Content Moderation in Presence of Fringe Platforms

Iván Rendo (TSE)

3rd Year Paper Presentation

Toulouse, 5 July 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 may 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 may 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)

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 may 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)

Broad question: consequences of **content moderation**?

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

Research Questions:

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

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

➡ What incentives do platforms have to self-regulate?

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

Main Features of the Model

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

A **Content Moderation** (K) bans any content $\theta > K$

Main Features of the Model

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

A **Content Moderation** (K) bans any content $\theta > K$

Users:

- Create + view content on platforms
- Common preferences for network size + quality of the platform
- **Heterogeneous preferences for unsafe content**

Main Features of the Model

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

A **Content Moderation** (K) bans any content $\theta > K$

Users:

- Create + view content on platforms
- Common preferences for network size + quality of the platform
- **Heterogeneous preferences for unsafe content**

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**

Main Features of the Model

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

A **Content Moderation** (K) bans any content $\theta > K$

Users:

- Create + view content on platforms
- Common preferences for network size + quality of the platform
- **Heterogeneous preferences for unsafe content**

2 Asymmetric **Platforms**:

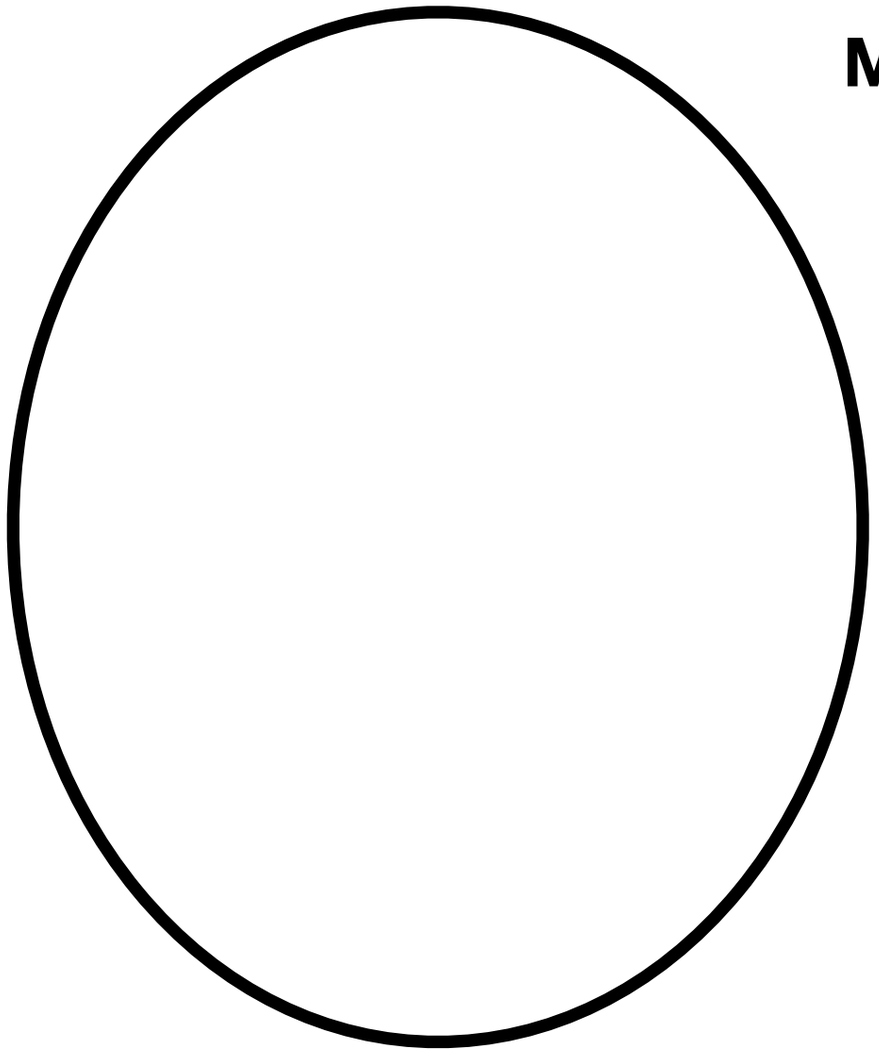
Twitter, Instagram, Facebook

- 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**

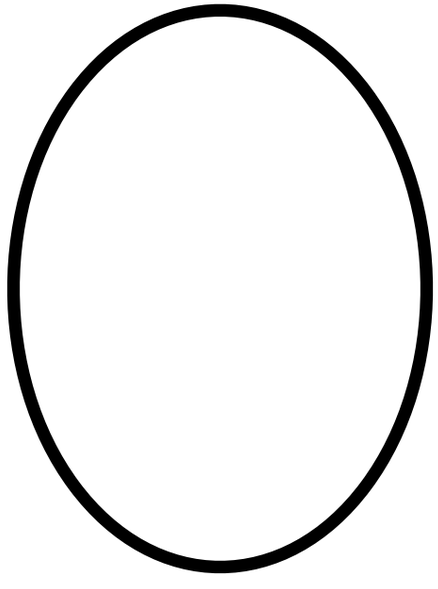
8Chan, Truth, Parler

Main User's Trade-off

Unsafe User

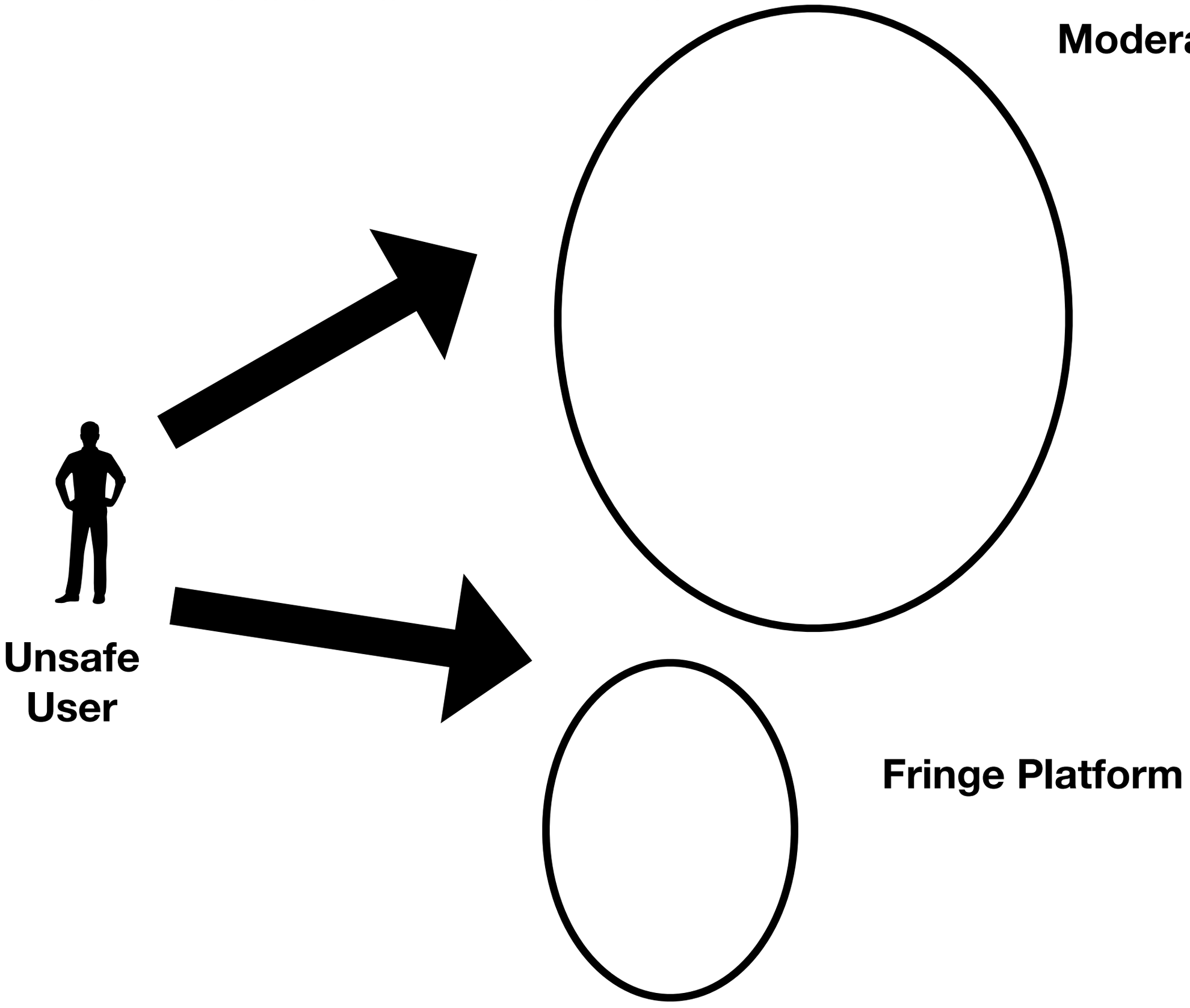


Moderated Platform



Fringe Platform

Main User's Trade-off



Main User's Trade-off

Moderated Platform

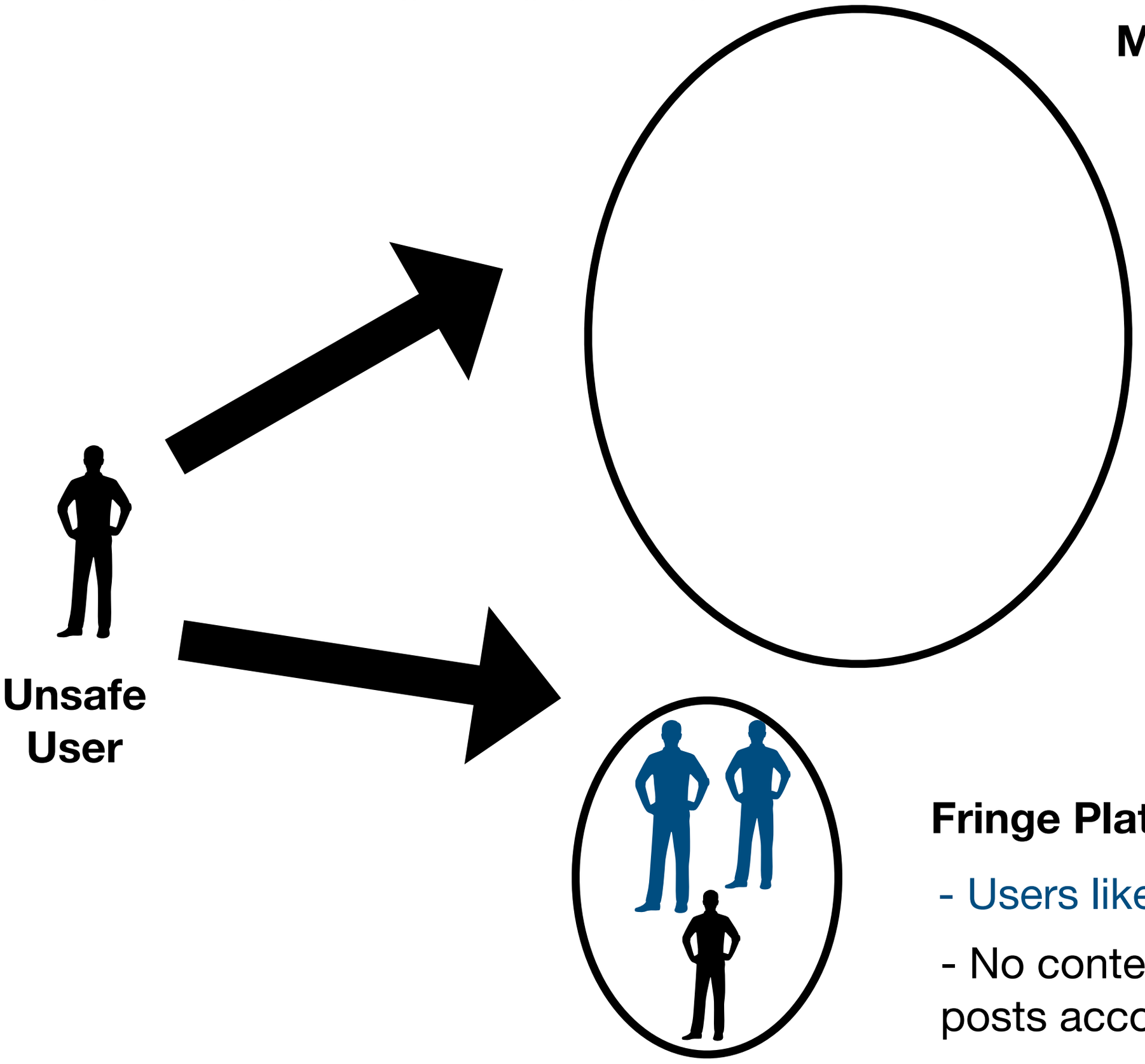
Unsafe User

Fringe Platform

- Users like him (in unsafe terms)

