringe** does not do anything

Main Mechanism

User

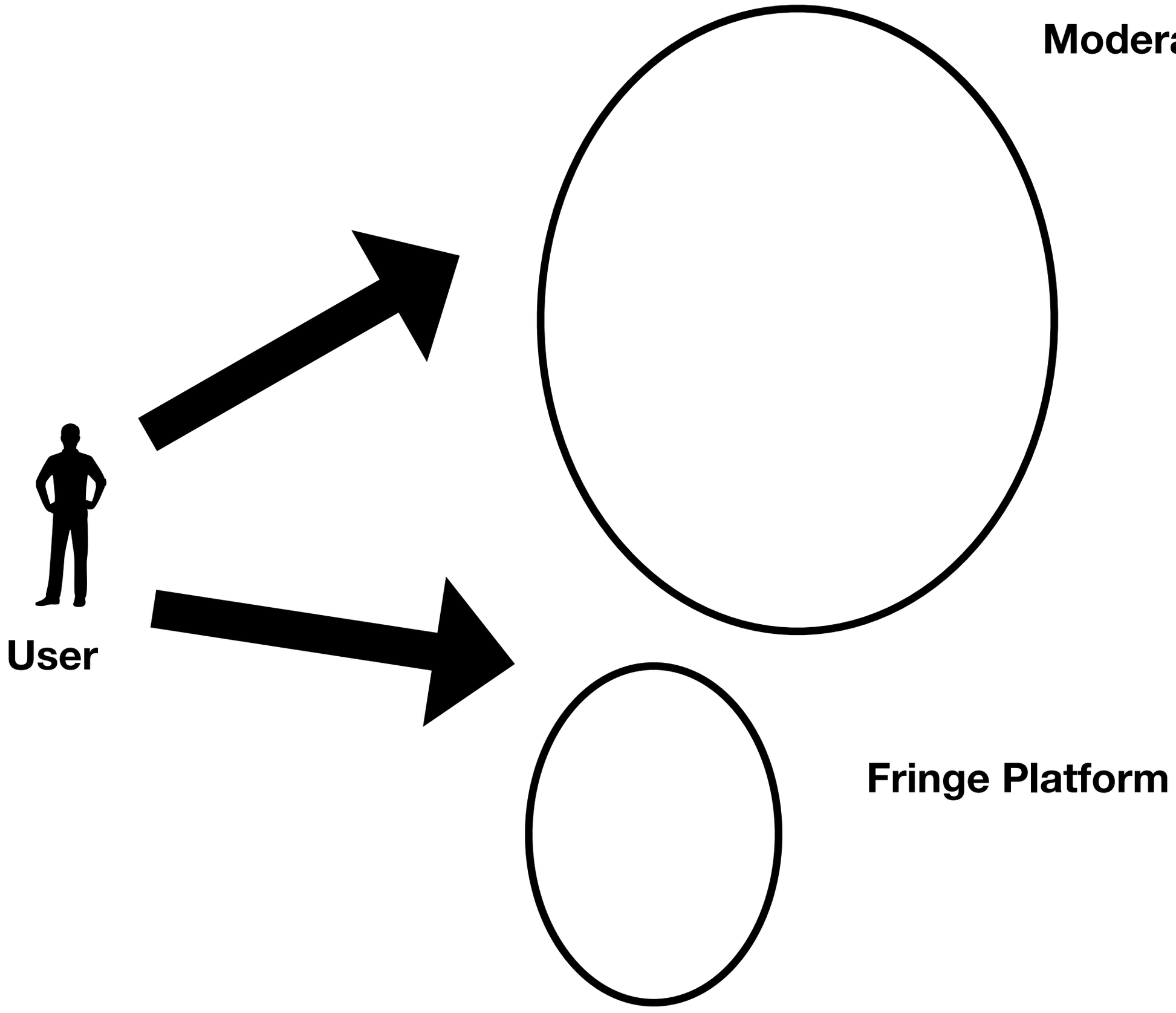


Moderated Platform

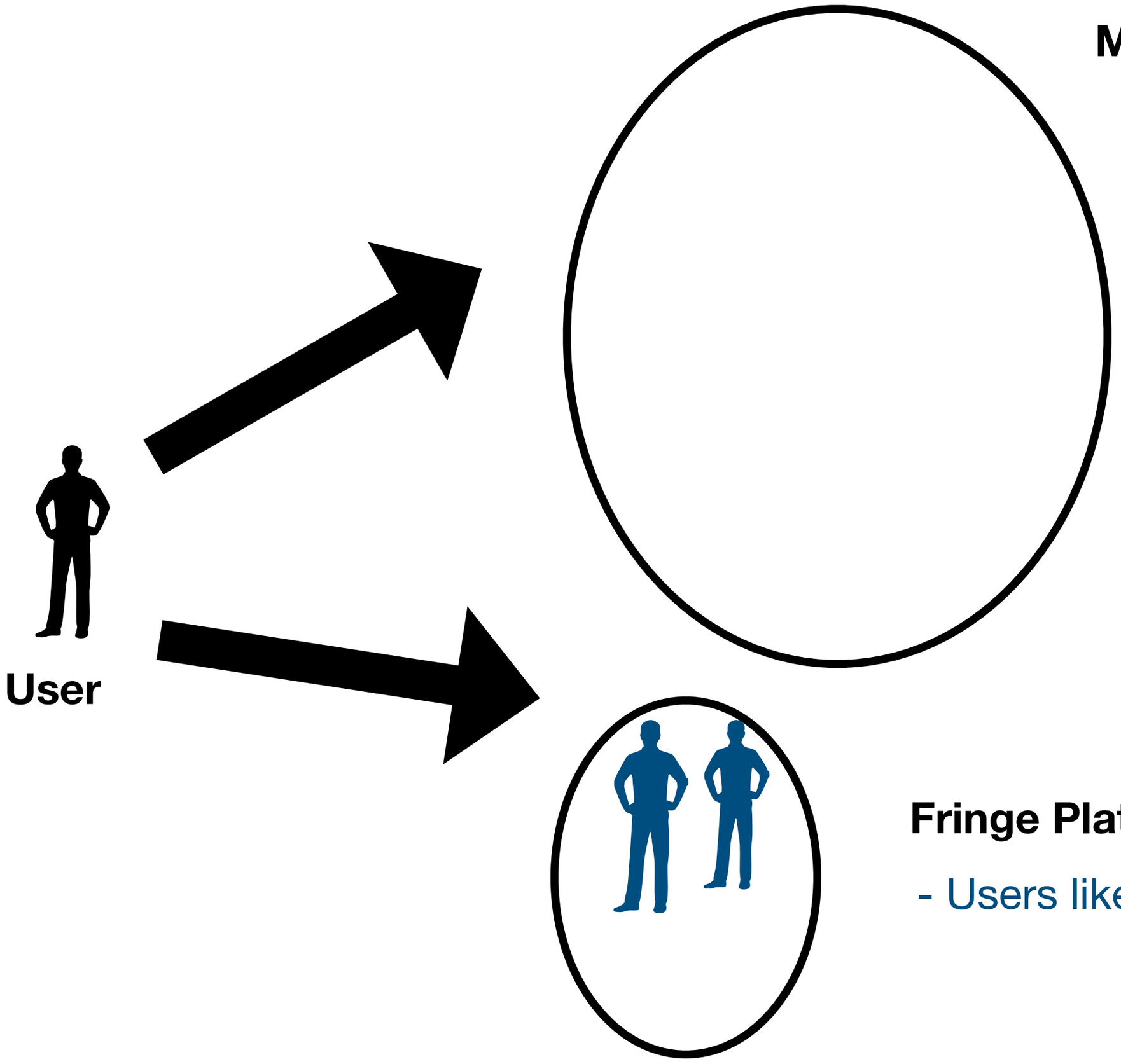


