Content Moderation in Presence of Fringe Platforms

Iván Rendo (TSE)

3rd Year Paper Presentation

Toulouse, 5 July 2024

Motivation

- **Online** extreme/**unsafe content** 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)
  - bullying, food disorders, pornography…

Motivation

- **Online** extreme/**unsafe content** 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)
  - bullying, food disorders, pornography…

➡ EU Response: **Digital Services Act**

Motivation

- **Online** extreme/**unsafe content** 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)
  - bullying, food disorders, pornography…

➡ EU Response: **Digital Services Act**

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

Motivation

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

(Hamiz and Ariza, 2022)

➡ EU Response: **Digital Services Act**

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

4Chan, Parler, Truth…

~ 6% of the US market

(Stocking et al., 2022)

Motivation

- **Online** extreme/**unsafe content** 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)
  - bullying, food disorders, pornography…

➡ EU Response: **Digital Services Act**

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

4Chan, Parler, Truth…

~ 6% of the US market (Stocking et al., 2022)

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

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

Motivation

- **Online** extreme/**unsafe content** 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)
  - bullying, food disorders, pornography…

➡ EU Response: **Digital Services Act**

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

4Chan, Parler, Truth…

~ 6% of the US market (Stocking et al., 2022)

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

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

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

Questions:

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

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

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

Questions:

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

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

➡ What incentives do platforms have to self-regulate?

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

Preview of the Main Results

1. More content moderation $\not\Rightarrow$ Less unsafety

2. W Large network effects, platform over-self-moderates

Preview of the Main Results

1. More content moderation $\not\Rightarrow$ Less unsafety

   Due to migration

2. W Large network effects, platform over-self-moderates

Preview of the Main Results

1. More content moderation $\not\Rightarrow$ Less unsafety

   Due to migration

2. W Large network effects, platform over-self-moderates

   Mainstream doesn’t internalizes what happens on the fringe

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**
  - Assumed $K_2 = 1$, No Content Moderation on the Fringe

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**
  - Assumed $K_2 = 1$, No Content Moderation on the Fringe
- 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**
  - Assumed $K_2 = 1$, No Content Moderation on the Fringe

- 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** $\theta_i > K$

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

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** $\theta_i > K$

Advertisers aversion to unsafe content

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

Price of ads

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** $\theta_i > K$

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

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** $\theta_i > K$

Advertisers aversion to unsafe content

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

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)

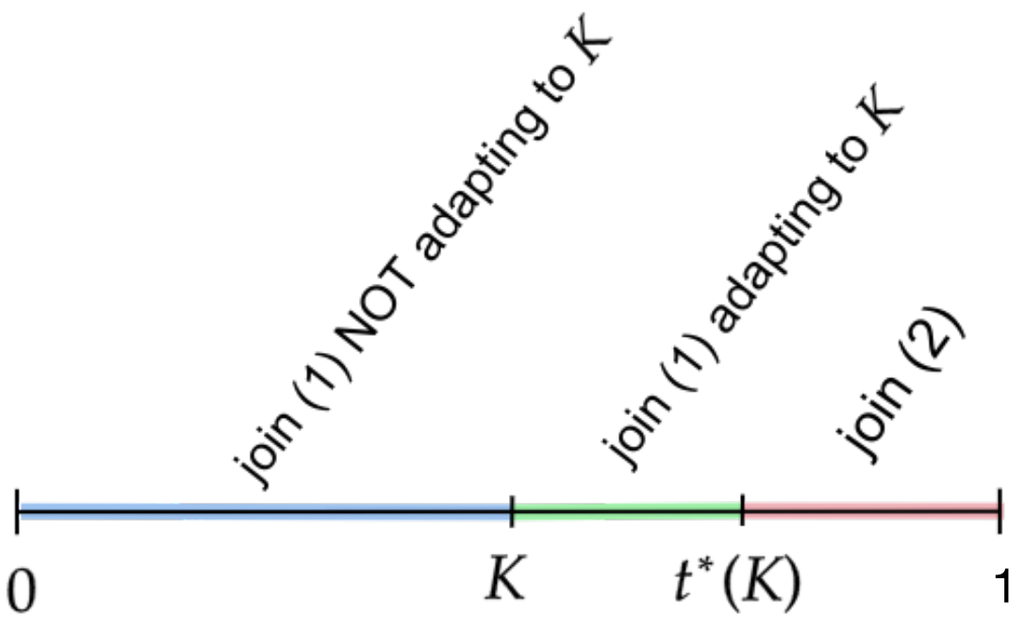
(Assumed) User i joins platform (1) iff $\theta_i < t^*$, otherwise, they join (2)