Main User's Trade-off

Moderated Platform

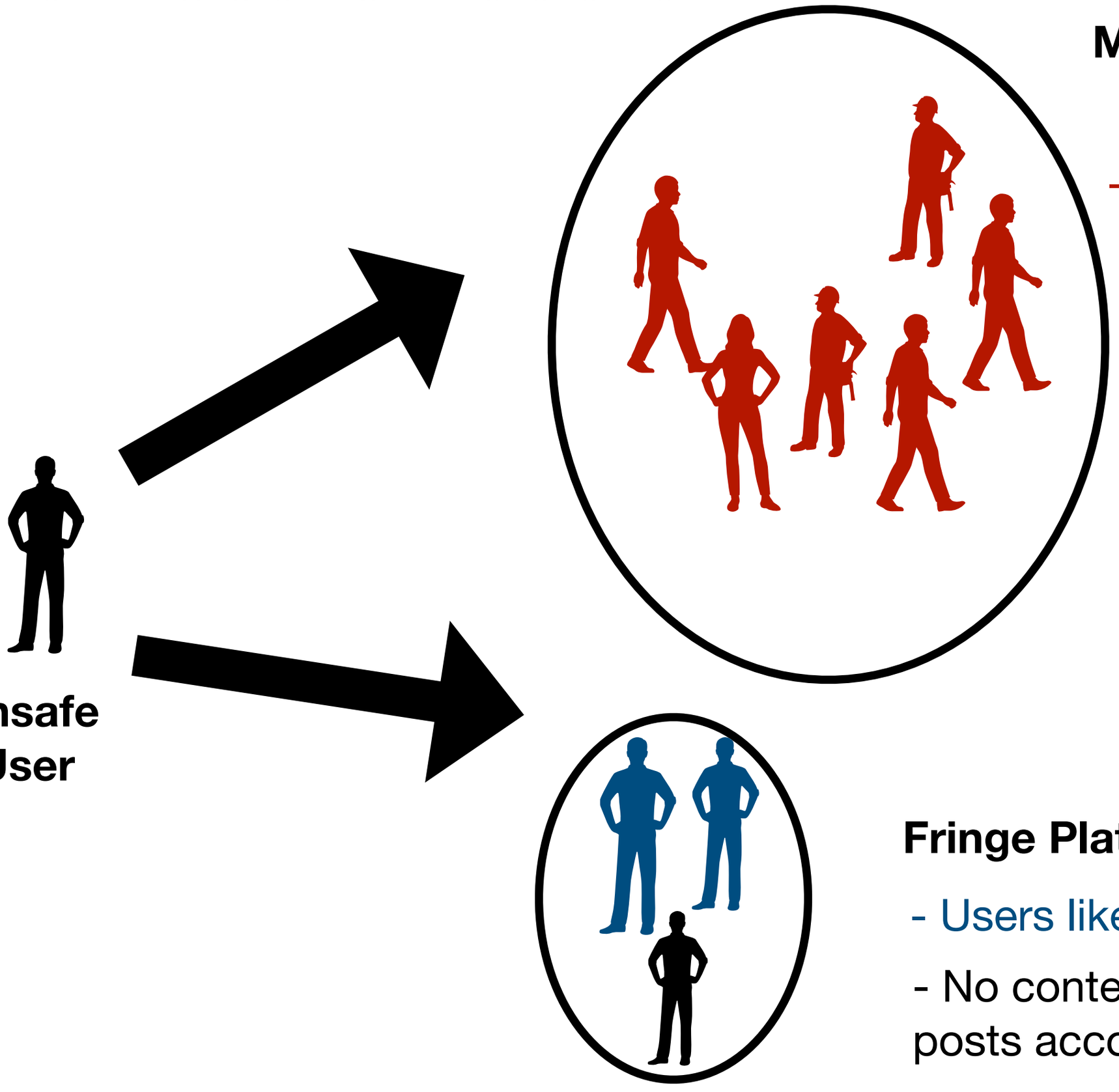


Unsafe User

Fringe Platform

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

Main User's Trade-off



Unsafe User

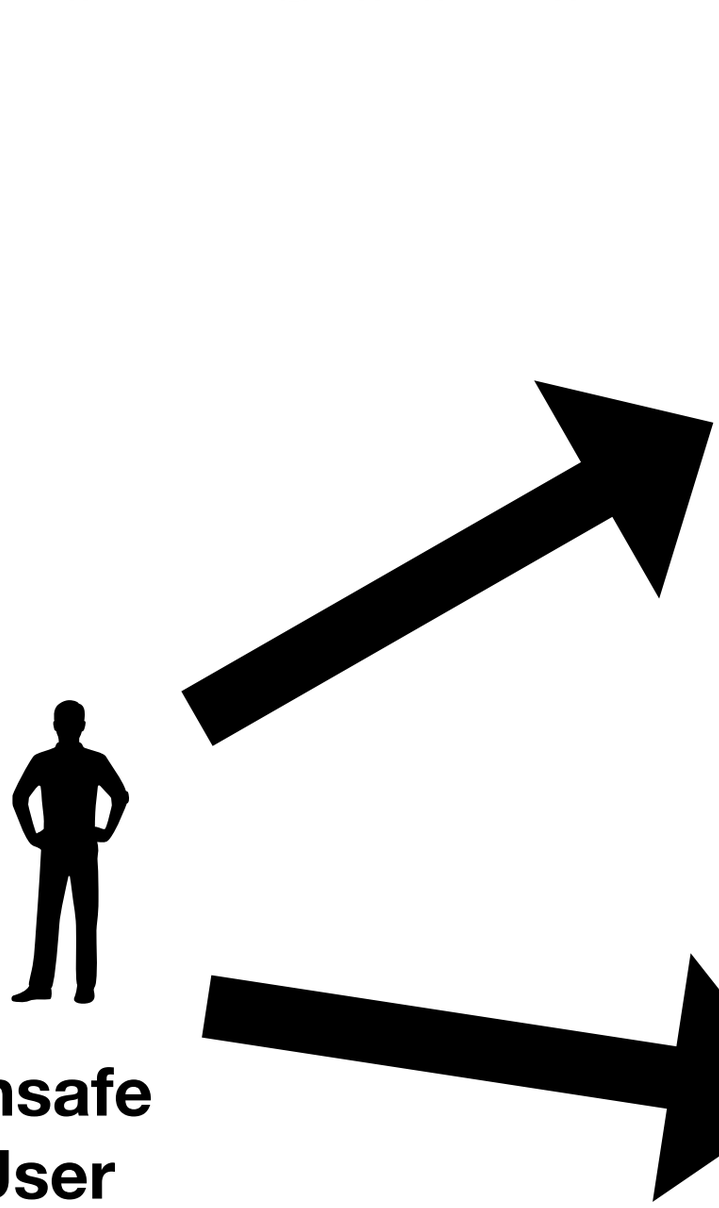
Moderated Platform

- More (and safer) users

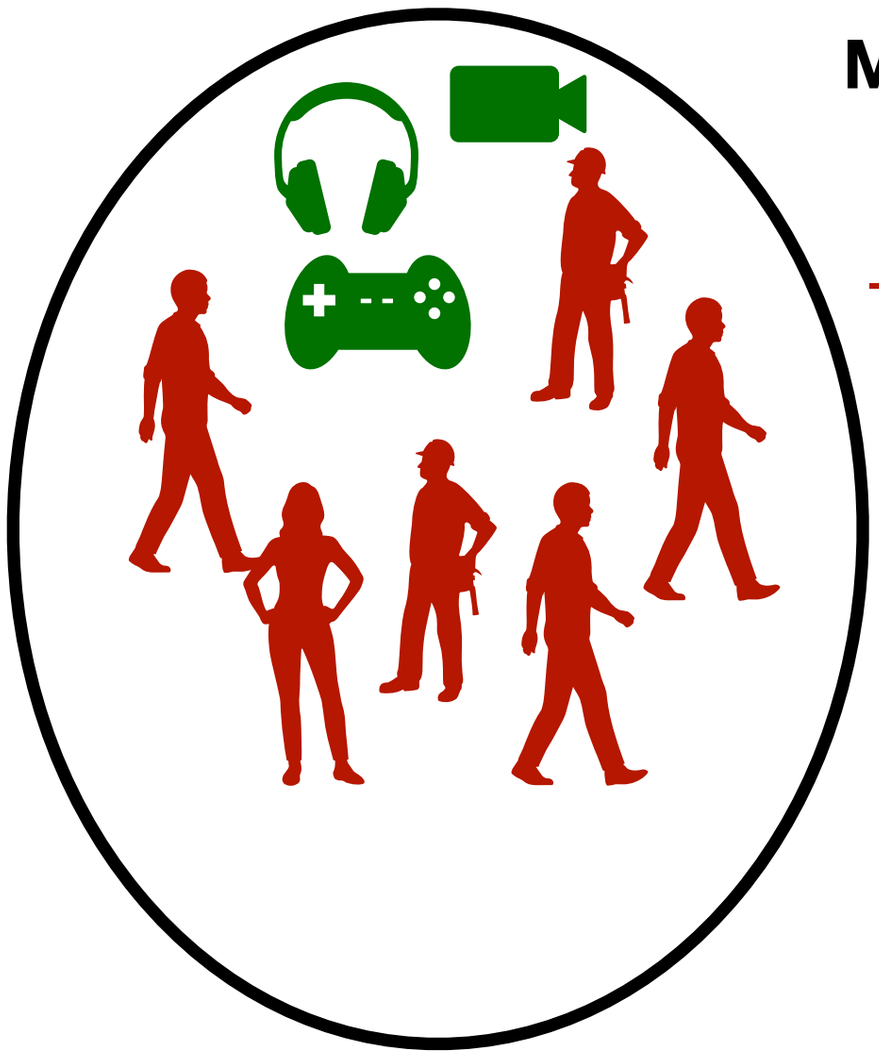
Fringe Platform

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

Main User's Trade-off



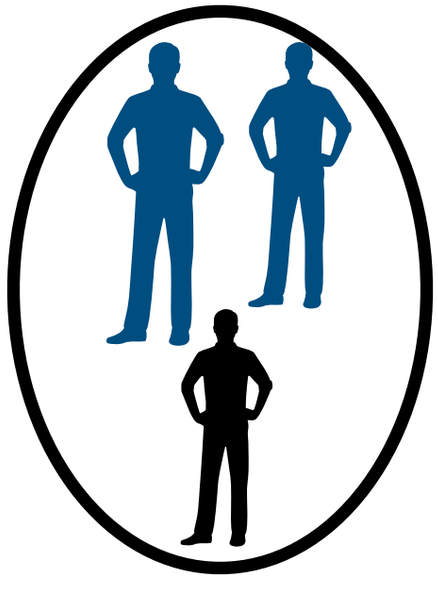
Unsafe User



Moderated Platform

- More (and safer) users

- More Features

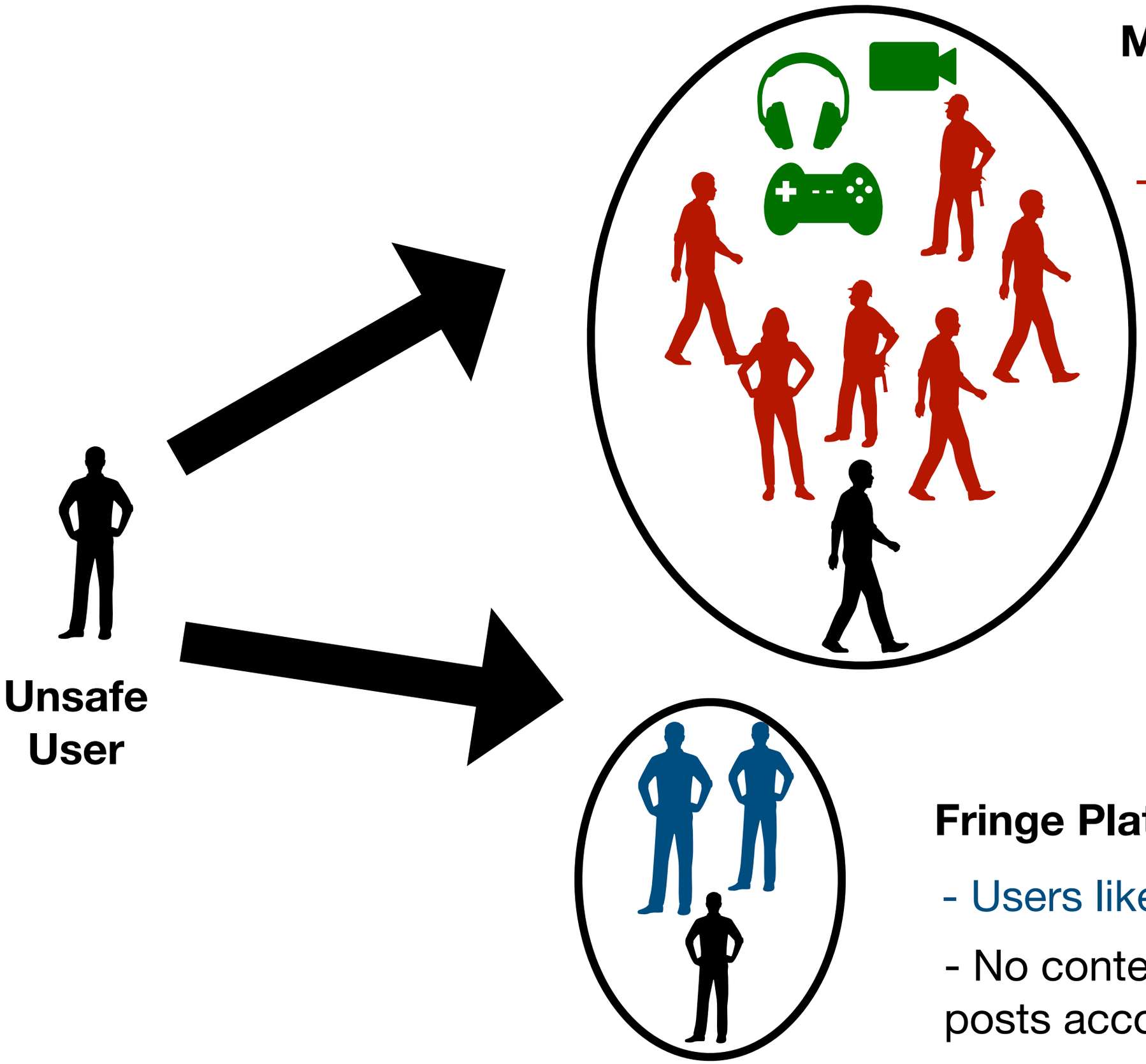


Fringe Platform

- Users like him (in unsafe terms)

- No content moderation:
posts according to his unsafety

Main User's Trade-off



Unsafe User

Moderated Platform

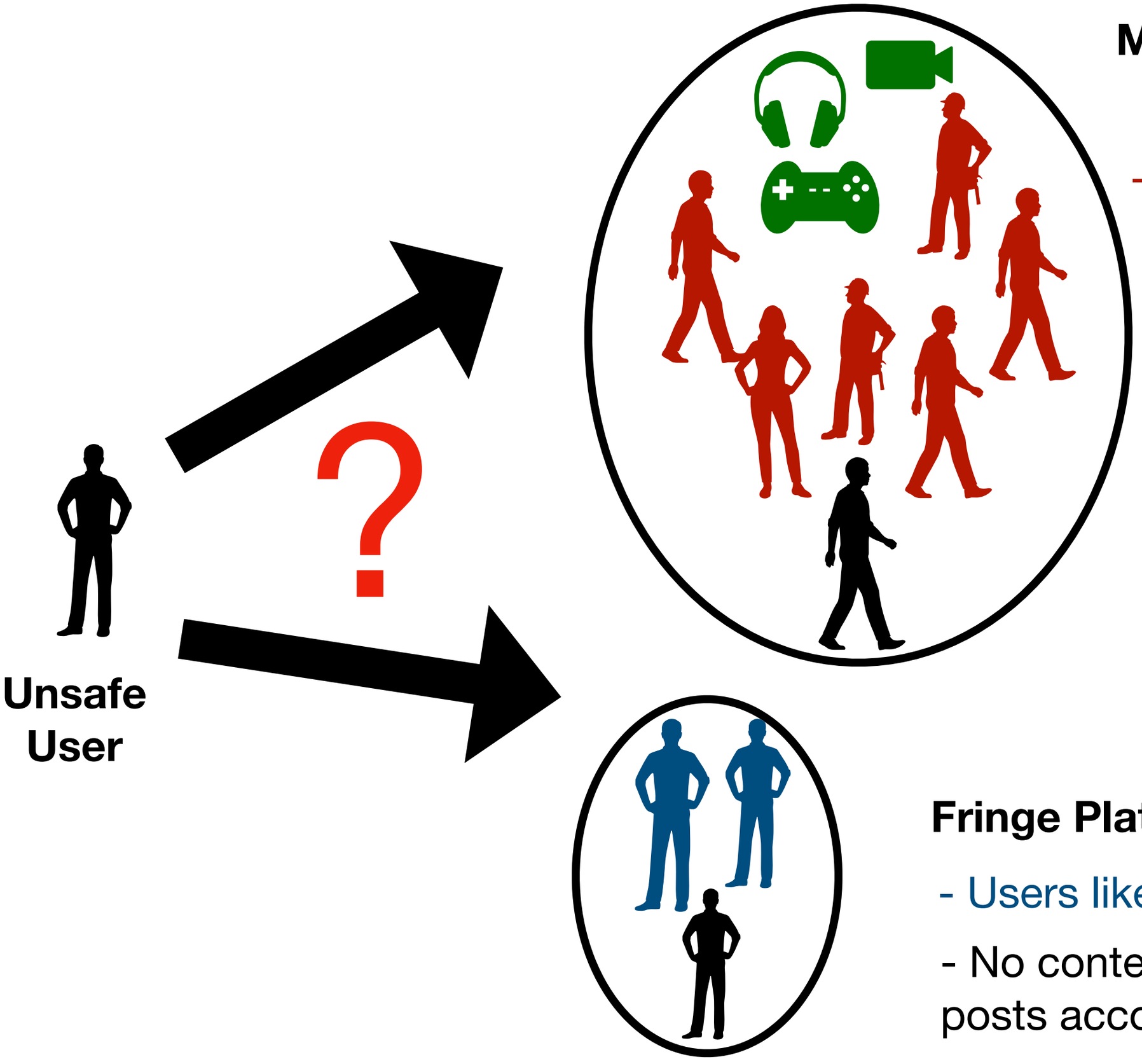
- More (and safer) users
- More Features
- Needs to respect the moderation policy:

self-censors

Fringe Platform

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

Main User's Trade-off



Unsafe User

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

Preview of the Main Results

Preview of the Main Results

1. **Prevalence of unsafe content:**
   - w/ **small** network effects: moderation **reduces** unsafety
   - w/ **large** network effects: moderation has a **u-shaped effect** on unsafety

Preview of the Main Results

1. **Prevalence of unsafe content:**
   - w/ **small** network effects: moderation **reduces** unsafety
   - w/ **large** network effects: moderation has a **u-shaped effect** on unsafety

2. **Policy:**
   - With **large** network effects, **too much self-regulation**
   - **Regulation** only **useful** with **small** network effects (or high competition)

Contribution (and literature)

Contribution (and literature)

- **Content Moderation Literature** (Madio & Quinn, 2024, Liu et al. 2021)
  1. On design: duopoly + endogenous content + homophily in unsafety
  2. On its consequences:
     - First to study (theoretically) migration, and therefore:
     - First to study **leakage**: ↑ Moderation may produce ↑ unsafe content

Contribution (and literature)

- **Content Moderation Literature** (Madio & Quinn, 2024, Liu et al. 2021)
  1. On design: duopoly + endogenous content + homophily in unsafety
  2. On its consequences:
     - First to study (theoretically) migration, and therefore:
     - First to study **leakage**: ↑ Moderation may produce ↑ unsafe content

- Analyzed potential mechanism behind empirical observations (Rizzi, 2023; Agarwal et al., 2022)
- Other Literature on Social Media (Zhang & Sarvary, 2012; Abreu & Jeon, 2019)

Roadmap

I. Model and Equilibrium

II. Policy

III. Extensions (if time allows)

- Multihoming
- Radicalization & Offline Violence

Model

Model

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

High θ = Unsafe content

Model

- 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** $(K_2 = 1)$

Model

- 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

Model

- 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

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

Model

- 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**

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

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

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

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

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

Users in the Platform

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

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

- 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

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

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

- 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

- 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

- 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

- 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

Advertisers

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

Advertisers

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

Are **averse** to unsafe content

Advertisers

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

Are **averse** to unsafe content

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

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

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** $\theta_i > K$

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

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** $\theta_i > K$

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

users in platform

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** $\theta_i > K$

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

$N_1(K)$: # users in platform

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

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** $\theta_i > K$

Advertisers aversion to unsafe content

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

users in platform

Price of ads

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

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

Timing

1. Platform (1) chooses K

Timing

1. Platform (1) chooses K

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

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)

