

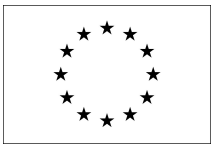
# Content Moderation in Presence of Fringe Platforms

Iván Rendo (TSE)



# Motivation

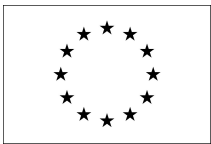
- Increased interest in **online** hateful/extreme/**unsafe content**:
  - E.g. spread of jihadism, bullying, food disorders...
  - Jiménez-Durán (2022) links online hate to **offline violence**



➡ EU Response: **Digital Services Act (DSA)**

# Motivation


- Increased interest in **online** hateful/extreme/**unsafe content**:
  - E.g. spread of jihadism, bullying, food disorders...
  - Jiménez-Durán (2022) links online hate to **offline violence**



➡ EU Response: **Digital Services Act (DSA)**

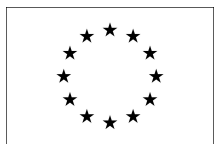
- Complementary views on content moderation:

# Motivation

- Increased interest in **online** hateful/extreme/**unsafe content**:
    - E.g. spread of jihadism, bullying, food disorders...
    - Jiménez-Durán (2022) links online hate to **offline violence**
-  ➡ EU Response: **Digital Services Act (DSA)**
- Complementary views on content moderation:
    - Duch-Brown's: **Constant unsafe content** across time

# Motivation

- Increased interest in **online** hateful/extreme/**unsafe content**:
  - E.g. spread of jihadism, bullying, food disorders...
  - Jiménez-Durán (2022) links online hate to **offline violence**

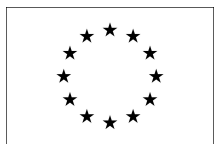


➡ EU Response: **Digital Services Act (DSA)**

- Complementary views on content moderation:
  - Duch-Brown's: **Constant unsafe content** across time
  - ➡ BUT more **impact** today: everyone in the same large platforms.

# Motivation

- Increased interest in **online** hateful/extreme/**unsafe content**:
  - E.g. spread of jihadism, bullying, food disorders...
  - Jiménez-Durán (2022) links online hate to **offline violence**

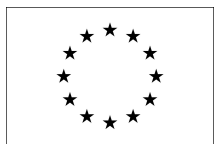


➡ EU Response: **Digital Services Act (DSA)**

- Complementary views on content moderation:
  - Duch-Brown's: **Constant unsafe content** across time
  - ➡ BUT more **impact** today: everyone in the same large platforms.
  - Lefouili & Madio ('22): ↑ regulation = ↑ **migration** to fringes ↓ Impact

# Motivation

- Increased interest in **online** hateful/extreme/**unsafe content**:
  - E.g. spread of jihadism, bullying, food disorders...
  - Jiménez-Durán (2022) links online hate to **offline violence**

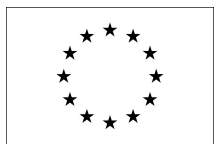


➡ EU Response: **Digital Services Act (DSA)**

- Complementary views on content moderation:
  - Duch-Brown's: **Constant unsafe content** across time
  - ➡ BUT more **impact** today: everyone in the same large platforms.
  - Lefouili & Madio ('22): ↑ regulation = ↑ **migration** to fringes ↓ Impact
  - ADL (Anti-Antisemitic lobby) viral video, trade-off:

# Motivation

- Increased interest in **online** hateful/extreme/**unsafe content**:
  - E.g. spread of jihadism, bullying, food disorders...
  - Jiménez-Durán (2022) links online hate to **offline violence**



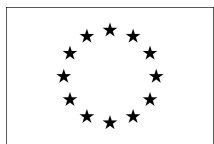
➡ EU Response: **Digital Services Act (DSA)**

- Complementary views on content moderation:
  - Duch-Brown's: **Constant unsafe content** across time
  - ➡ BUT more **impact** today: everyone in the same large platforms.
  - Lefouili & Madio ('22): ↑ regulation = ↑ **migration** to fringes ↓ Impact
  - ADL (Anti-Antisemitic lobby) viral video, trade-off:
    - ➡ ↑ **moderation** on X = ↑ **migration** = ↑ radicalization, hate...



# Motivation

- Increased interest in **online** hateful/extreme/**unsafe content**:
  - E.g. spread of jihadism, bullying, food disorders...
  - Jiménez-Durán (2022) links online hate to **offline violence**



➡ EU Response: **Digital Services Act (DSA)**

- Complementary views on content moderation:
  - Duch-Brown's: **Constant unsafe content** across time
  - ➡ BUT more **impact** today: everyone in the same large platforms.
  - Lefouili & Madio ('22): ↑ regulation = ↑ **migration** to fringes ↓ Impact
  - ADL (Anti-Antisemitic lobby) viral video, trade-off:
    - ➡ ↑ **moderation** on X = ↑ **migration** = ↑ radicalization, hate...