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**



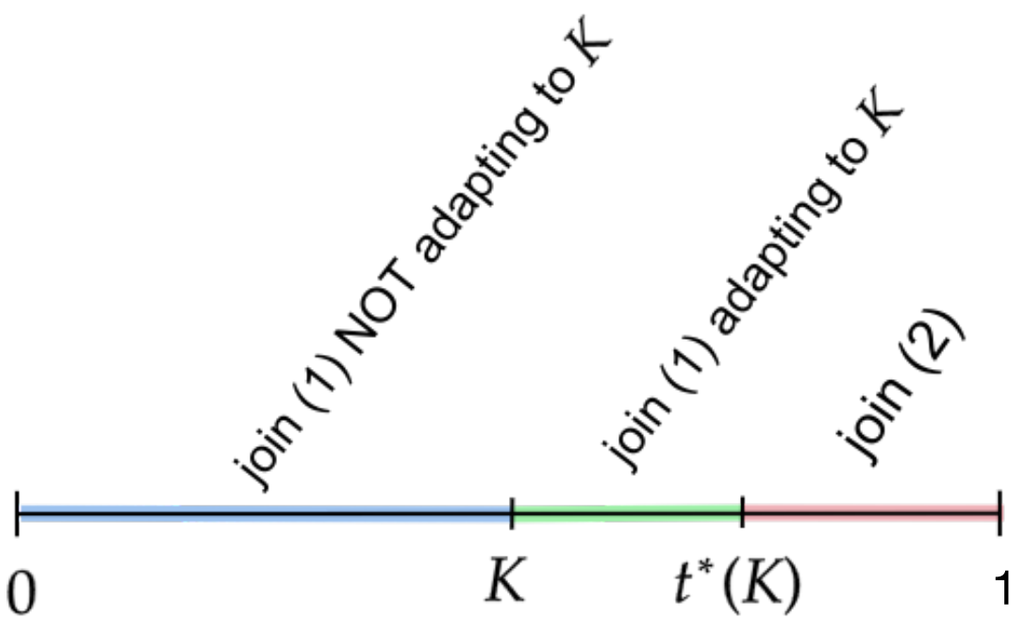
Users Unsafety θ_i

Threshold Equilibrium (subgame for given K)

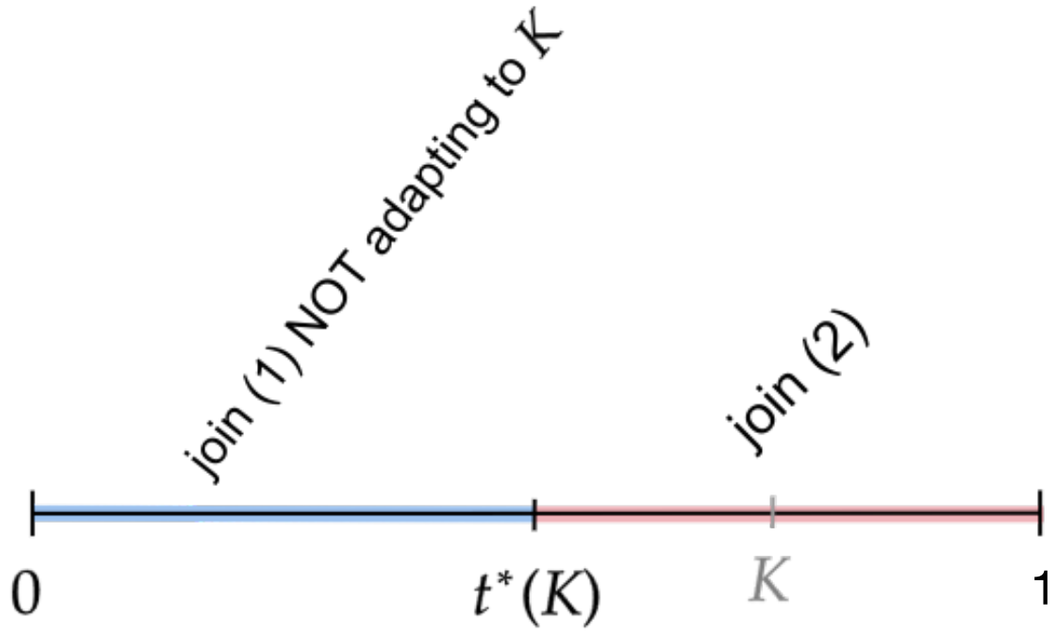
(Assumed) User i joins platform (1) iff $\theta_i < t^*$, otherwise, they join (2)