Threshold Equilibrium (subgame for given K)

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

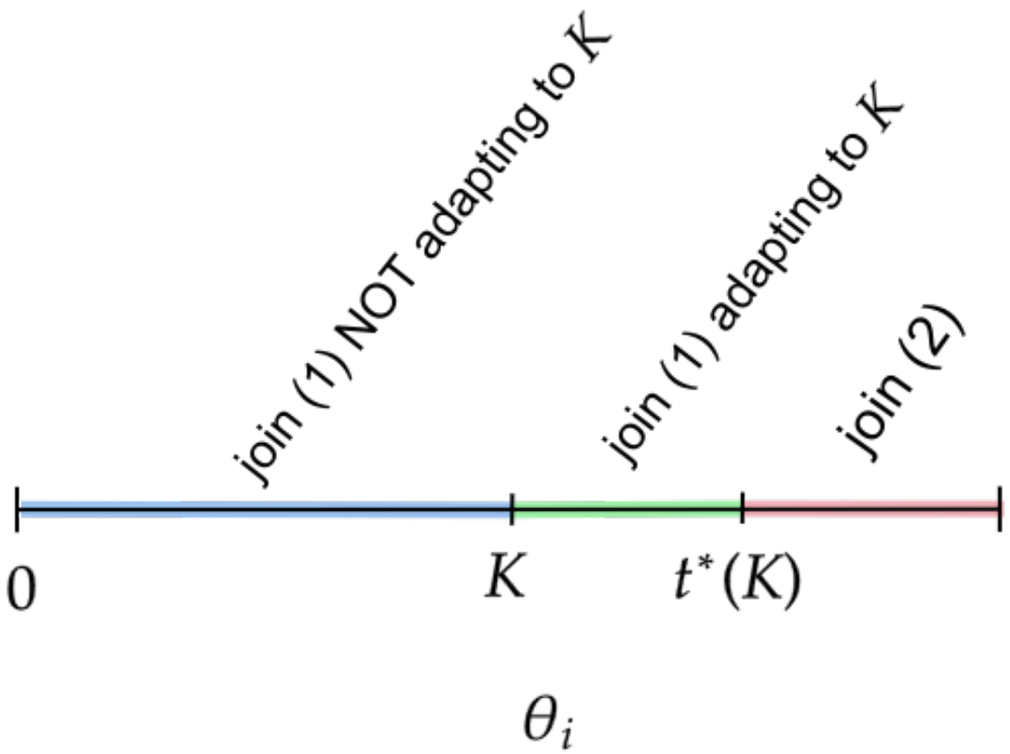
Under some assumptions on α, Δ; and given K,
there exist a **unique** threshold **equilibrium**

Threshold Equilibrium (subgame for given K)

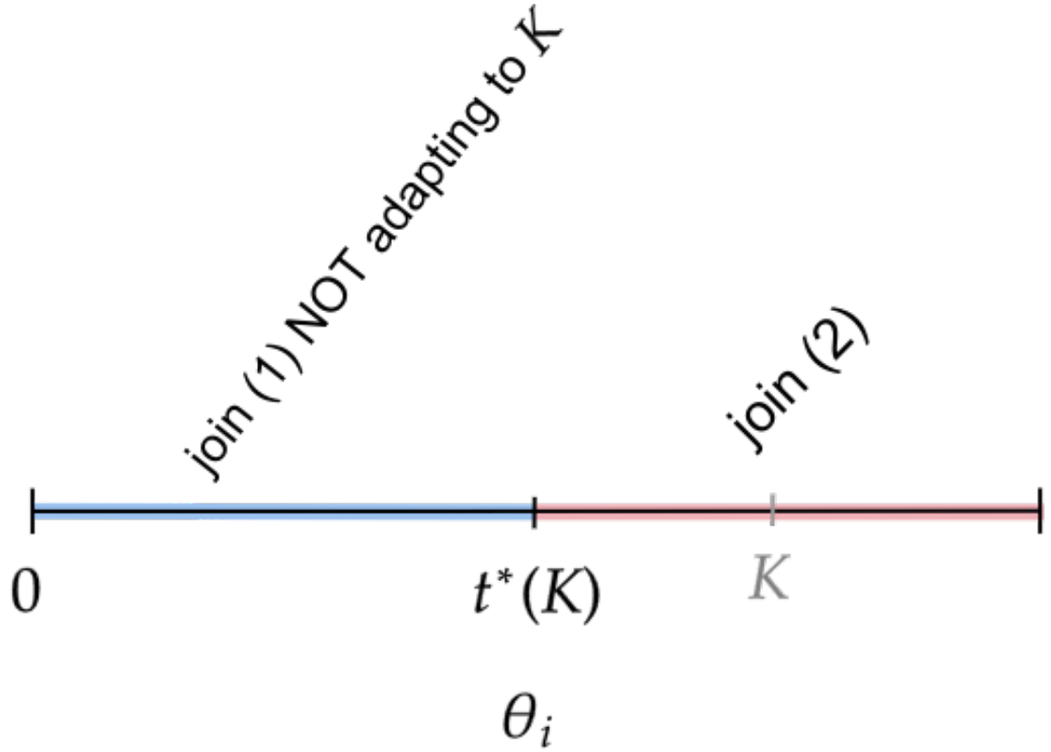
(Assumed) User i joins platform (1) iff $\theta_i < t^*$, otherwise, they join (2)

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

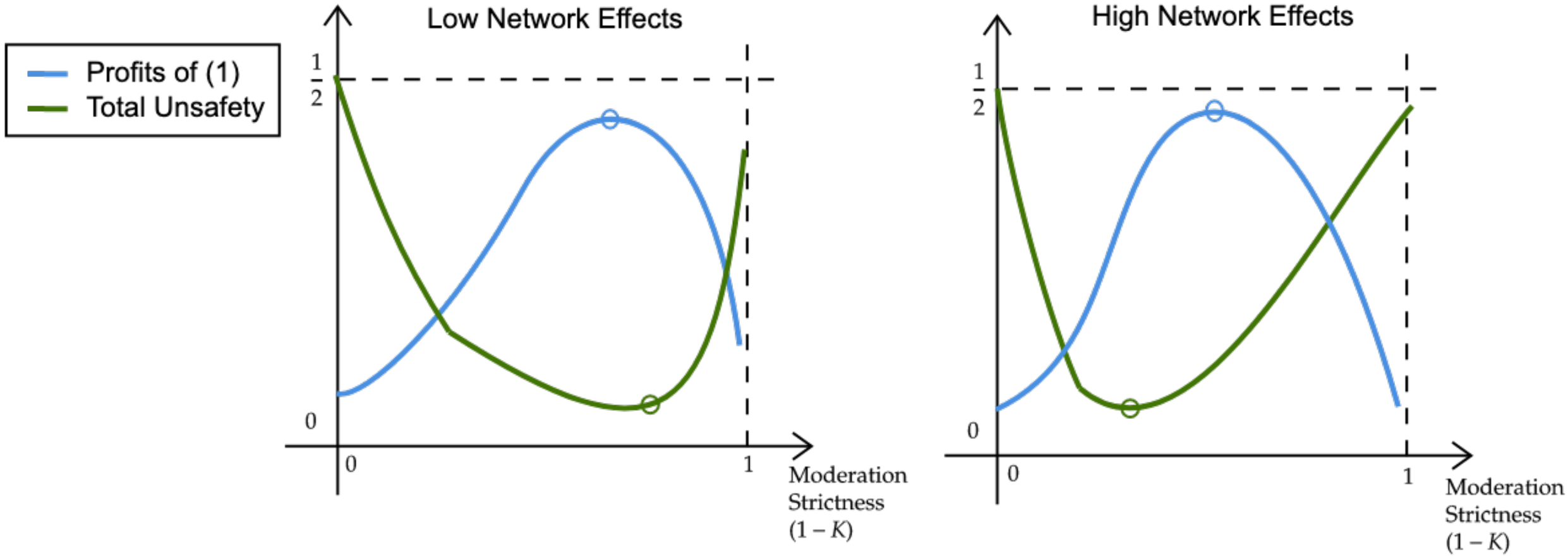
Low K (strict policy)