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

# Today

# Today

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

**Content Moderation on Content Unsafety**

through **Migration\*** to a **fringe** platform... for an ad-funded platform

# Today

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

## **Content Moderation on Content Unsafety**

through **Migration**\* to a **fringe** platform... for an ad-funded platform

- ➡ How **migration** is affected by content moderation **policies**
- ➡ How **unsafe content** is affected by **migration**
- ➡ What **incentives** do platforms have to **self-regulate**
- ➡ Characterize the **optimal regulation** to **minimize** unsafe content

# Main Features of the Model

# Main Features of the Model

## Users:

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

# Main Features of the Model

## Users:

- Create + read 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)
- An **Unmoderated** one, lower quality platform: **no content moderation**

# Main Features of the Model

## Users:

- Create + read 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)
- An **Unmoderated** one, lower quality platform: **no content moderation**  
8Chan, Truth, Parler



# Main Features of the Model

## Users:

- Create + read 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)
- An **Unmoderated** one, lower quality platform: **no content moderation**

8Chan, Truth, Parler

- **Endogenous composition** ~ migration
  - Users' trade-off: network size, quality, (un)safe content
  - Platform's trade-off: participation, unsafe content

# Preview of the Main Results

# Preview of the Main Results

## 1. Prevalence of unsafe content:

- i. **U-shaped** in moderation intensity, w **large** network effects
- ii. **Decreasing** in moderation intensity, w **small** network effects

# Preview of the Main Results

## 1. Prevalence of unsafe content:

- i. **U-shaped** in moderation intensity, w **large** network effects
- ii. **Decreasing** in moderation intensity, w **small** network effects

## 2. Policy:

- **Incentives misalignment** b/ platform & regulator (min unsafe content)
- Imposing a **minimal** content moderation intensity (policy):
  - i. Only useful with small network effects (or high competition)
  - ii. Otherwise, always **superfluous** (not binding)

# Roadmap

## I. Model

- Equilibrium

## **II. Policy Discussion**

## III. Extensions

- Multihoming
- Radicalization & Offline Violence

**THEORY**

# Model

# Model

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



# Model

- A unit mass of **users**, heterogeneous in their preferences for unsafe content:  $\theta_i \sim U(0,1)$
- 2 **platforms**  $j = 1,2$ 
  - with  $K_j = \text{max unsafety level allowed}$   $(K_2 = 1)$

# Model

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

# Model

- A unit mass of **users**, heterogeneous in their preferences for unsafe content:  $\theta_i \sim U(0,1)$
- **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 unsafety  $\theta_i^C$ 
$$\theta_i^C = \min\{\theta_i, K_j\}$$

# Model

- A unit mass of **users**, heterogeneous in their preferences for unsafe content:  $\theta_i \sim U(0,1)$
- **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 unsafety  $\theta_i^C$ 
$$\theta_i^C = \min\{\theta_i, K_j\}$$
- User  $i$  in platform  $j$  **reads** a random sample of the content, of avg unsafety  $\bar{\theta}_j$

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

# Model

- A unit mass of **users**, heterogeneous in their preferences for unsafe content:  $\theta_i \sim U(0,1)$
- **2 platforms**  $j = 1,2$ 
  - with  $K_j = \text{max unsafety level allowed}$  ( $K_2 = 1$ )
- User  $i$  in platform  $j$  **creates** 1 piece of content of unsafety  $\theta_i^C$ 
$$\theta_i^C = \min\{\theta_i, K_j\}$$
- User  $i$  in platform  $j$  **reads** a random sample of the content, of avg unsafety  $\bar{\theta}_j$

$$\bar{\theta}_j = \int_{i \in j} \theta_i^C di \quad = \text{average unsafety of content in platform } j$$



- Platform 1, **moderated**, is intrinsically better than 2, **unmoderated**

- 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:

$$U_1(\theta_i) = \alpha N_1 - |\theta_i - \bar{\theta}_1| + \Delta$$
$$U_2(\theta_i) = \alpha N_2 - |\theta_i - \bar{\theta}_2|$$

Strength of network effects

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:

$$\begin{aligned}
 U_1(\theta_i) &= \alpha N_1 - |\theta_i - \bar{\theta}_1| + \Delta \\
 U_2(\theta_i) &= \alpha N_2 - |\theta_i - \bar{\theta}_2|
 \end{aligned}$$

# Users in the Platform      Average “Unsafety” of the Content  
 Strength of network effects      Quality Premium of the Moderated

User  $i$  joins (only!) the platform that maximizes their utility

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

$$\begin{aligned}
 U_1(\theta_i) &= \alpha N_1 - |\theta_i - \bar{\theta}_1| + \Delta \\
 U_2(\theta_i) &= \alpha N_2 - |\theta_i - \bar{\theta}_2|
 \end{aligned}$$

# Users in the Platform      Average “Unsafety” of the Content  
 Strength of network effects      Quality Premium of the Moderated

User  $i$  joins (only!) the platform that maximizes their utility

Rk: No outside option!



# Advertisers

# Advertisers

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

# Advertisers

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

Are **averse** to unsafe content

# Advertisers

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

Are **averse** to unsafe content

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

# Advertisers

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

Are **averse** to unsafe content

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

## Moderated Platform

- The **moderated** platform (1) chooses a **content moderation policy**

$K := K_1 \in [0,1]$ : perfectly and costlessly **bans any content**  $\theta_i > K$

# Advertisers

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

Are **averse** to unsafe content

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

## Moderated Platform

- The **moderated** platform (1) chooses a **content moderation policy**

$K := K_1 \in [0,1]$ : perfectly and costlessly **bans any content**  $\theta_i > K$

Platform (1) **maximizes** revenues:

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

# Advertisers

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

Are **averse** to unsafe content

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

## Moderated Platform

- The **moderated** platform (1) chooses a **content moderation policy**

$K := K_1 \in [0,1]$ : perfectly and costlessly **bans any content**  $\theta_i > K$

Platform (1) **maximizes** revenues:

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

# users in platform



# Advertisers

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

Are **averse** to unsafe content

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

## Moderated Platform

- The **moderated** platform (1) chooses a **content moderation policy**

$K := K_1 \in [0,1]$ : perfectly and costlessly **bans any content**  $\theta_i > K$

Platform (1) **maximizes** revenues:

$$\Pi(K) = \underbrace{N_1(K)}_{\substack{\text{\# users in platform}}} \times \underbrace{(1 - b\bar{\theta}_1(K))}_{\text{Price of ads}}$$