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

If policy lenient enough…

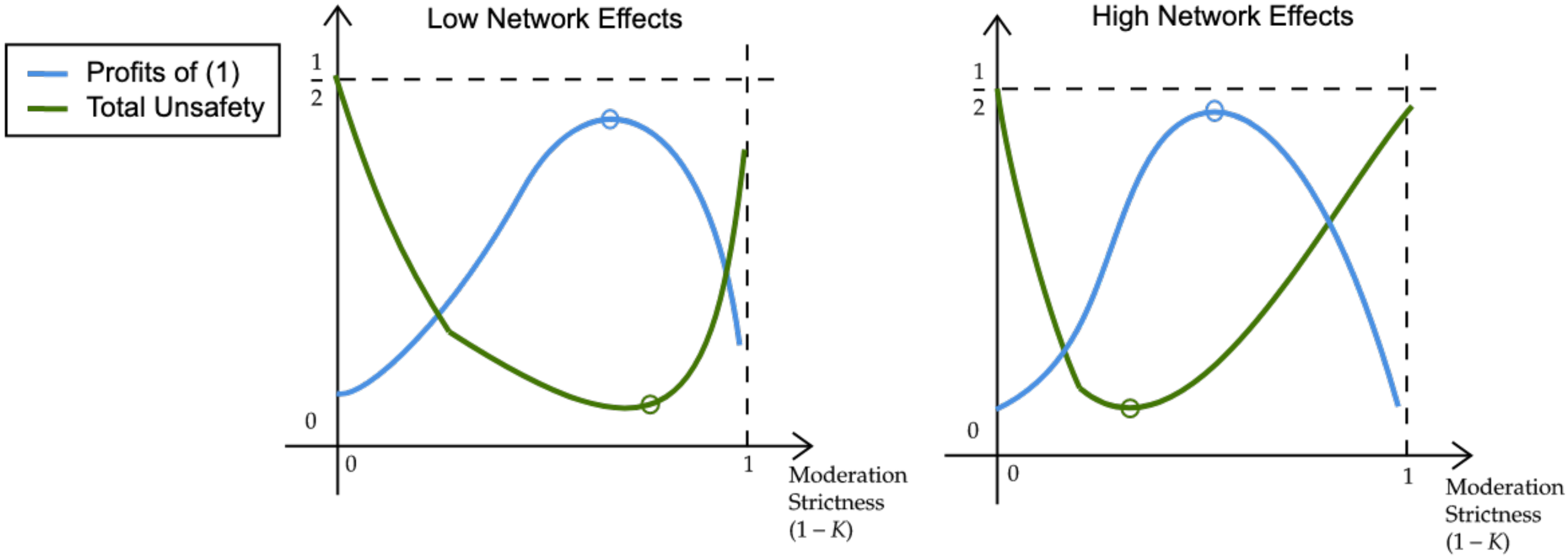


Users Unsafety θ_i

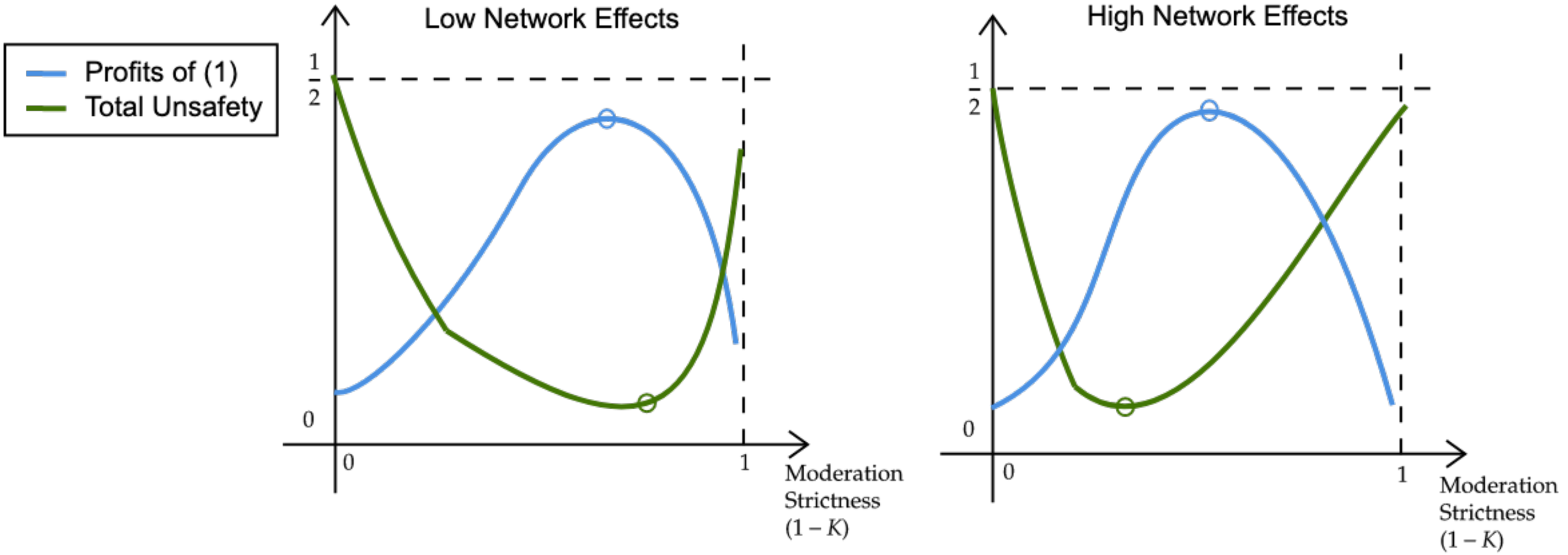


Users Unsafety θ_i

Characterization of the Equilibrium