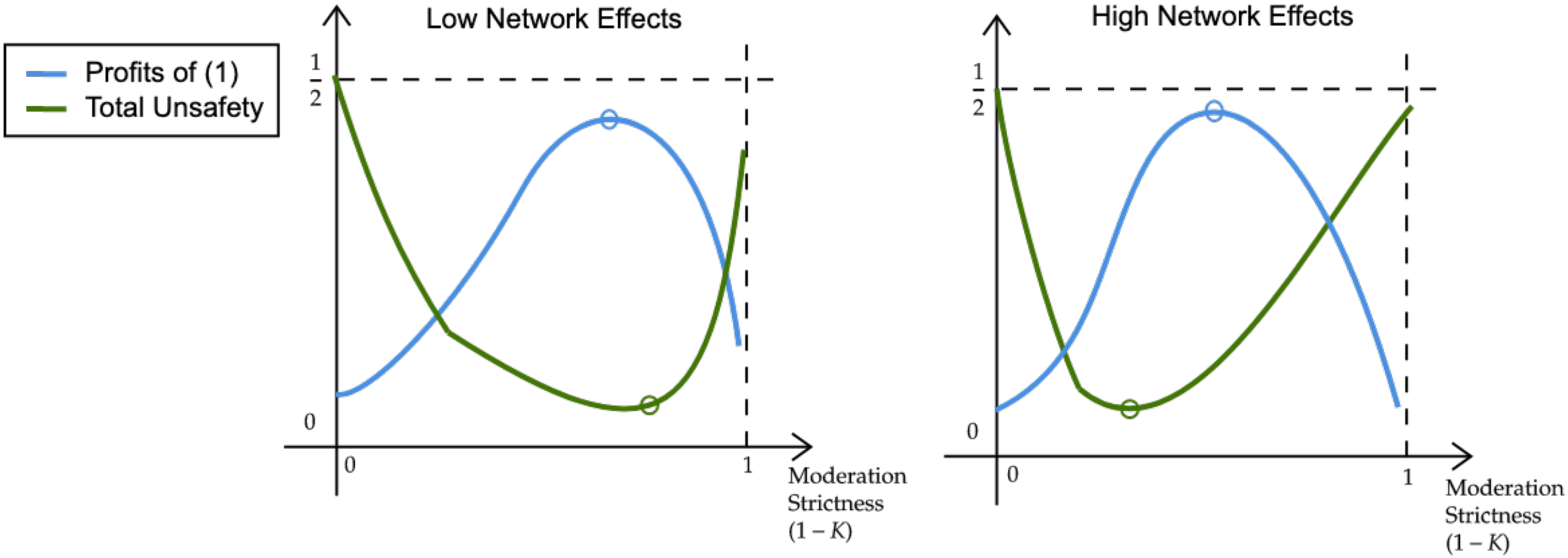
High K (lenient policy)



Characterization of the Equilibrium

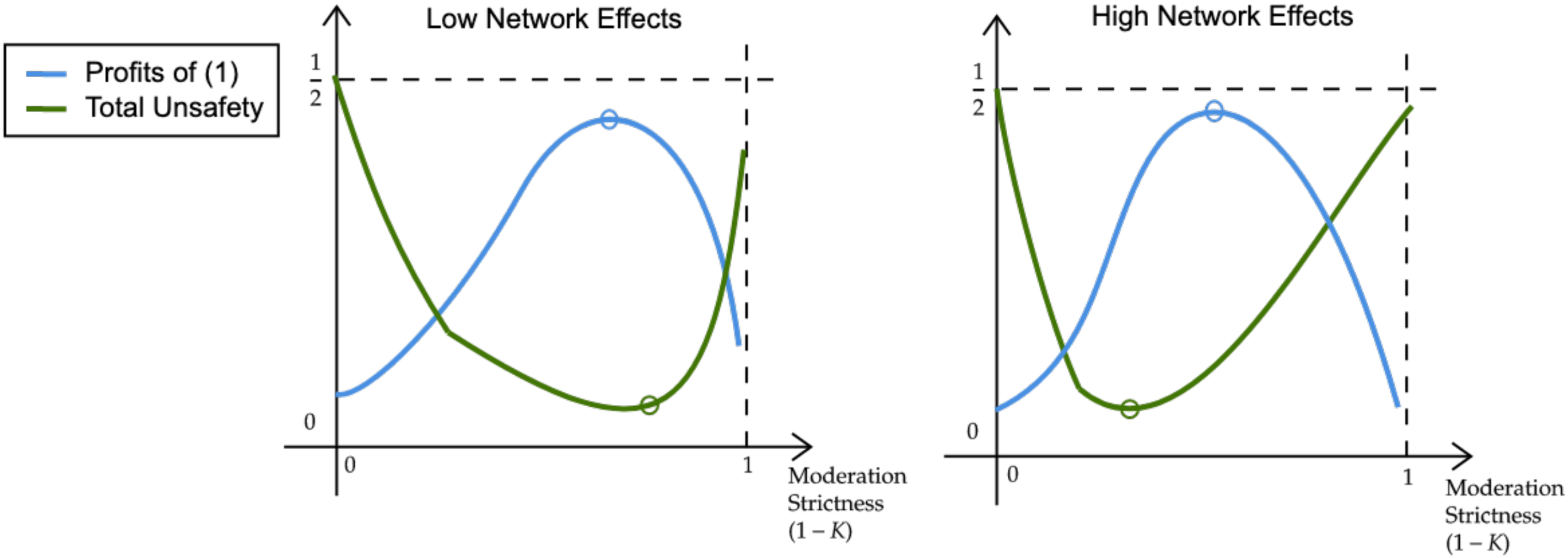


Characterization of the Equilibrium



Comparative statics: (excluding corner solutions)

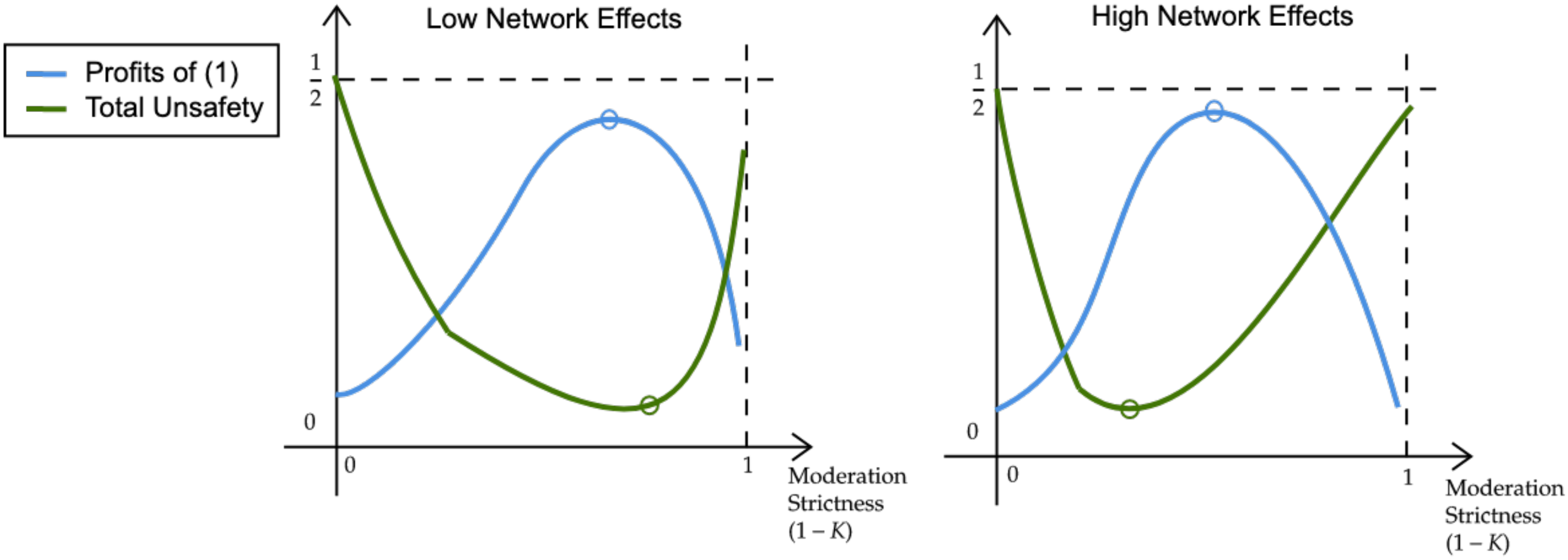
Characterization of the Equilibrium



Comparative statics: (excluding corner solutions)