# Advertisers

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

Are **averse** to unsafe content

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

## Moderated Platform

- The **moderated** platform (1) chooses a **content moderation policy**

$K := K_1 \in [0,1]$ : perfectly and costlessly **bans any content**  $\theta_i > K$

Platform (1) **maximizes** revenues:

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 **regulated** platform (1)

Are **averse** to unsafe content

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

## Moderated Platform

- The **moderated** platform (1) chooses a **content moderation policy**

$K := K_1 \in [0,1]$ : perfectly and costlessly **bans any content**  $\theta_i > K$

Platform (1) **maximizes** revenues:

$$\Pi(K) = \underbrace{N_1(K)}_{\substack{\text{\# users in platform}}} \times \underbrace{\left(1 - \underbrace{b\bar{\theta}_1(K)}_{\substack{\text{Advertisers aversion} \\ \text{to unsafe content}}}\right)}_{\substack{\text{Price of ads}}} \underbrace{\quad}_{\substack{\text{Average content} \\ \text{unsafety}}}$$

# Advertisers

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

Are **averse** to unsafe content

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

## Moderated Platform

- The **moderated** platform (1) chooses a **content moderation policy**

$K := K_1 \in [0,1]$ : perfectly and costlessly **bans any content**  $\theta_i > K$

Platform (1) **maximizes** revenues:

$$\Pi(K) = \underbrace{N_1(K)}_{\substack{\text{\# users in platform}}} \times \underbrace{\left(1 - \underbrace{b\bar{\theta}_1(K)}_{\substack{\text{Advertisers aversion} \\ \text{to unsafe content}}}\right)}_{\substack{\text{Price of ads}}} \underbrace{\quad}_{\substack{\text{Average content} \\ \text{unsafety}}}$$

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

# Timing

# Timing

1. The moderated platform (1) chooses the content moderation policy  $K$  and commits to it

# Timing

1. The moderated platform (1) chooses the content moderation policy  $K$  and commits to it

# Timing

1. The moderated platform (1) chooses the content moderation policy  $K$  and commits to it
2. All the users simultaneously choose whether to join platform (1) *xor* (2) depending on their  $\theta_i$



# Timing

1. The moderated platform (1) chooses the content moderation policy  $K$  and commits to it
2. All the users simultaneously choose whether to join platform (1) *xor* (2) depending on their  $\theta_i$

# Timing

1. The moderated platform (1) chooses the content moderation policy  $K$  and commits to it
2. All the users simultaneously choose whether to join platform (1) *xor* (2) depending on their  $\theta_i$
3. Agents derive the corresponding payoffs from the composition of the social network

# Threshold Equilibrium (for given $K$ )

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

# Threshold Equilibrium (for given $K$ )

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

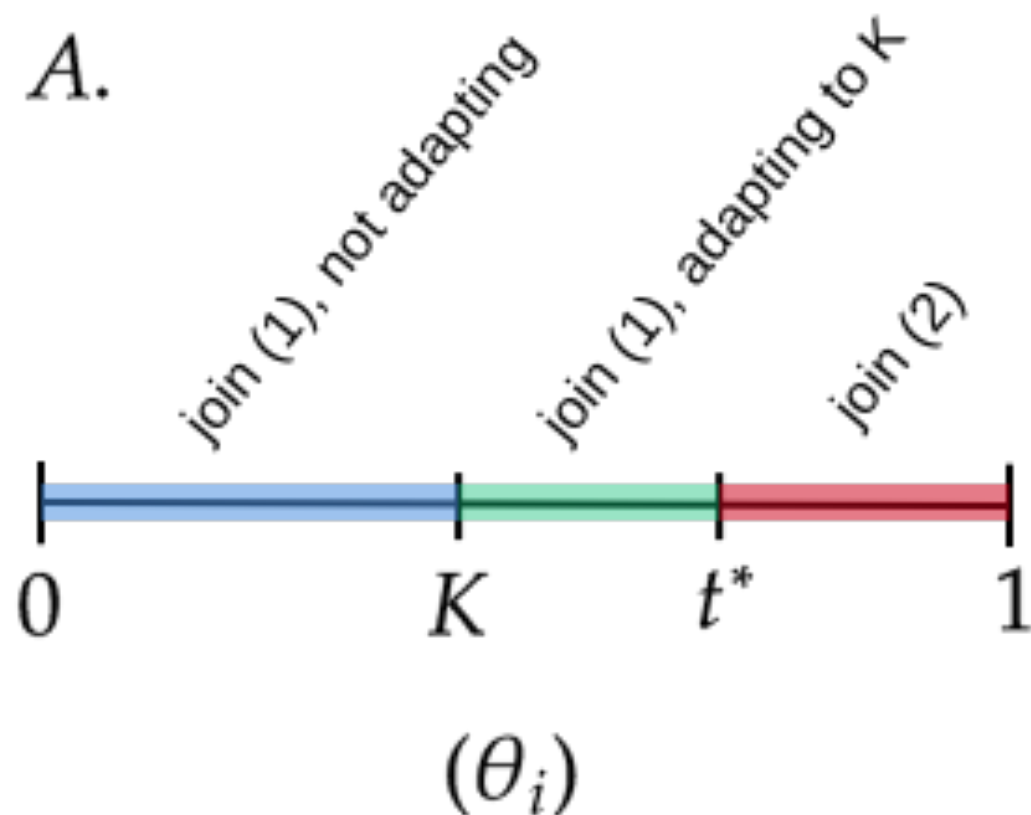
Under some assumptions on  $\alpha, \Delta$ ; for any  $K$ , there exist a **unique threshold equilibrium**, which takes one of these two forms:

# Threshold Equilibrium (for given $K$ )

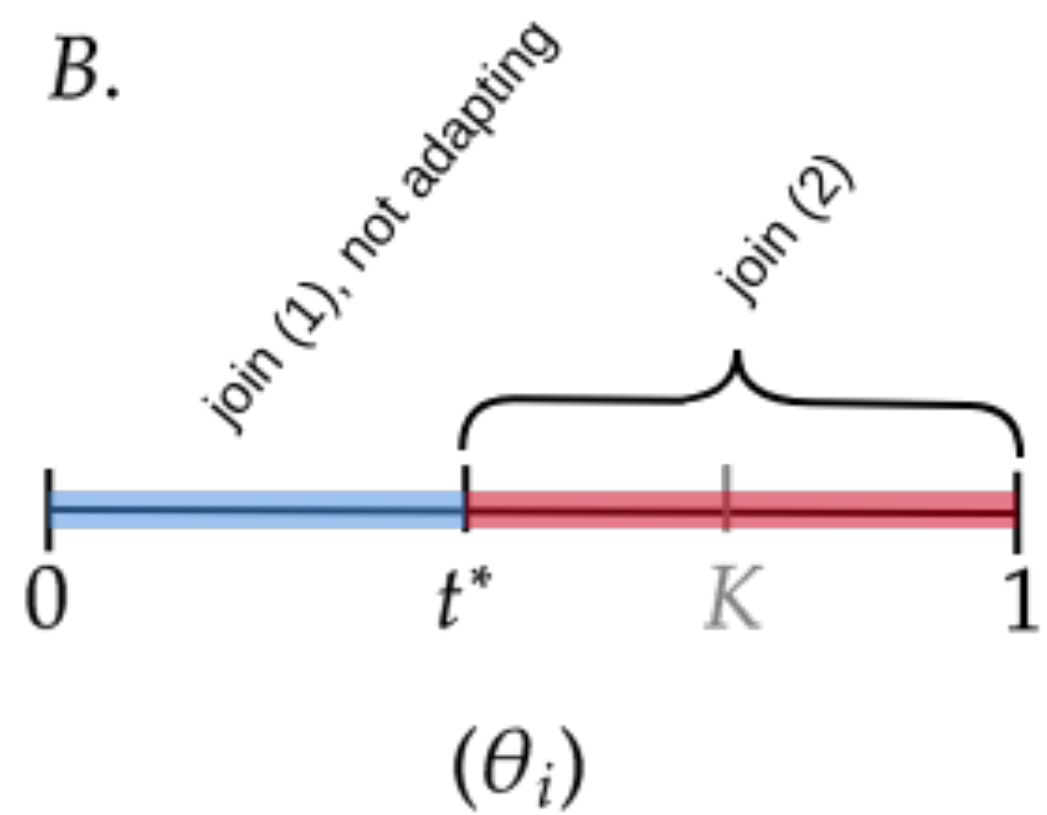
User  $i$  joins platform (1) iff  $\theta_i < t^*$ , otherwise, they join (2)

Under some assumptions on  $\alpha, \Delta$ ; for any  $K$ , there exist a **unique threshold equilibrium**, which takes one of these two forms:

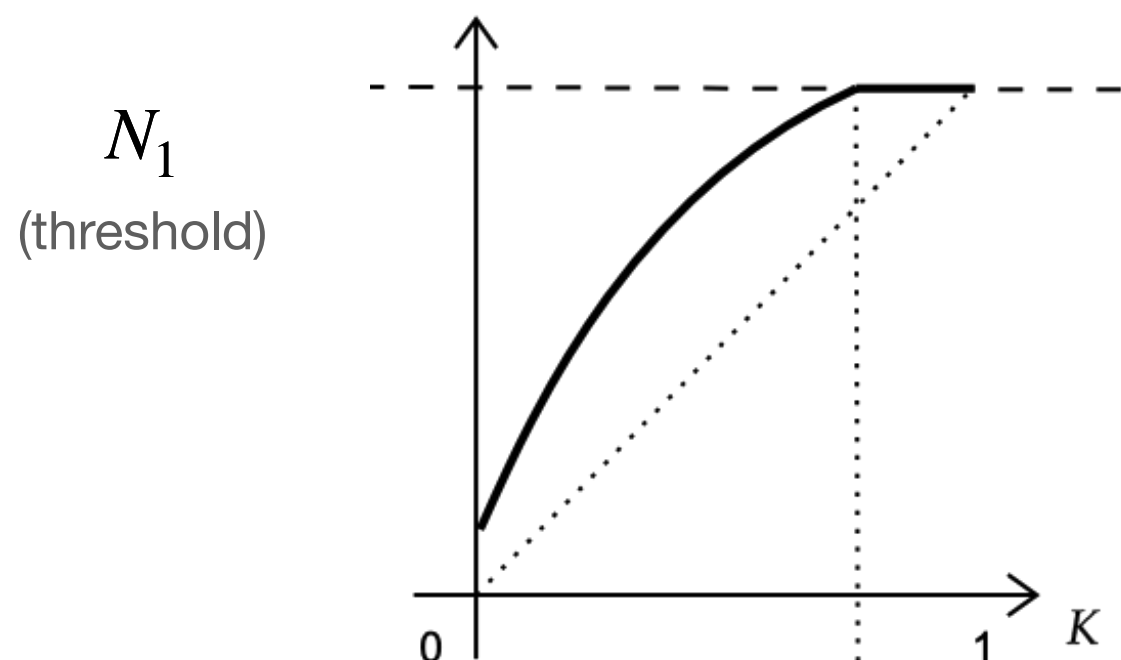
Low  $K$  (strict policy)



High  $K$  (lenient policy)