Fringe Platform

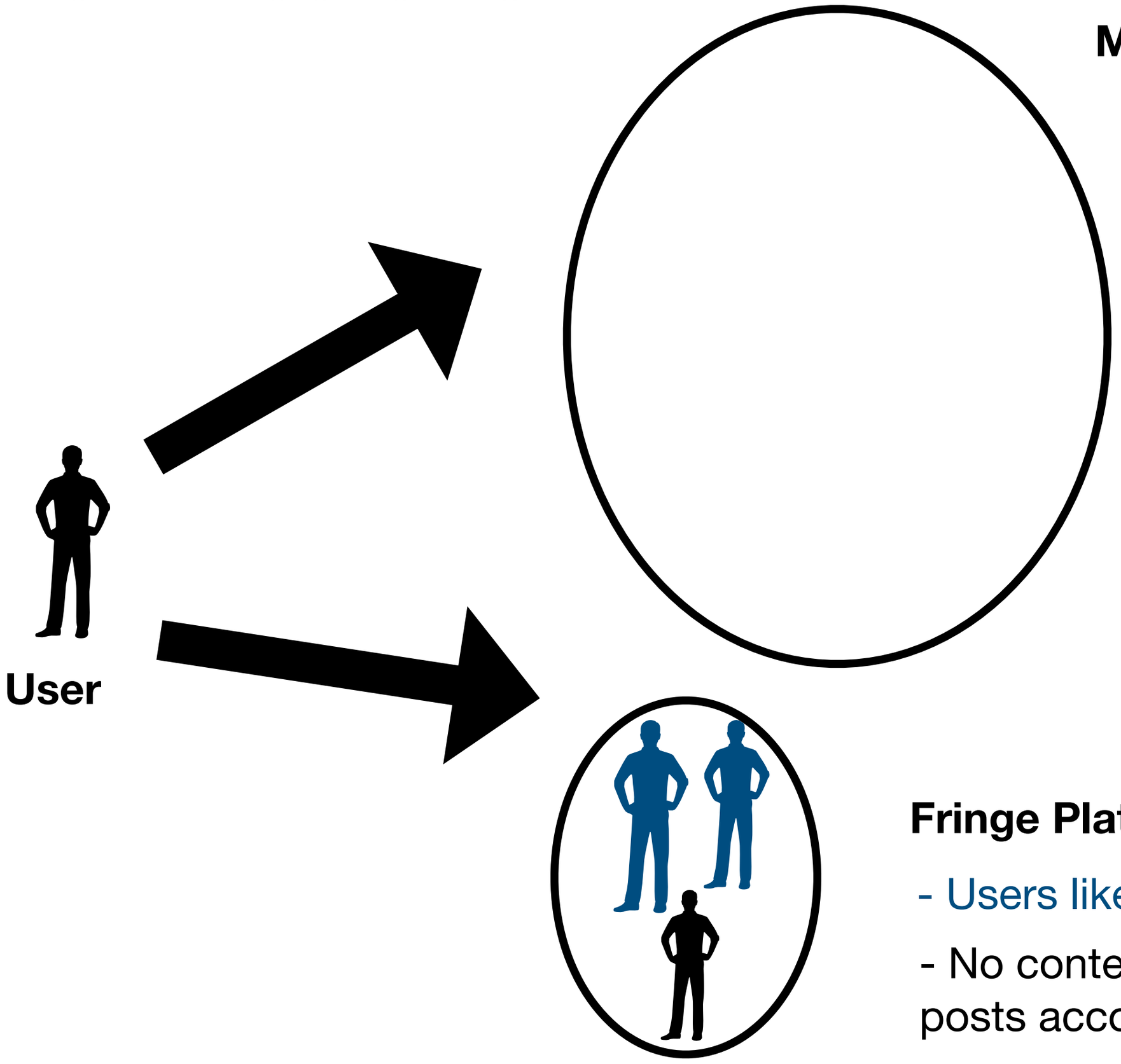
Main Mechanism



Main Mechanism



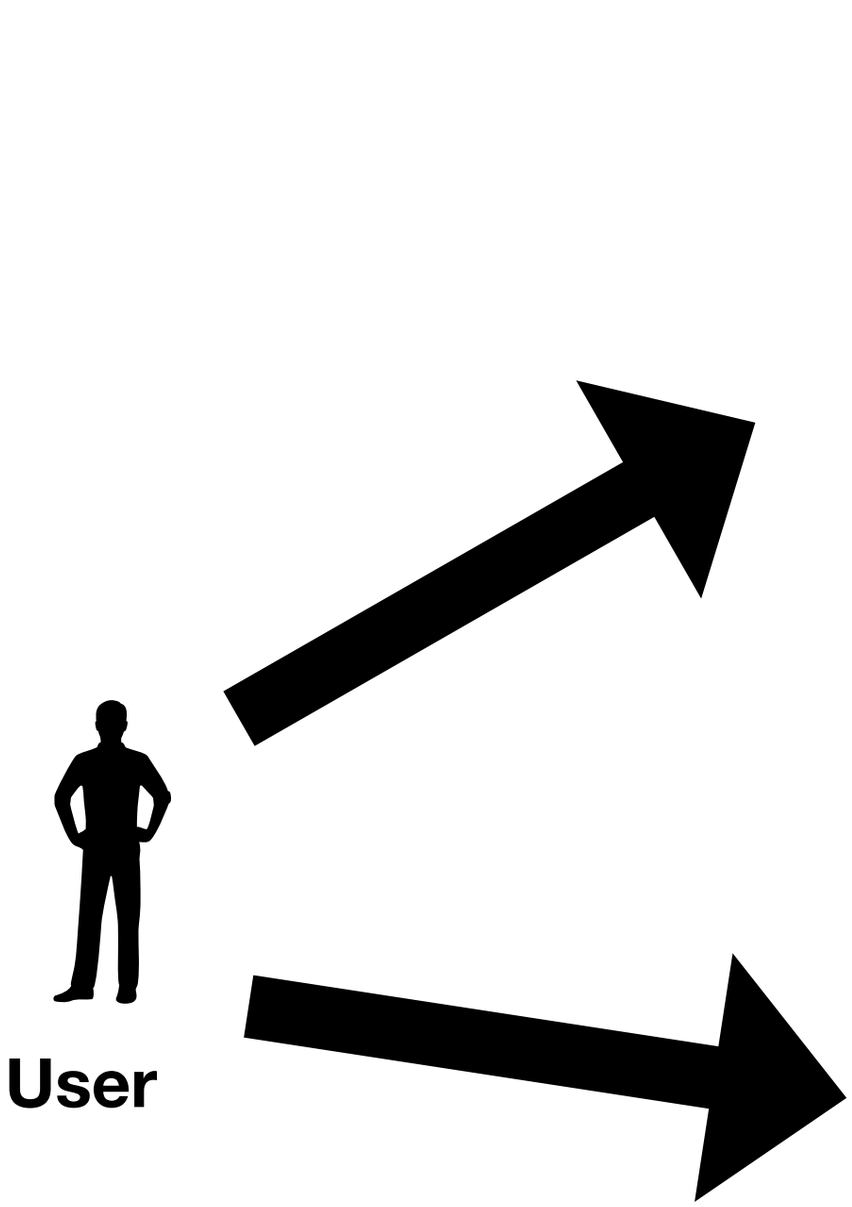
Main Mechanism