I) As N.E. ↑, moderation strictness ↓ for **platform** and **regulator**

Characterization of the Equilibrium

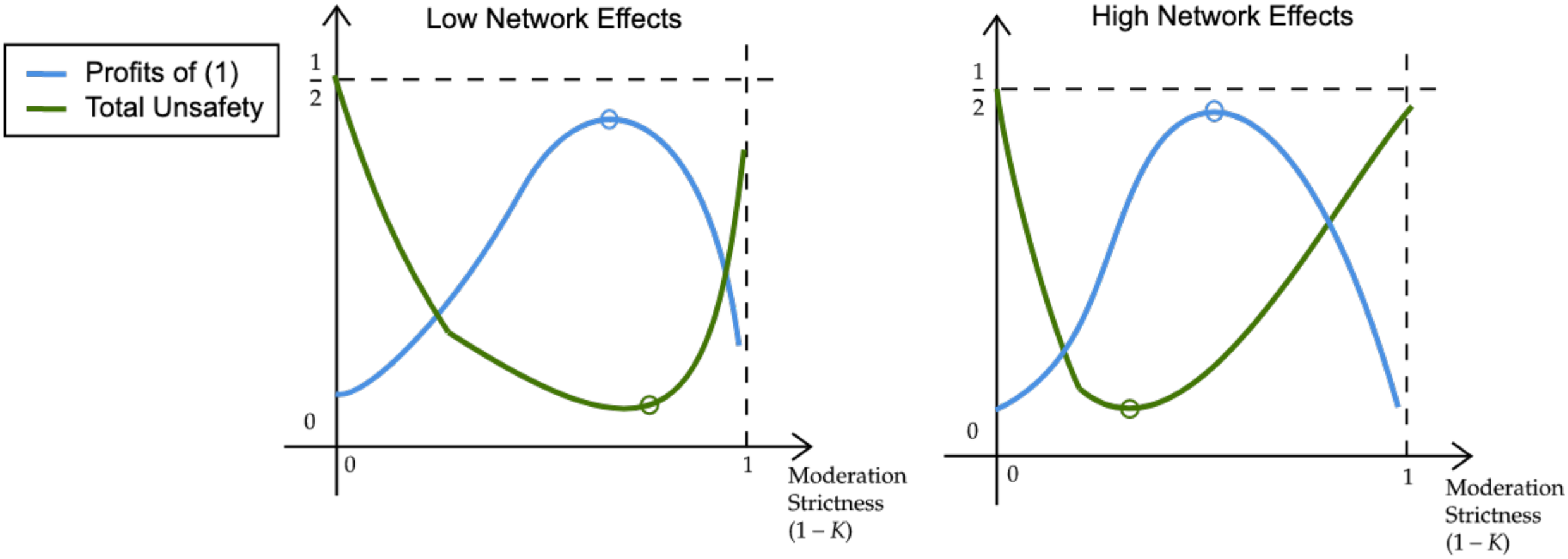


Comparative statics: (excluding corner solutions)

I) As N.E. ↑, moderation strictness ↓ for **platform** and **regulator**

II) It decreases **more** for the **regulator**

Characterization of the Equilibrium



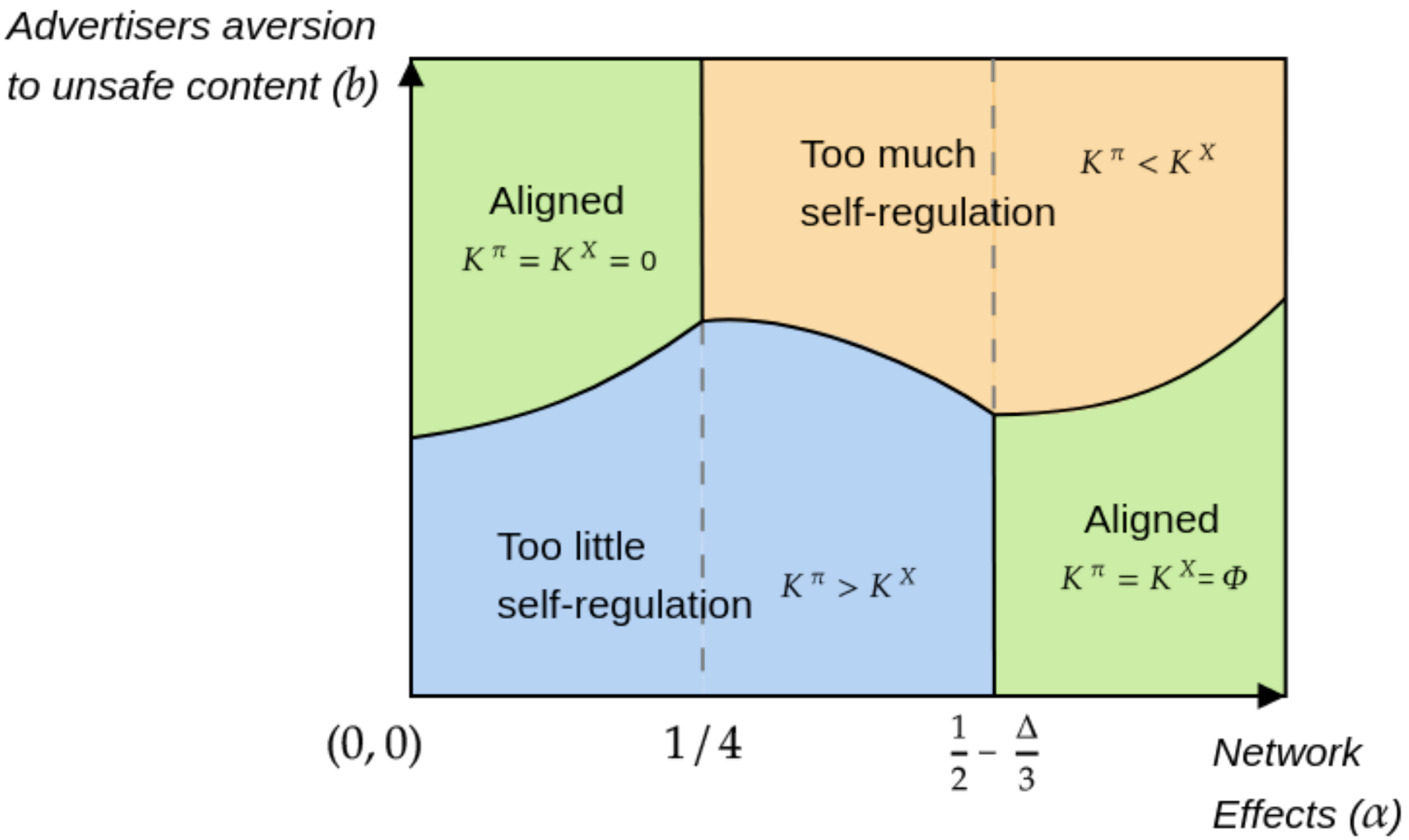
Comparative statics: (excluding corner solutions)