# Characterization of the Equilibrium



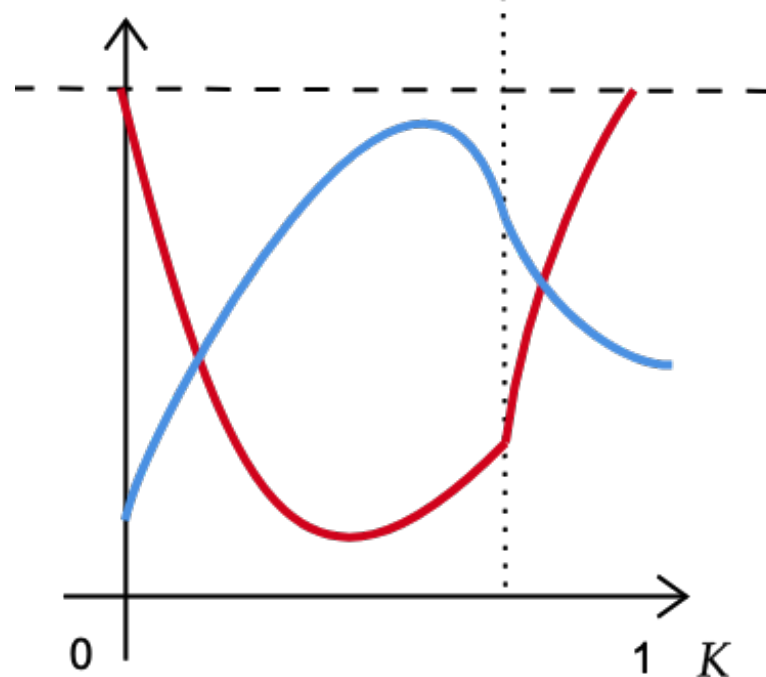
$\uparrow \alpha, \uparrow \Delta$

~ more attractive moderated platform (market could tip!)

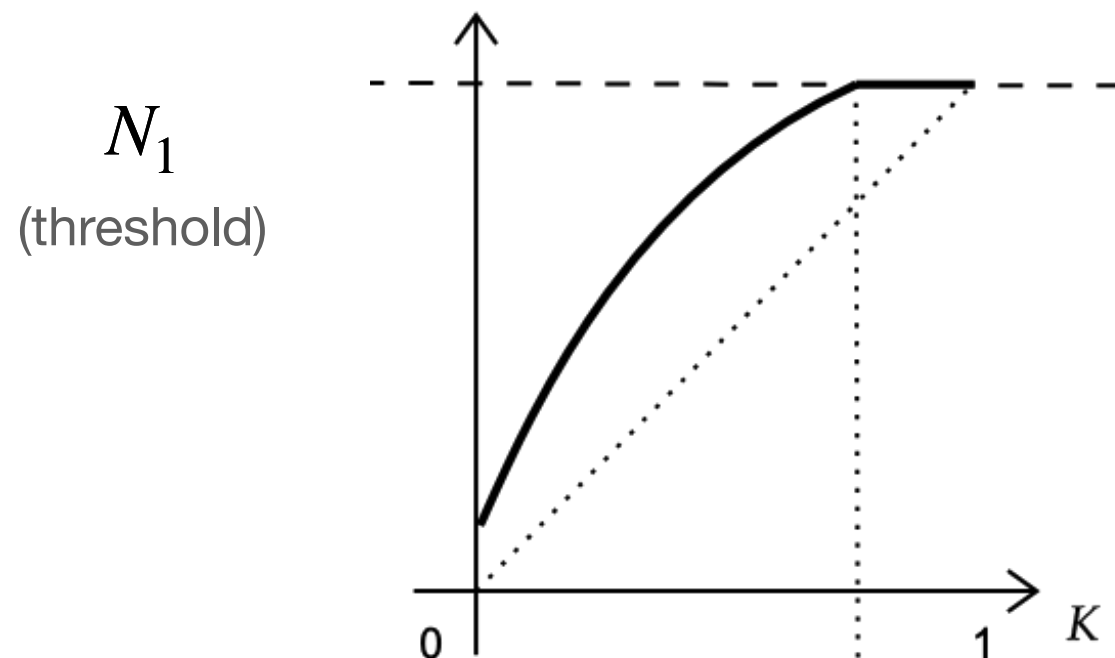
$$\frac{d^2 N_1}{dK d\alpha} > 0$$

Content moderation affects participation more with high network effects

Platform's  
profit  
Total unsafety  
Level



# Characterization of the Equilibrium



$\uparrow \alpha, \uparrow \Delta$

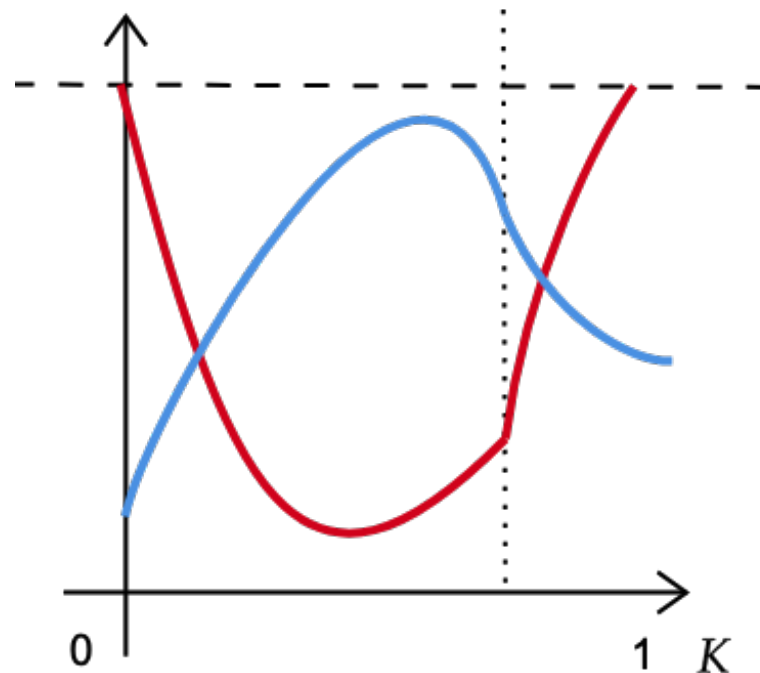
~ more attractive moderated platform (market could tip!)