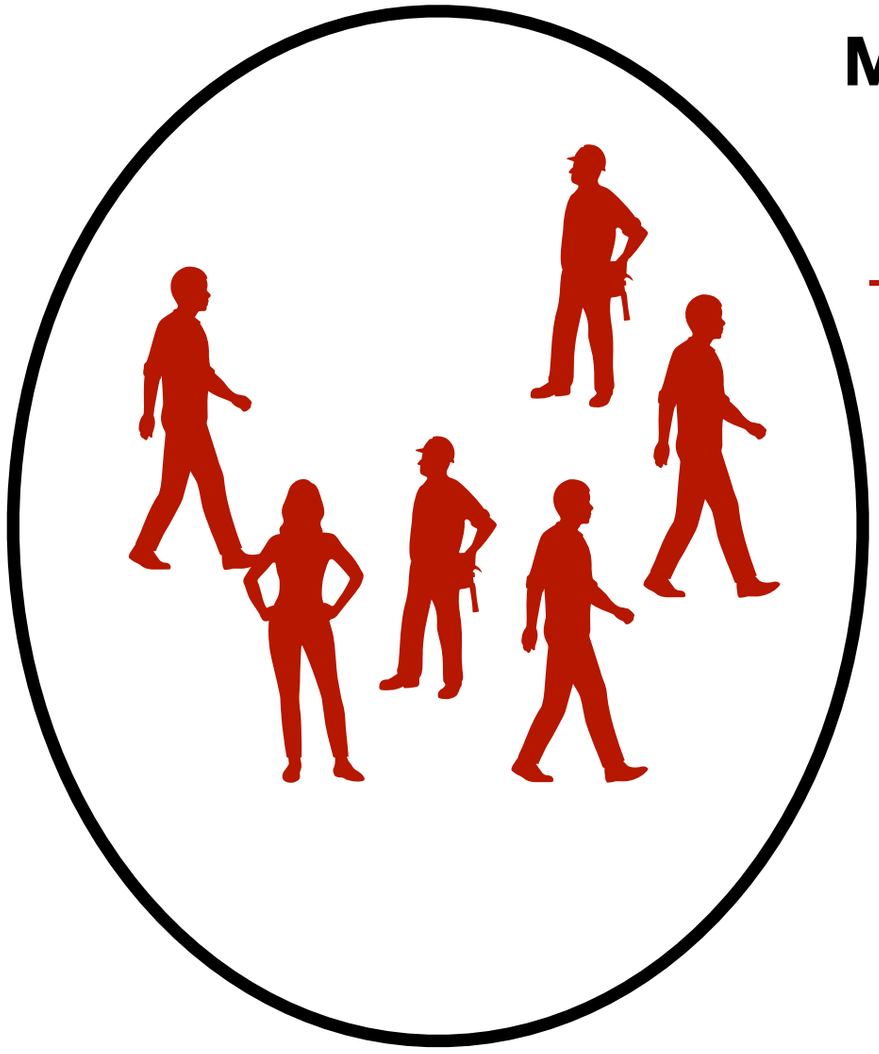


Fringe Platform

- Users like him (in unsafe terms)
- No content moderation:
posts according to his unsafety