I) As N.E. ↑, moderation strictness ↓ for **platform** and **regulator**

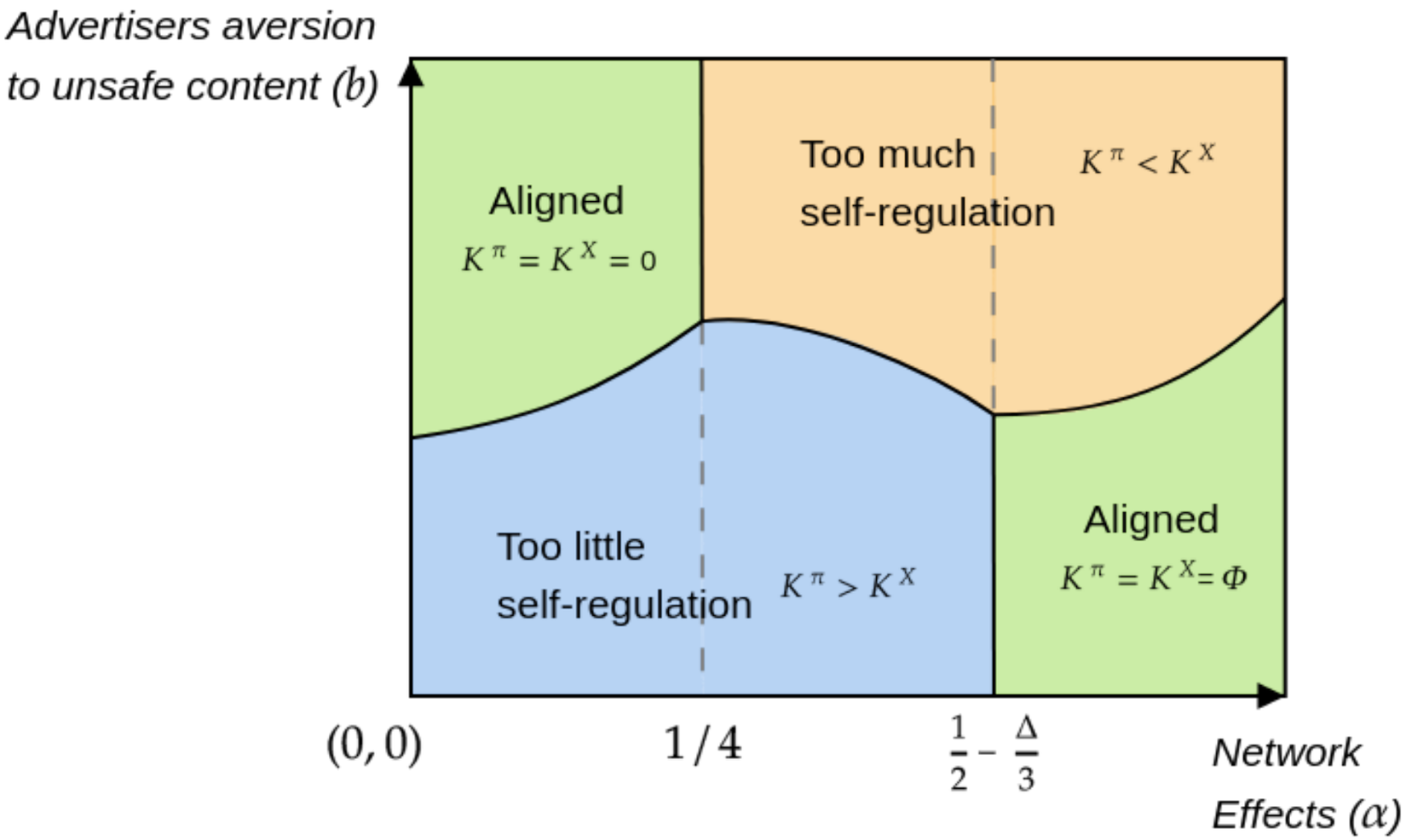
II) It decreases **more** for the **regulator**

III) As symmetry ↑, strictness ↓ for **platform** but ↑ for **regulator**

Policy (imposing a minimal content moderation)

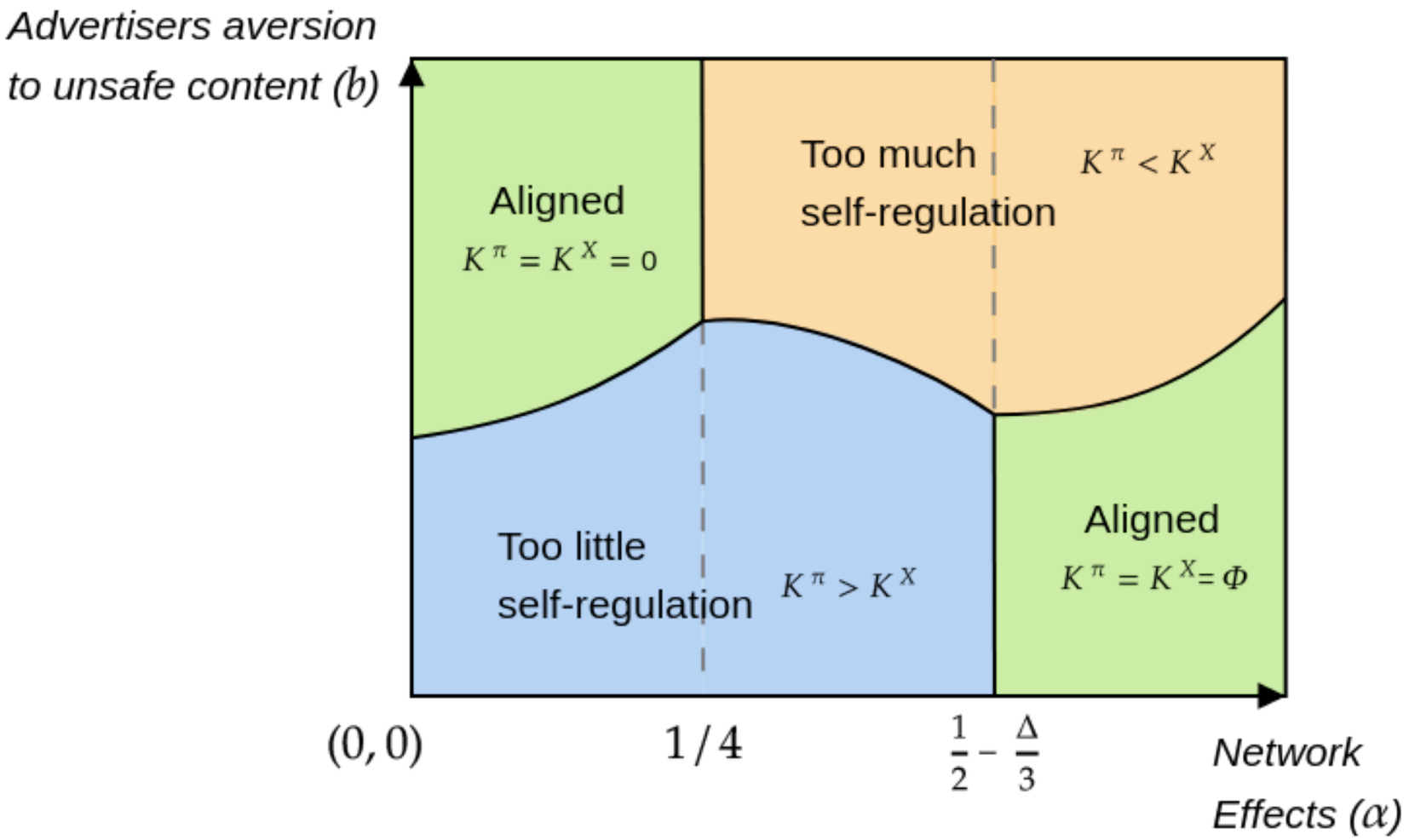


Policy (imposing a minimal content moderation)



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

Policy (imposing a minimal content moderation)



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.

Extensions

MULTIHOMING

OFFLINE VIOLENCE

Extensions

MULTIHOMING

1. Multihoming **reduces** incentives of the platform to moderate content

Multihoming $\approx$

Soften Network Effects = $\downarrow\alpha$

OFFLINE VIOLENCE

Extensions

MULTIHOMING

1. Multihoming **reduces** incentives of the platform to moderate content

Multihoming $\approx$

Soften Network Effects = $\downarrow\alpha$

$\uparrow$

Result in the literature
(cf. Crémer et al 2019's report for the EU)

OFFLINE VIOLENCE

Extensions

MULTIHOMING

1. Multihoming **reduces** incentives of the platform to moderate content

Multihoming $\approx$

Soften Network Effects = $\downarrow\alpha$

$\uparrow$

Result in the literature
(cf. Crémer et al 2019's report for the EU)

2. Optimal moderation to min unsafety:

- **More Lenient** than w singlehoming with **high network effects**
- **Stricter** otherwise

OFFLINE VIOLENCE

Extensions

MULTIHOMING

1. Multihoming **reduces** incentives of the platform to moderate content

Multihoming $\approx$

Soften Network Effects = $\downarrow\alpha$

↑

Result in the literature
(cf. Crémer et al 2019's report for the EU)

2. Optimal moderation to min unsafety:

- **More Lenient** than w singlehoming with **high network effects**
- **Stricter** otherwise

OFFLINE VIOLENCE

Model Extension:

t=3. Users preferences align (oppose) unsafety of the content they read

t=4. Prob[violence] increases (decreases) with new preference for unsafety

Extensions

MULTIHOMING

1. Multihoming **reduces** incentives of the platform to moderate content

Multihoming $\approx$

Soften Network Effects = $\downarrow\alpha$

↑

Result in the literature
(cf. Crémer et al 2019's report for the EU)

2. Optimal moderation to min unsafety:

- **More Lenient** than w singlehoming with **high network effects**
- **Stricter** otherwise

OFFLINE VIOLENCE

Model Extension:

t=3. Users preferences align (oppose) unsafety of the content they read

t=4. Prob[violence] increases (decreases) with new preference for unsafety

Main Result:

Moderate content moderation can reduce (increase) offline violence

Users are attracted to safer platforms and converge to the safer content they find there

Conclusion

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)

(Near) Future

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)

(Near) Future

- **Structural** Empirical Model, we need counterfactuals

Working on that

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)

(Near) Future

- **Structural** Empirical Model, we need counterfactuals

Working on that

- Other *non-IO* applications

e.g. Cancel culture (~Tirole's safe spaces)