$$\frac{d^2 N_1}{dK d\alpha} > 0$$

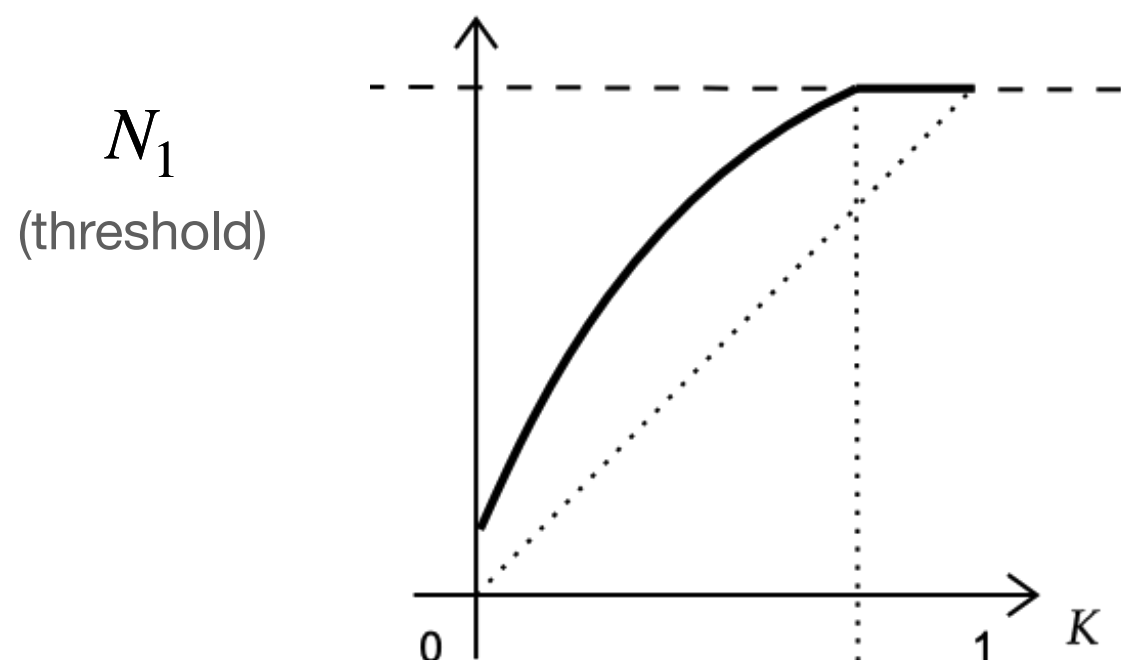
Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$

Platform's  
profit  
Total unsafety  
Level



# Characterization of the Equilibrium



$$\uparrow \alpha, \uparrow \Delta$$

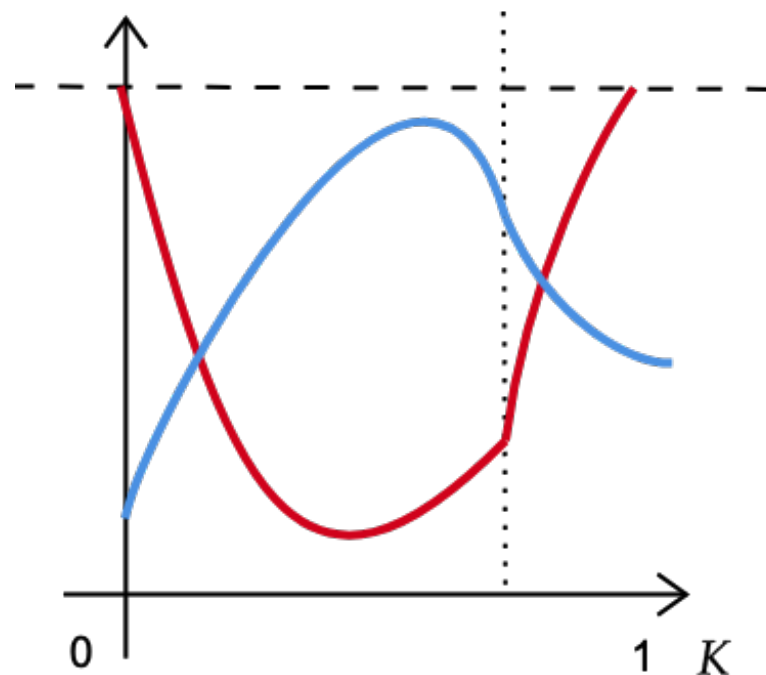
~ more attractive moderated platform (market could tip!)