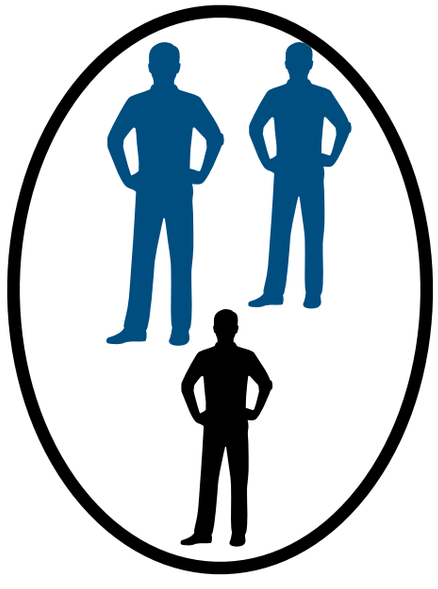
Main Mechanism





Moderated Platform

- More (and safer) users

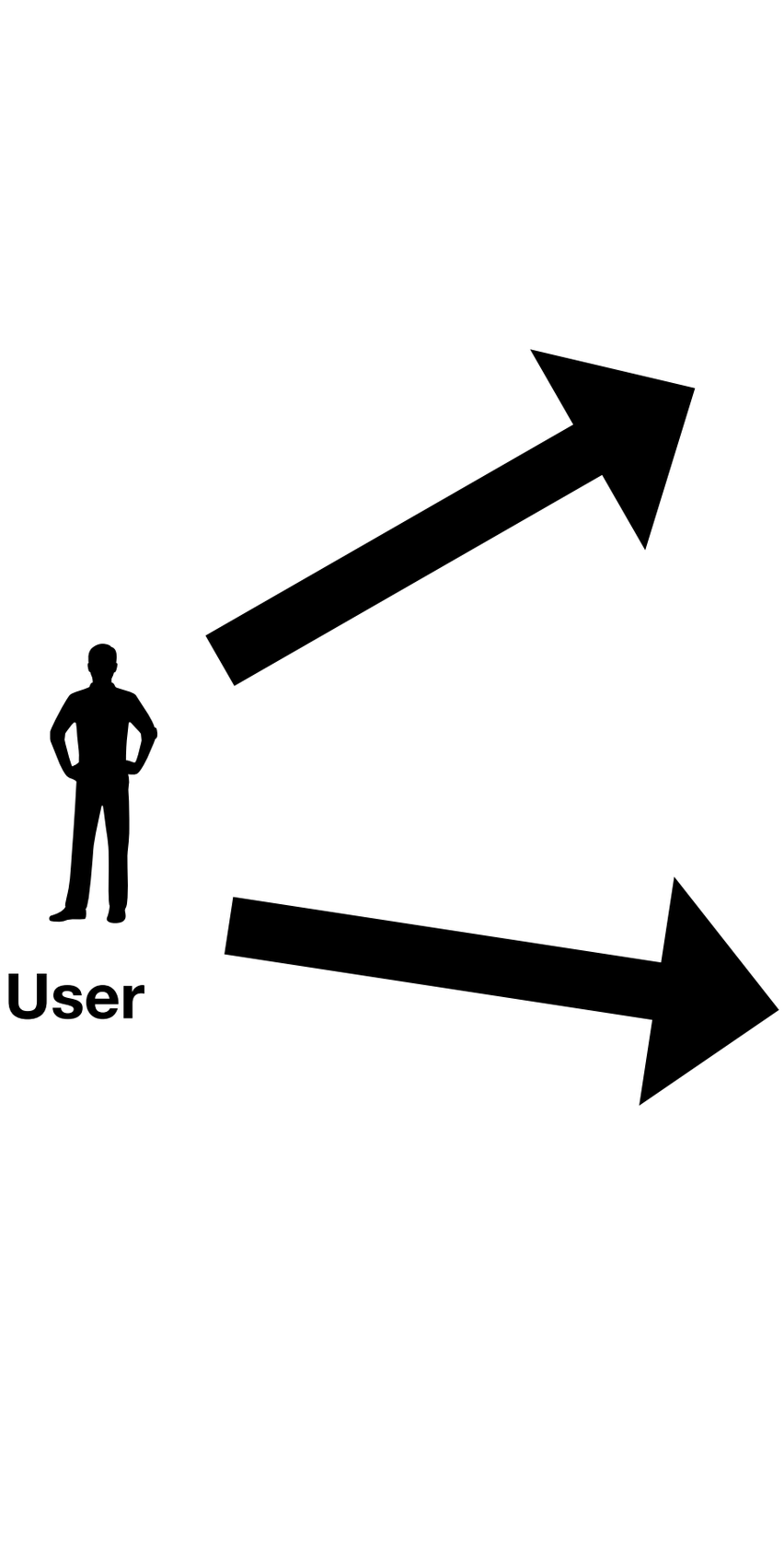


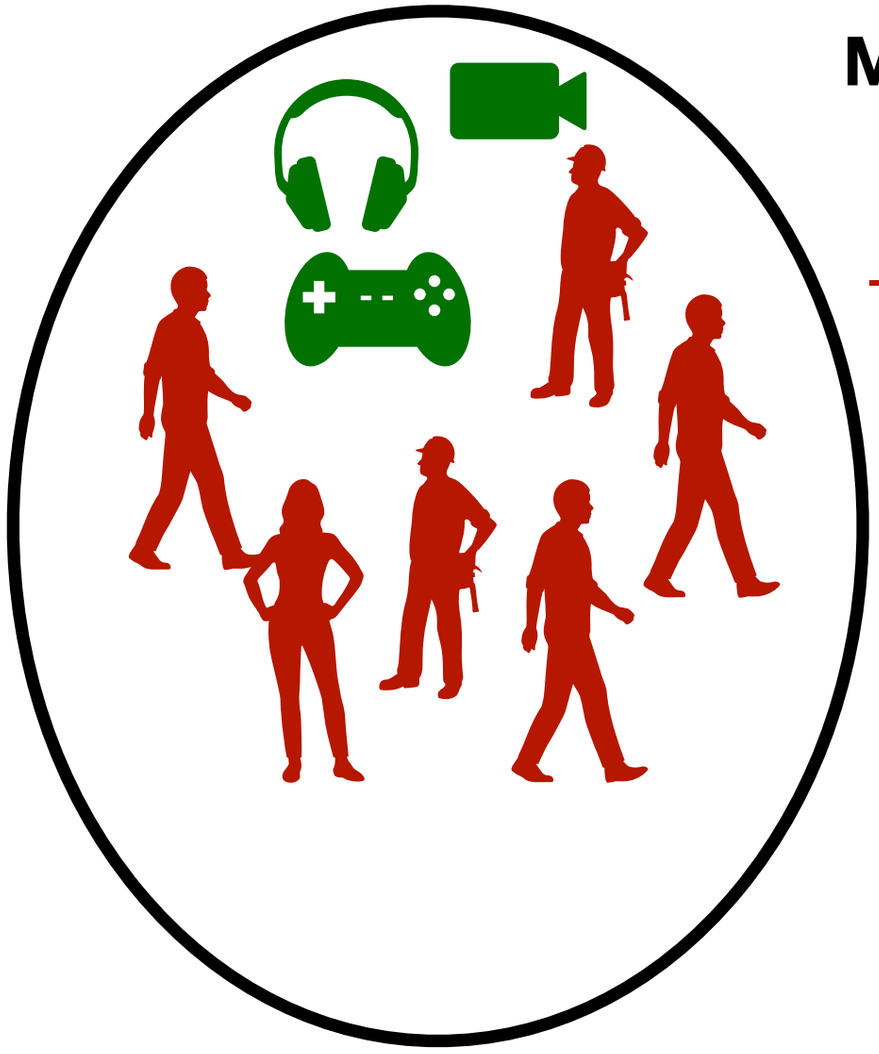
Fringe Platform

- Users like him (in unsafe terms)