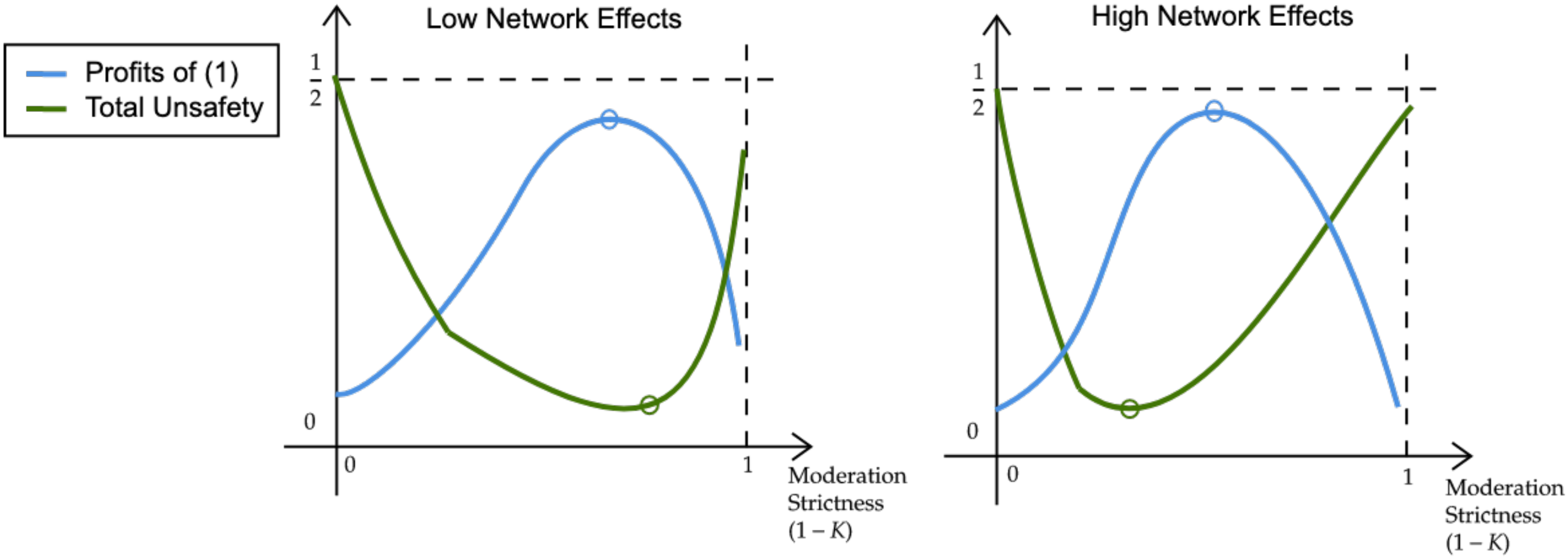


Characterization of the Equilibrium



Comparative statics: (excluding corner solutions)

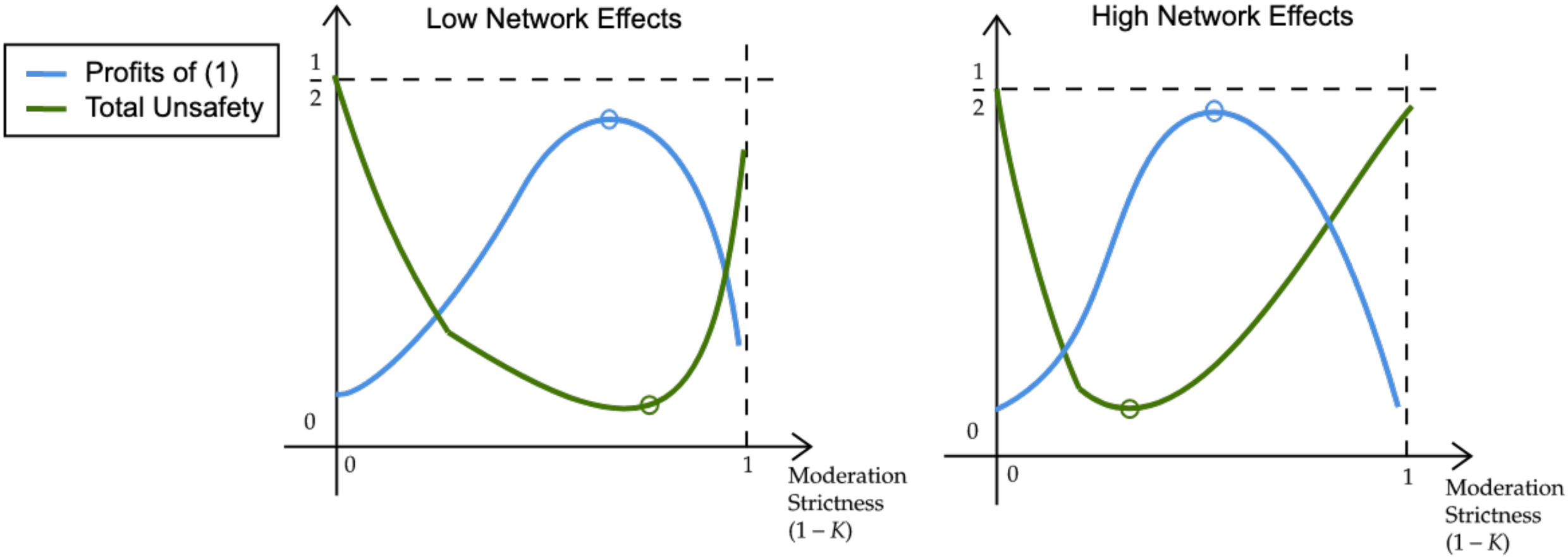
Characterization of the Equilibrium



Comparative statics: (excluding corner solutions)