$$\frac{d^2 N_1}{dK d\alpha} > 0$$

Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$

Platform's  
profit  
Total unsafety  
Level

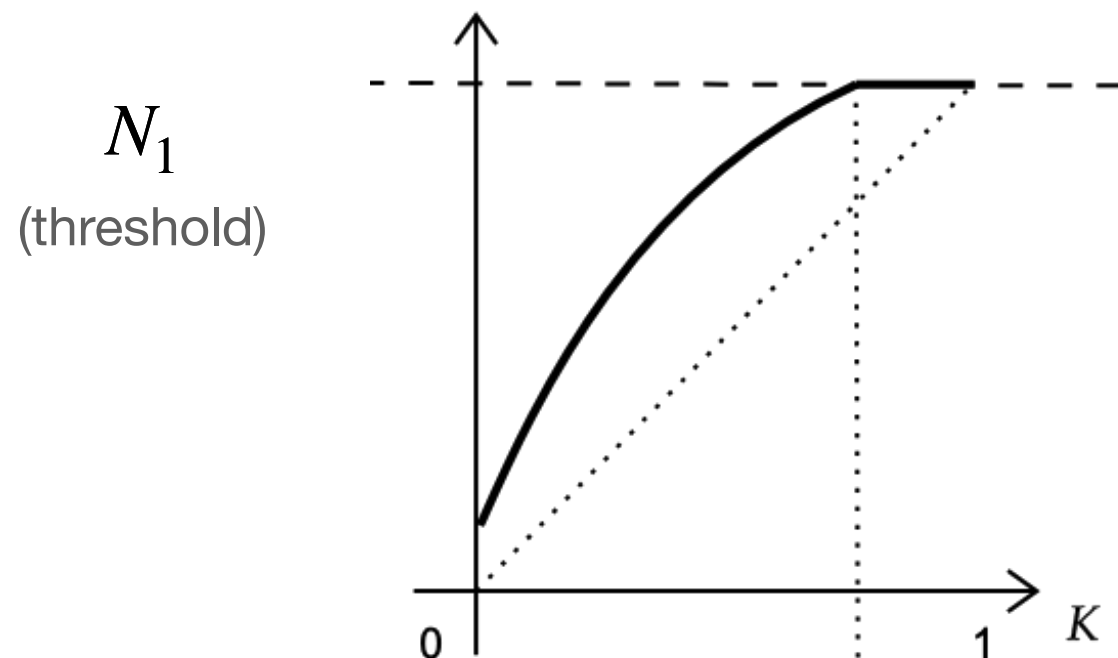


**In general, total unsafety non-monotonic!**

**Key:** mass of users willing to self-censor varies



# Characterization of the Equilibrium



$\uparrow \alpha, \uparrow \Delta$

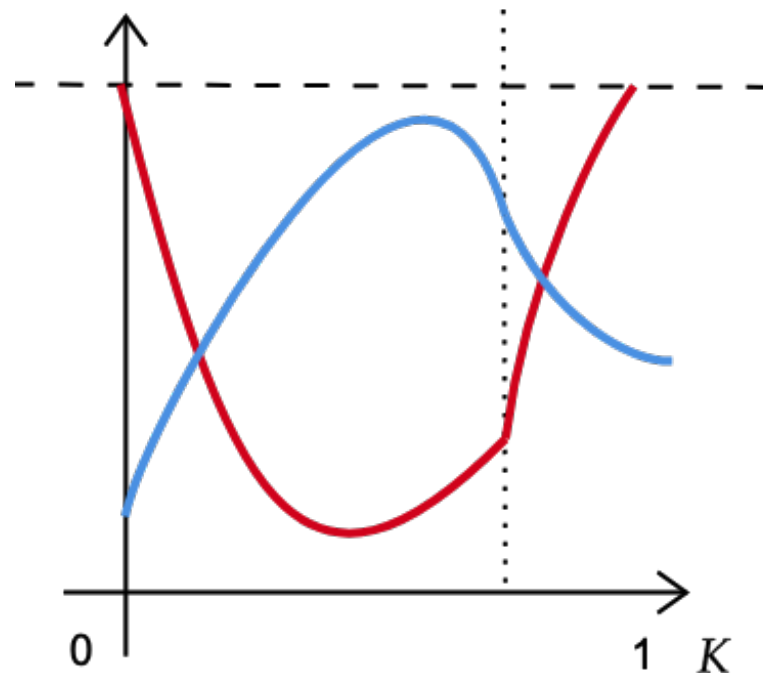
~ more attractive moderated platform (market could tip!)

$$\frac{d^2 N_1}{dK d\alpha} > 0$$

Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$

Platform's profit  
Total unsafety  
Level



**In general, total unsafety non-monotonic!**

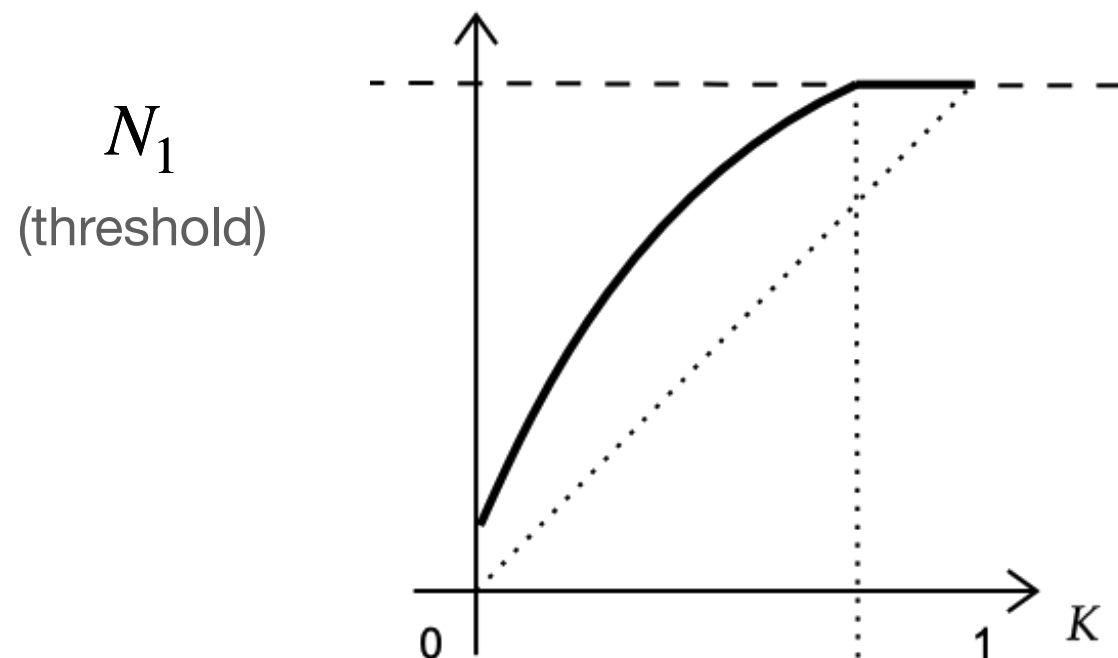
**Key:** mass of users willing to self-censor varies