- No content moderation:
posts according to his unsafety

Main Mechanism

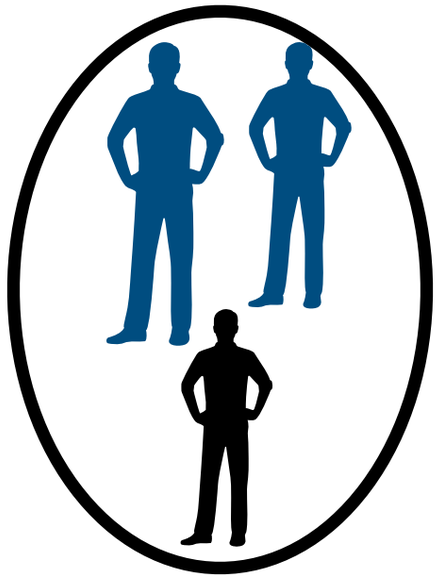




Moderated Platform

- More (and safer) users

- More Features

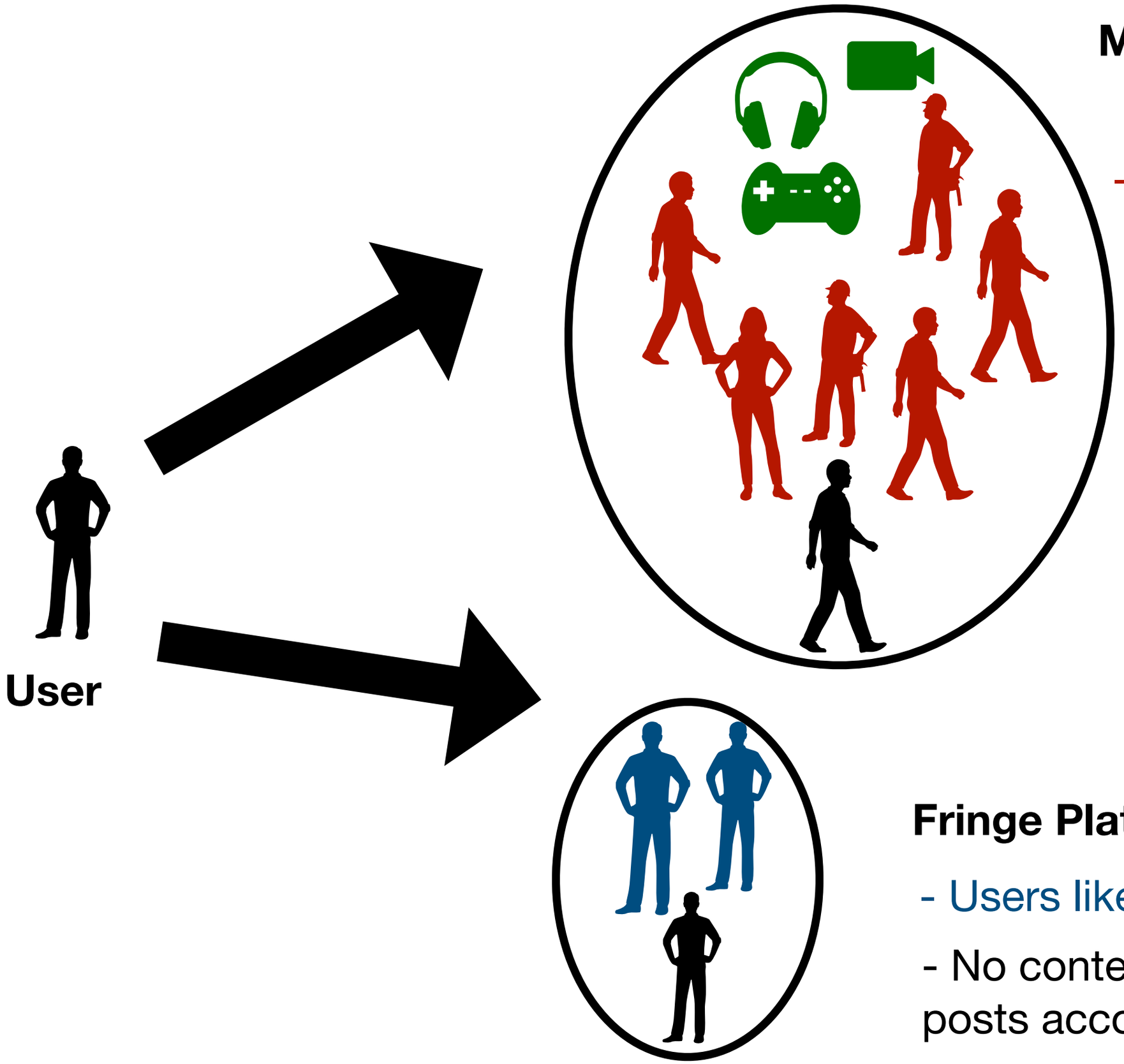


Fringe Platform

- Users like him (in unsafe terms)

- No content moderation:
posts according to his unsafety

Main Mechanism



Moderated Platform

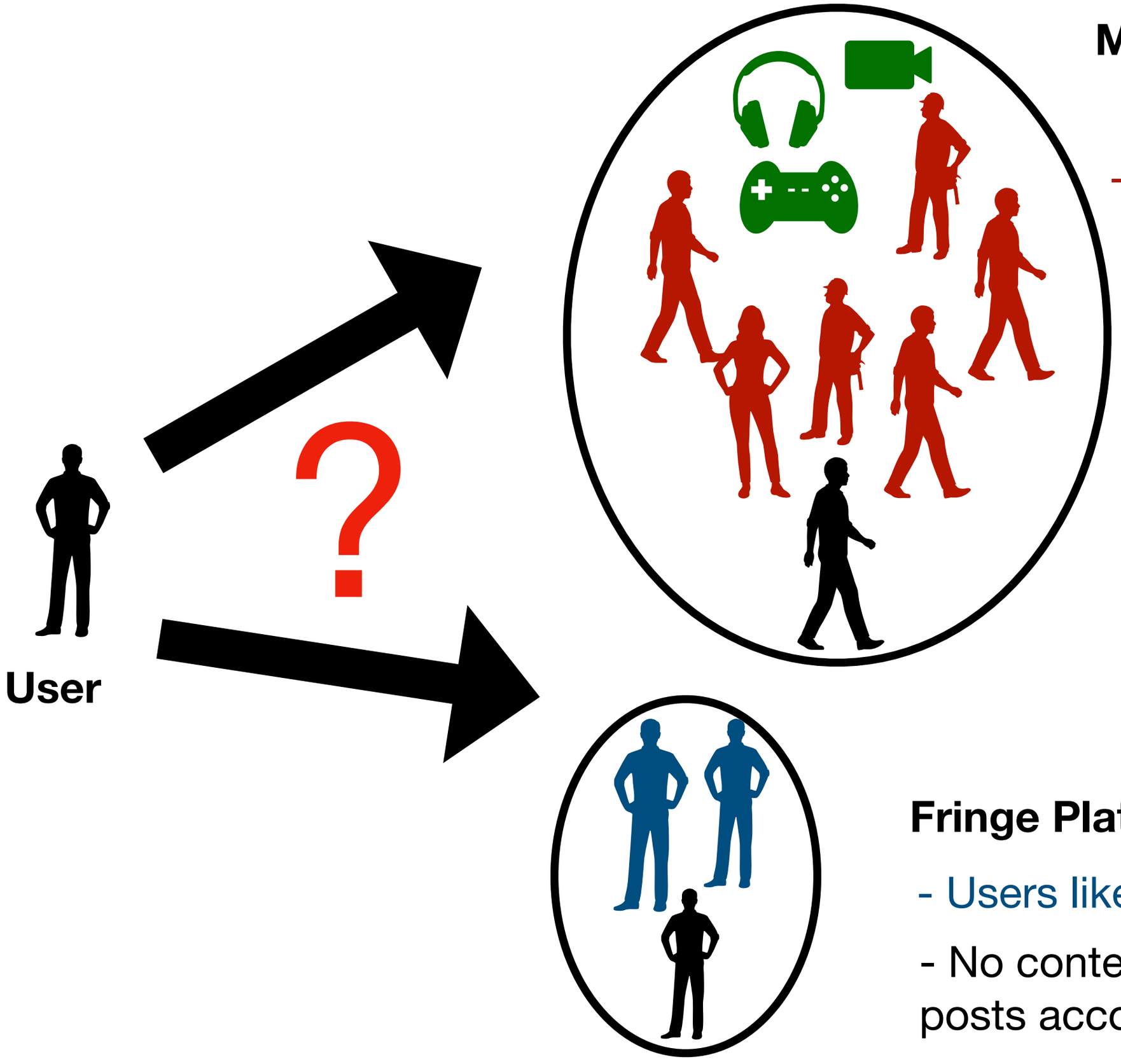
- More (and safer) users
- More Features
- Needs to respect the moderation policy:

self-censors

Fringe Platform

- Users like him (in unsafe terms)
- No content moderation: posts according to his unsafety

Main Mechanism



Moderated Platform

- More (and safer) users
- More Features
- Needs to respect the moderation policy:

self-censors

Fringe Platform

- Users like him (in unsafe terms)
- No content moderation:

posts according to his unsafety

(Technical version)

Characterization of the Equilibrium

Total Content Unsafety is U-shaped in Moderation Level

⇨ (Technical version)

Characterization of the Equilibrium

Total Content Unsafety is U-shaped in Moderation Level

Exists an unique policy that maximizes profits of the platform

Exists an unique policy that minimizes unsafety level

(Technical version)

Characterization of the Equilibrium

Total Content Unsafety is U-shaped in Moderation Level

Exists an unique policy that maximizes profits of the platform

Exists an unique policy that minimizes unsafety level

Comparative statics:

I)

II)

III)

(Technical version)

Characterization of the Equilibrium

Total Content Unsafety is U-shaped in Moderation Level

Exists an unique policy that maximizes profits of the platform

Exists an unique policy that minimizes unsafety level

Comparative statics:

I) Incentives to moderate increase with the strength of network effects

…both for the platform and the regulator

II)

III)

(Technical version)

Characterization of the Equilibrium

Total Content Unsafety is U-shaped in Moderation Level

Exists an unique policy that maximizes profits of the platform

Exists an unique policy that minimizes unsafety level

Comparative statics:

I) Incentives to moderate increase with the strength of network effects
…both for the platform and the regulator

II) **But** they increase more for the regulator than the platform

III)

(Technical version)

Characterization of the Equilibrium

Total Content Unsafety is U-shaped in Moderation Level

Exists an unique policy that maximizes profits of the platform

Exists an unique policy that minimizes unsafety level

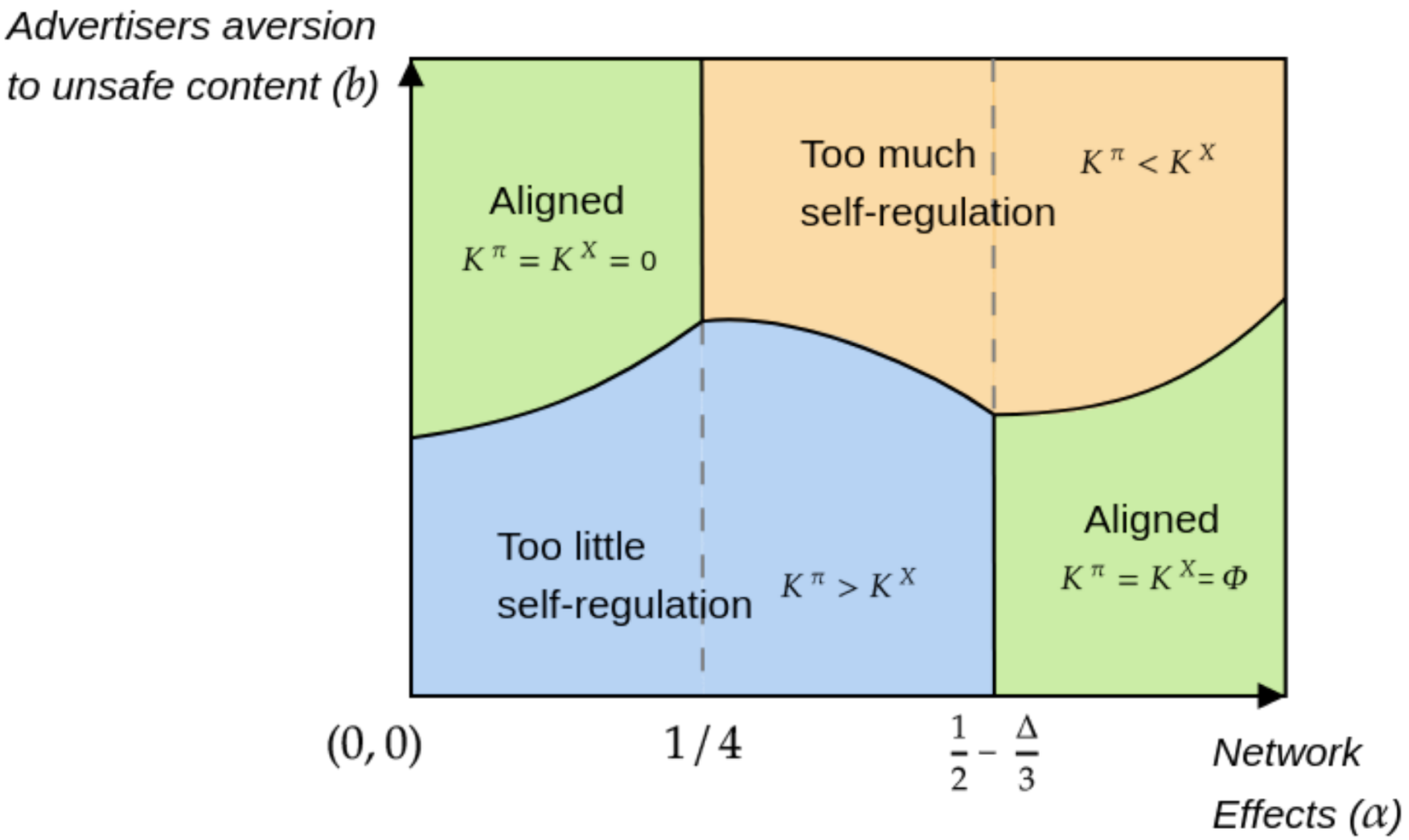
Comparative statics:

I) Incentives to moderate increase with the strength of network effects
…both for the platform and the regulator

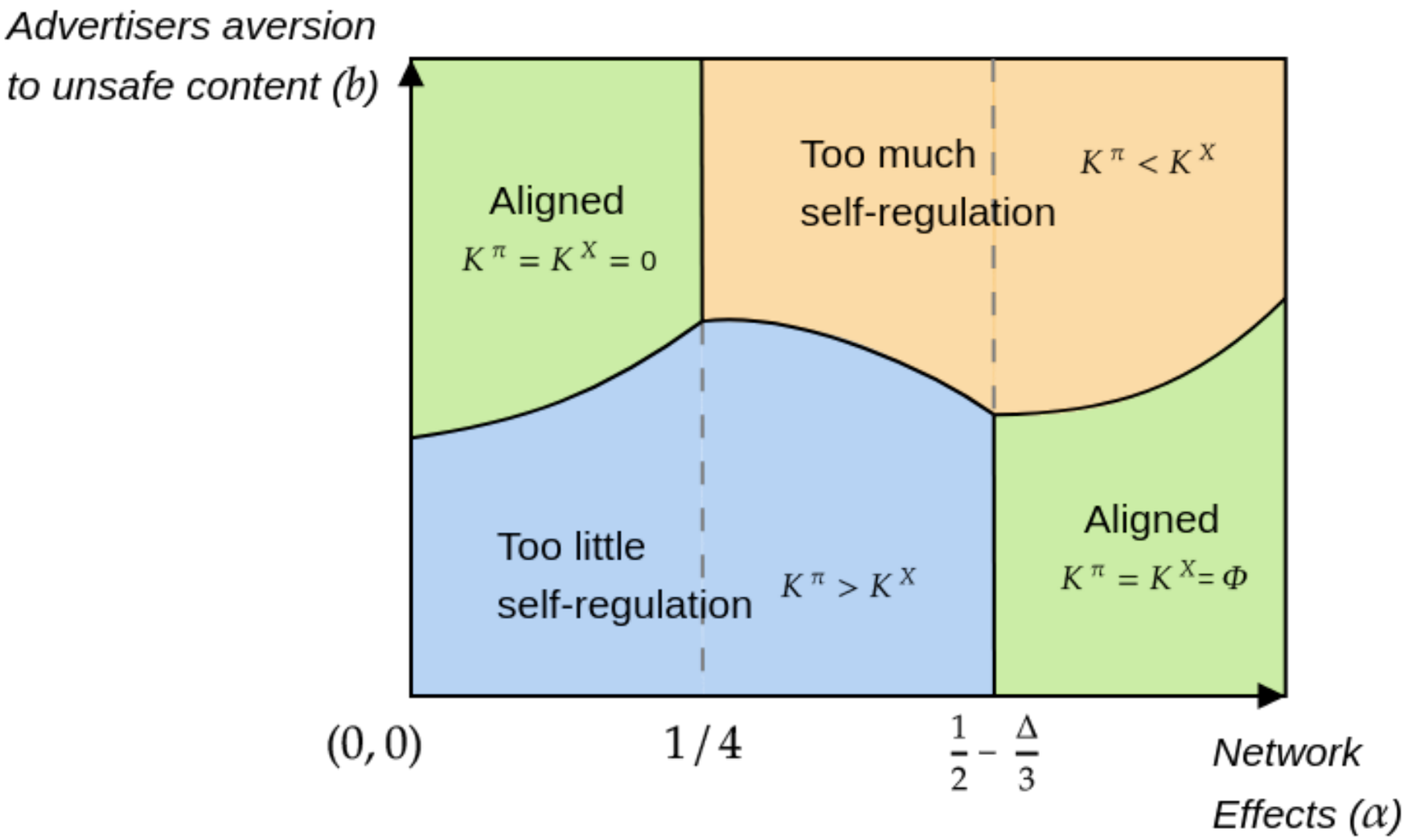
II) **But** they increase more for the regulator than the platform