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

Characterization of the Equilibrium

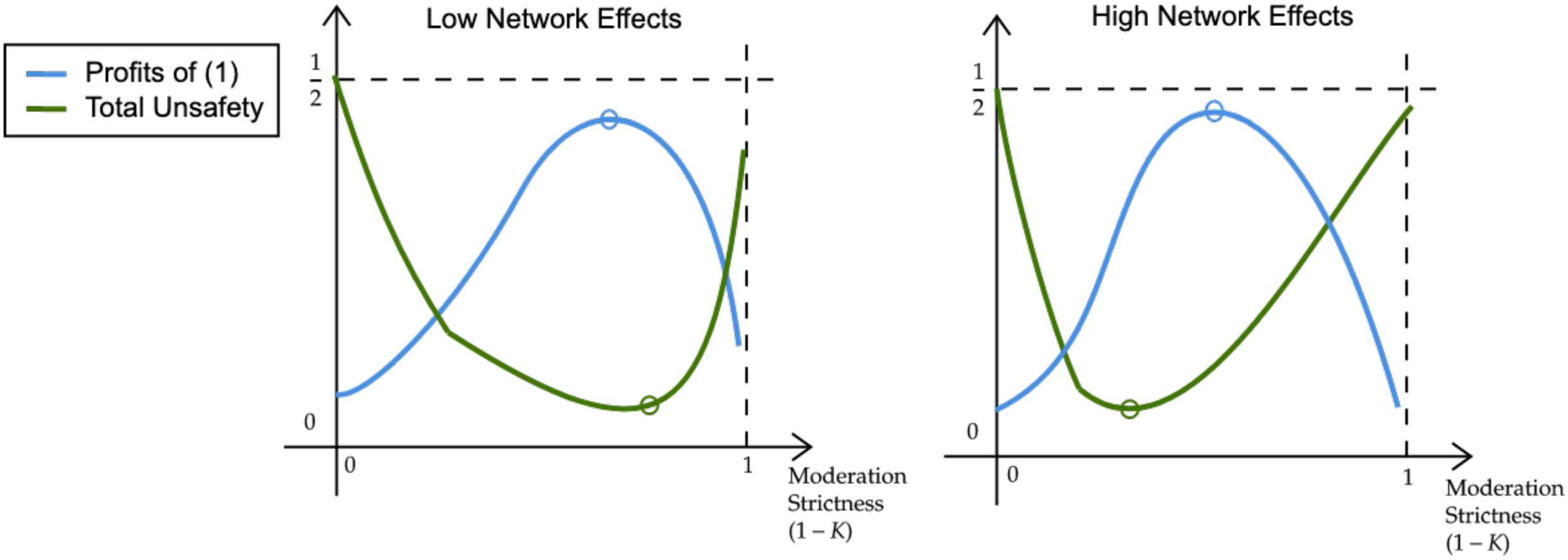


Comparative statics: (excluding corner solutions)