III) The **lower** the **competition,** the **more** the platform wants to moderate
the **less** the regulator wants to moderate

Policy (imposing a minimal moderation policy)

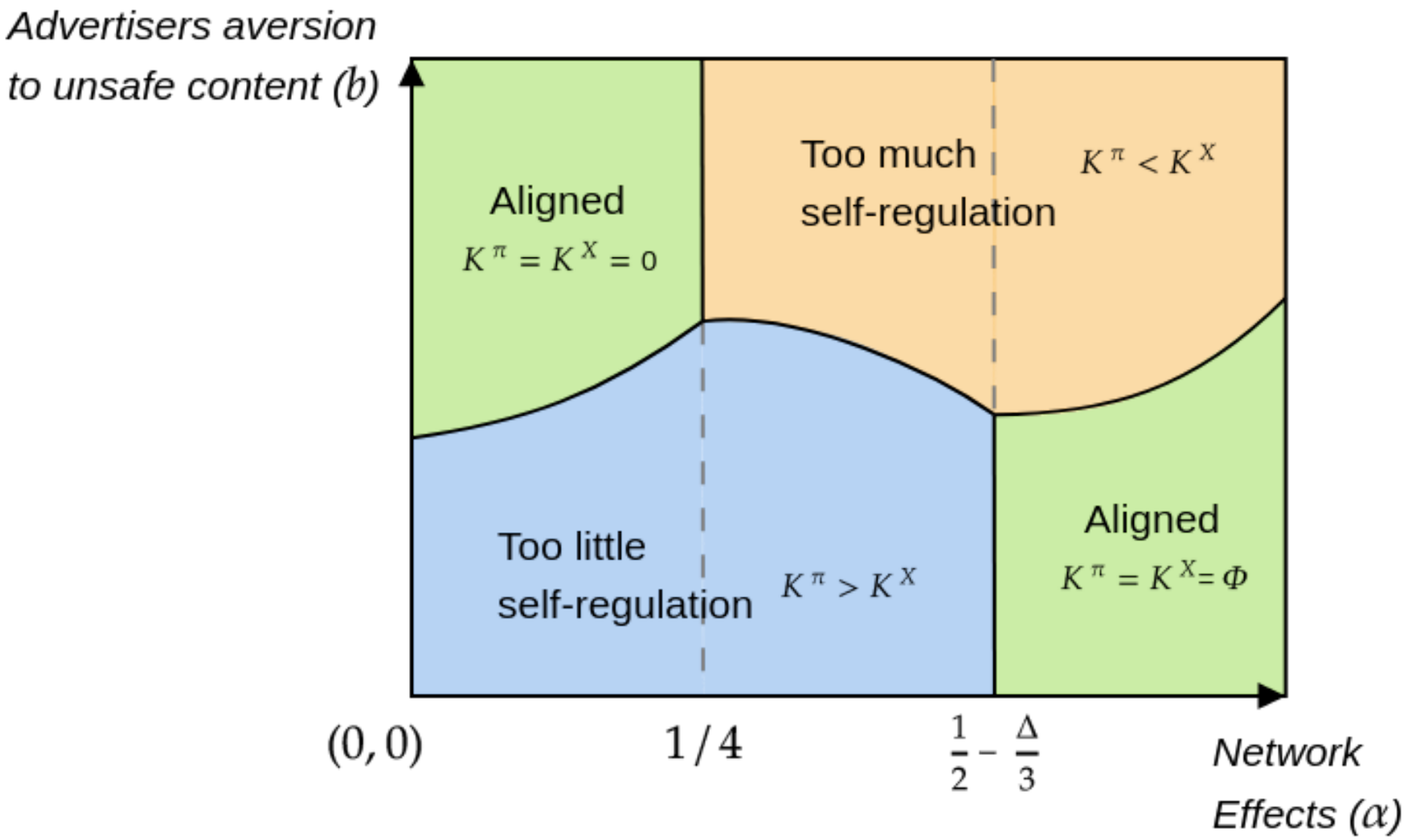


Policy (imposing a minimal moderation policy)



Blue Area:
Beneficial for the regulator to impose a minimal moderation policy

Policy (imposing a minimal moderation policy)



Blue Area:
Beneficial for the regulator to impose a minimal moderation policy

Orange Area: such a policy wouldn't bind. Regulators would like to impose a maximal moderation policy to attract users from the fringe platform.

Conclusion

Main takeaways:

- Potential **migration** reshapes the economic incentives of the agents
- Minimal content moderation policy ONLY if the outside option is bad enough
    - Partly due (thanks) to network effects (what I explored here)

Conclusion

Main takeaways:

- Potential **migration** reshapes the economic incentives of the agents
- Minimal content moderation policy ONLY if the outside option is bad enough
    - Partly due (thanks) to network effects (what I explored here)

(In the paper)

- **Extensions:** Multihoming, Offline Violence, 3 platforms

Conclusion

Main takeaways:

- Potential **migration** reshapes the economic incentives of the agents
- Minimal content moderation policy ONLY if the outside option is bad enough
    - Partly due (thanks) to network effects (what I explored here)

(In the paper)

- **Extensions:** Multihoming, Offline Violence, 3 platforms

Working on…

- Empirical, **structural**, project (to run counterfactuals)
- Other *non-IO* applications

    Cancel culture (~Tirole's safe spaces)

Thanks!

ivan.rendo@tse-fr.eu

Model (Technical)

Back

- A unit mass of **users**, heterogeneous in their preferences for unsafe content: $\theta_i \sim U(0,1)$. High θ = Unsafe content

- 2 **platforms** $j = 1,2$
  - with K_j = **max unsafety level allowed** $(K_2 = 1)$

- User i in platform j **creates** 1 piece of content of type θ_i^C

$$\theta_i^C = \min\{\theta_i, K_j\}$$

- User i in platform j **reads** a random sample of the content, of avg type $\bar{\theta}_j$

$$\bar{\theta}_j = \int_{i \in j} \theta_i^C \mathrm{d}i$$

= average type of content in platform j

- Platform 1, **moderated**, is intrinsically better than 2, **unmoderated**
- Utilities of user i joining $j = 1,2$ are defined as:

Users in the Platform

Average "Unsafety" of the Content

$$U_1(\theta_i) = \alpha N_1 - |\theta_i - \bar{\theta}_1| + \Delta$$

$$U_2(\theta_i) = \alpha N_2 - |\theta_i - \bar{\theta}_2|$$

Quality Premium of the Moderated

Strength of network effects

Users single-home

Rk: No outside option!