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

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

Characterization of the Equilibrium



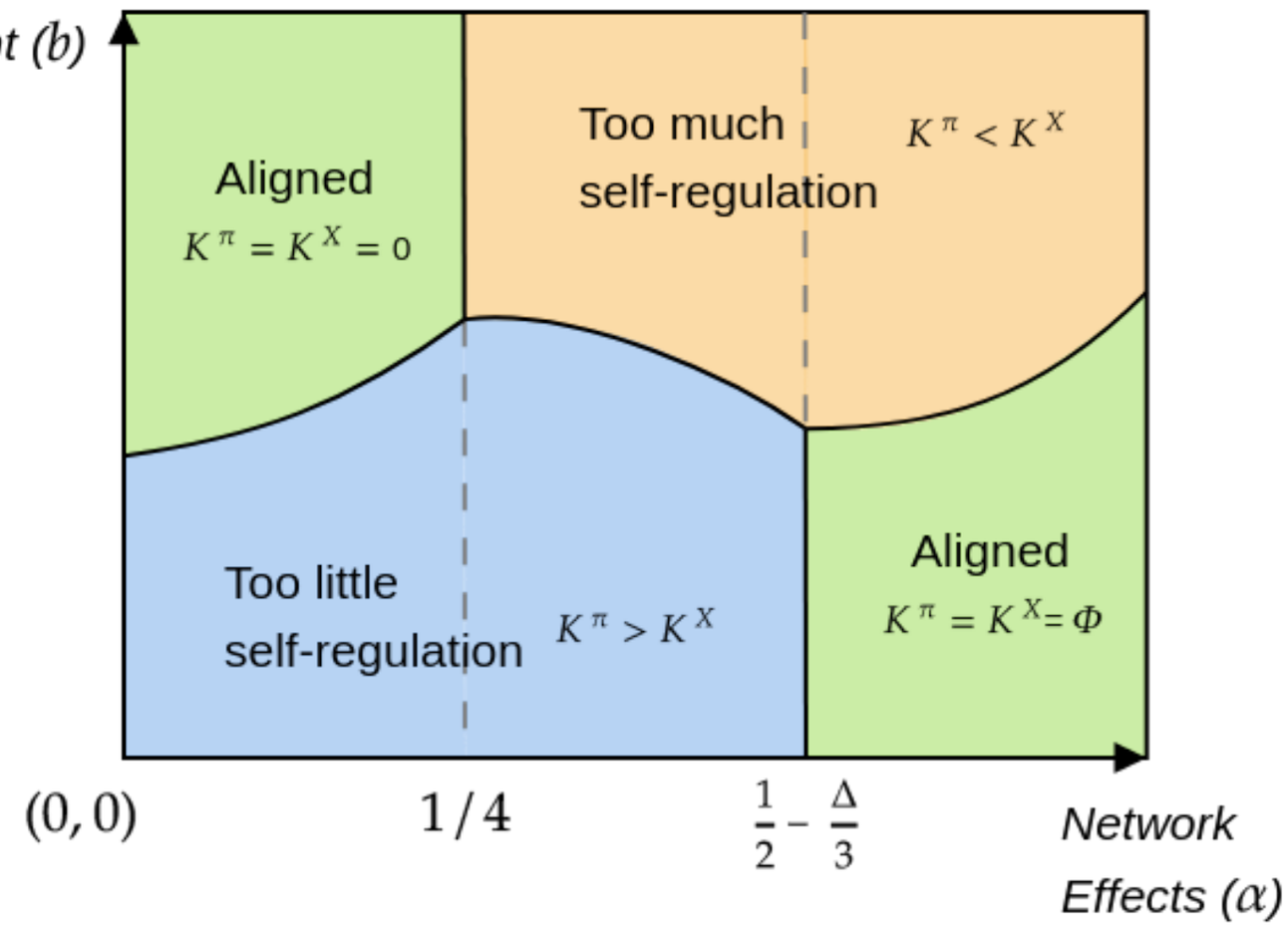
Comparative statics: (excluding corner solutions)

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

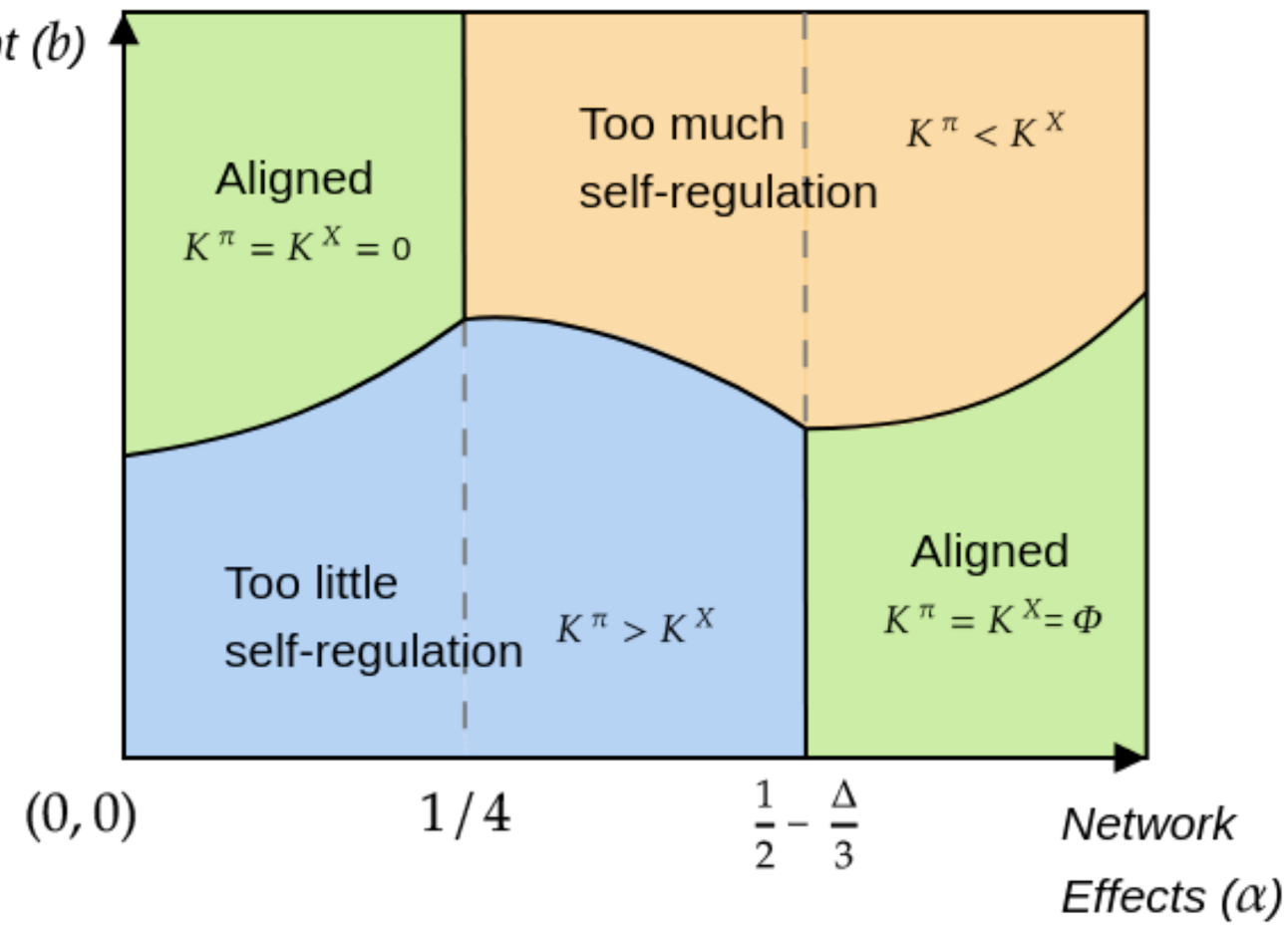
II) It decreases **more** for the **regulator**

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

Policy

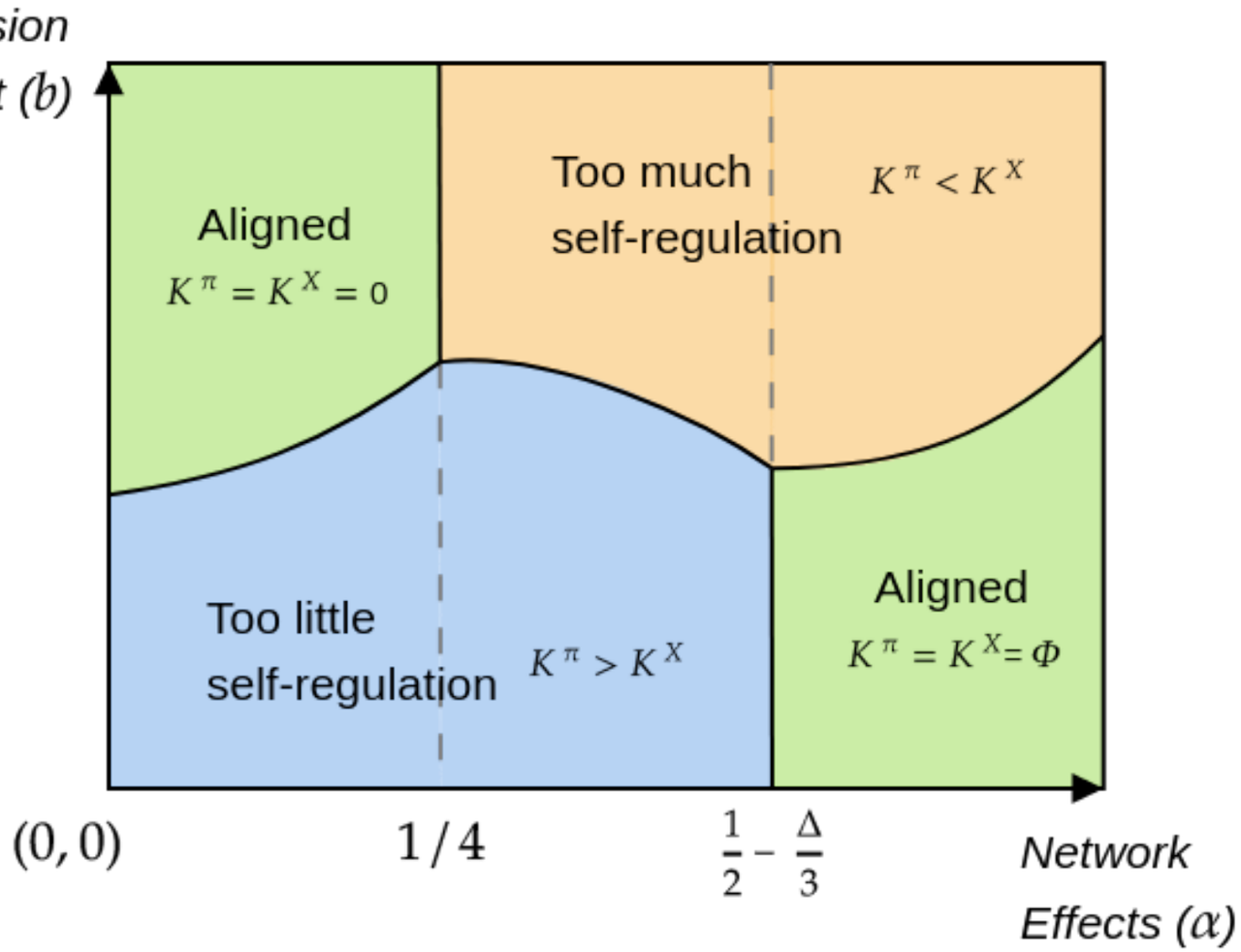

Advertisers aversion
to unsafe content (b)
Aligned
$K^{\pi} = K^{X} = 0$
Too much
self-regulation
$K^{\pi} < K^{X}$
Too little
self-regulation
$K^{\pi} > K^{X}$
Aligned
$K^{\pi} = K^{X} = \Phi$
(0, 0)
1 / 4
$\frac{1}{2} - \frac{\Delta}{3}$
Network
Effects (α)