Advertisers

Buy a fixed amount of ads in the **moderated** platform (1)

Are **averse** to unsafe content

Price of ads: $1 - b\bar{\theta}_1$

Moderated Platform

- Platform (1) chooses a **content moderation policy**

$K \in [0,1]$: perfectly and costlessly **bans any content** $\theta_i > K$

$$\Pi(K) = N_1(K) \times (1 - b\bar{\theta}_1(K))$$

Advertisers aversion to unsafe content

Average content unsafety

users in platform

Price of ads

…platform (2) just exists with $K_2 = 1$

Timing

1. Platform (1) chooses K

2. Users choose which platform to join. I focus on threshold equilibria

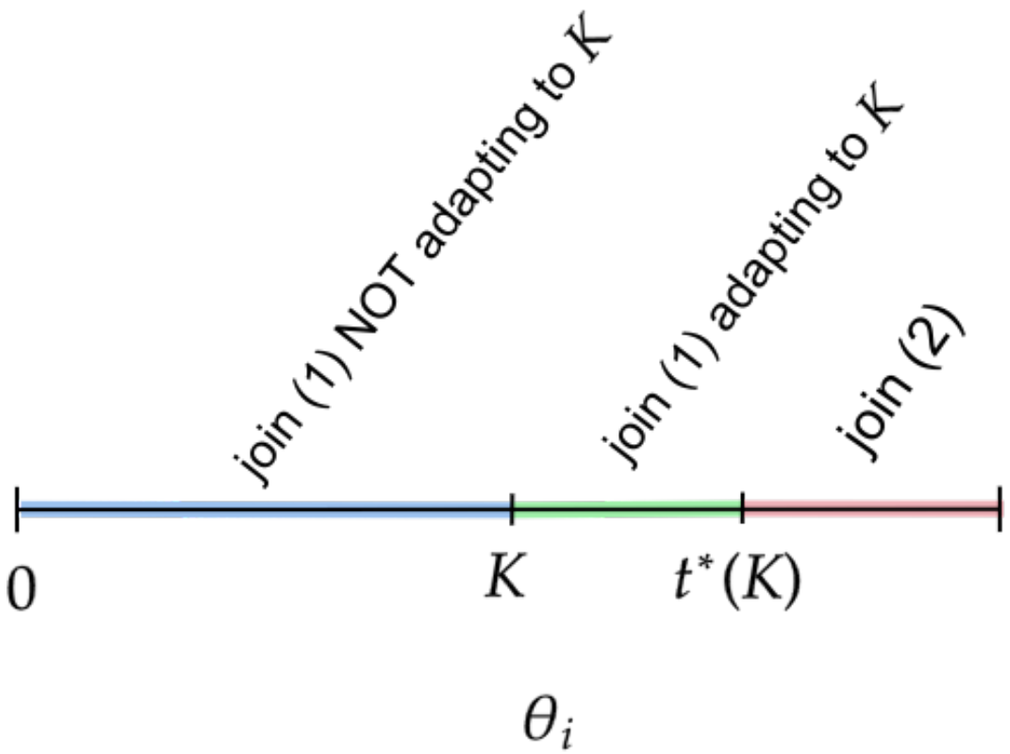
3. Profits and payoffs are realized

Threshold Equilibrium (subgame for given K)

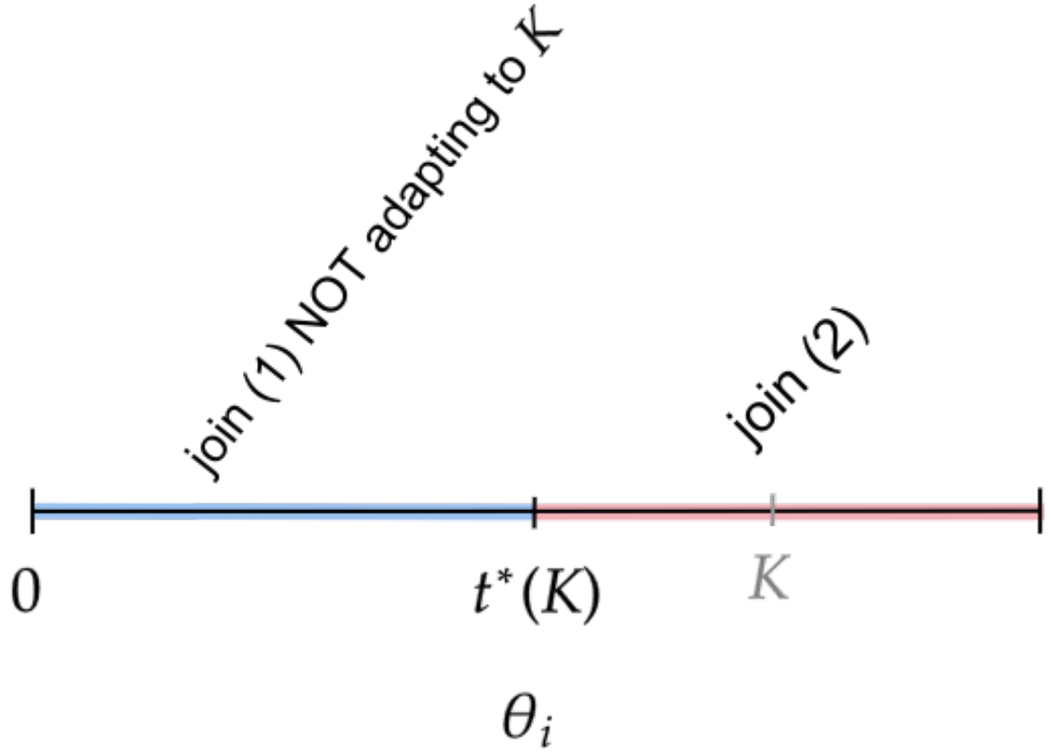
(Assumed) User i joins platform (1) iff $\theta_i < t^*$, otherwise, they join (2)

Under some assumptions on α, Δ; and given K,
there exist a **unique** threshold **equilibrium**

Low K (strict policy)



High K (lenient policy)



Characterization of the Equilibrium

Excluding corner solutions:

$\exists!$ $K^{\pi}(\alpha, \Delta)$ maximizing **profits** of the firm

$\exists!$ $K^{X}(\alpha, \Delta)$ minimizing total **unsafety**

Comparative statics:

I) $$\frac{\mathrm{d}}{\mathrm{d}\alpha} K^{X}(\alpha, \Delta) > \frac{\mathrm{d}}{\mathrm{d}\alpha} K^{\pi}(\alpha, \Delta, b) > 0$$ Policy! (next slide)

II) $$\frac{\mathrm{d}}{\mathrm{d}\Delta} K^{\pi}(\alpha, \Delta, b) < 0 \qquad \frac{\mathrm{d}}{\mathrm{d}\Delta} K^{X}(\alpha, \Delta) > 0$$