Policy



Imposing a minimal content moderation:

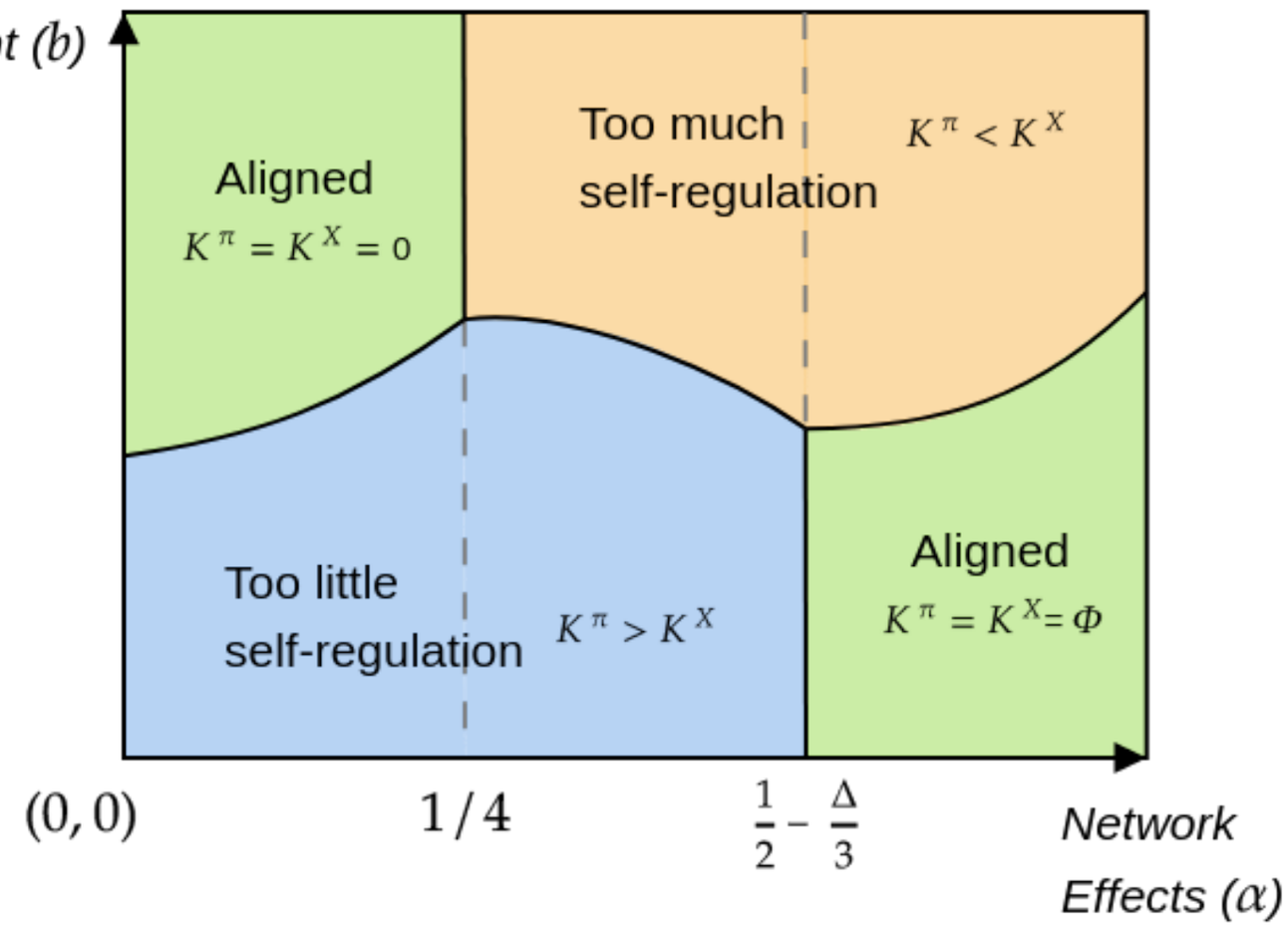
Policy



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

Extensions

MULTIHOMING

Multihoming **increases** incentives of the **platform** and the **regulator (more)** to moderate content

Extensions

MULTIHOMING

Multihoming **increases** incentives of the **platform** and the **regulator (more)** to moderate content

1. Multihoming $\approx$
   Soften Network Effects = $\downarrow\alpha$

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

Extensions

MULTIHOMING

Multihoming **increases** incentives of the **platform** and the **regulator (more)** to moderate content

1. Multihoming $\approx$
   Soften Network Effects = $\downarrow\alpha$

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

OFFLINE VIOLENCE

Extensions

MULTIHOMING

Multihoming **increases** incentives of the **platform** and the **regulator (more)** to moderate content

1. Multihoming $\approx$
   Soften Network Effects = $\downarrow\alpha$

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

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

Multihoming **increases** incentives of the **platform** and the **regulator (more)** to moderate content

1. Multihoming $\approx$
   Soften Network Effects = $\downarrow\alpha$

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

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

THANKS!

P.S. Working on a **Structural** Empirical Model

Feel free to reach out!