**Incentives misalignment:**

**Regulator:** Unsafety (and participation) in both platforms

**Platform :** Unsafety and participation in his platform

# Characterization of the Equilibrium



$$\uparrow \alpha, \uparrow \Delta$$

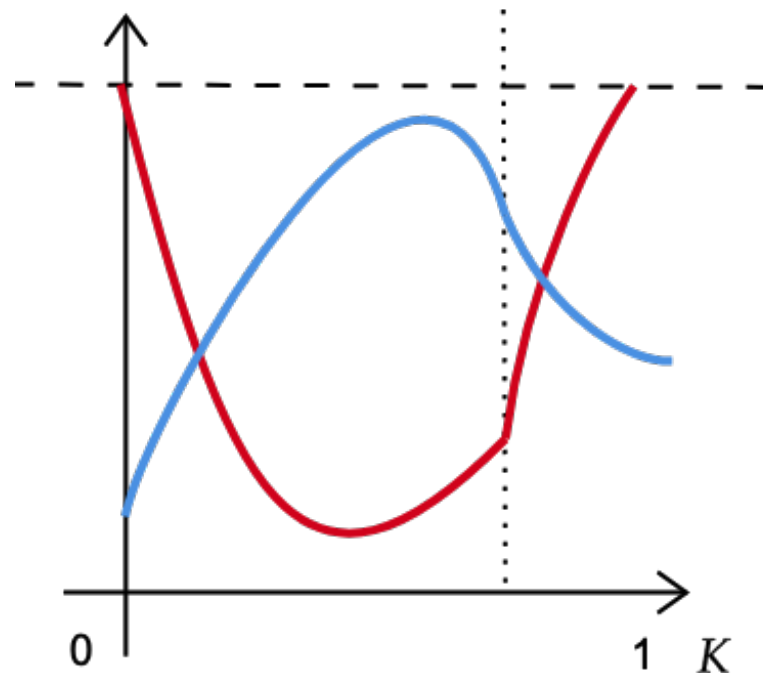
~ more attractive moderated platform (market could tip!)

$$\frac{d^2 N_1}{dK d\alpha} > 0$$

Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$

Platform's  
profit  
Total unsafety  
Level



**In general, total unsafety non-monotonic!**

**Key:** mass of users willing to self-censor varies

**Incentives misalignment:**

**Regulator:** Unsafety (and participation) in both platforms

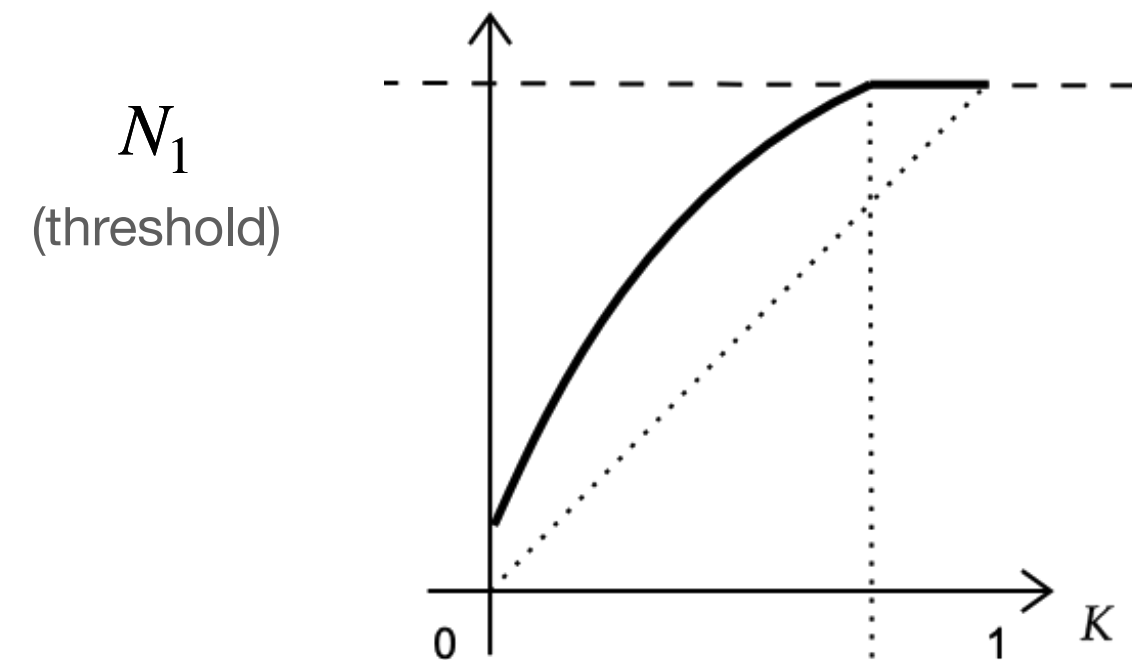
**Platform :** Unsafety and participation in his platform

**High Network Effects: Stricter or Lenient Moderation?**

**Regulator:** Lenient! wants users to self-censor IN the moderated platform

**Platform:** Stricter! Prefers smaller size and clean content due to advertisers

# Characterization of the Equilibrium



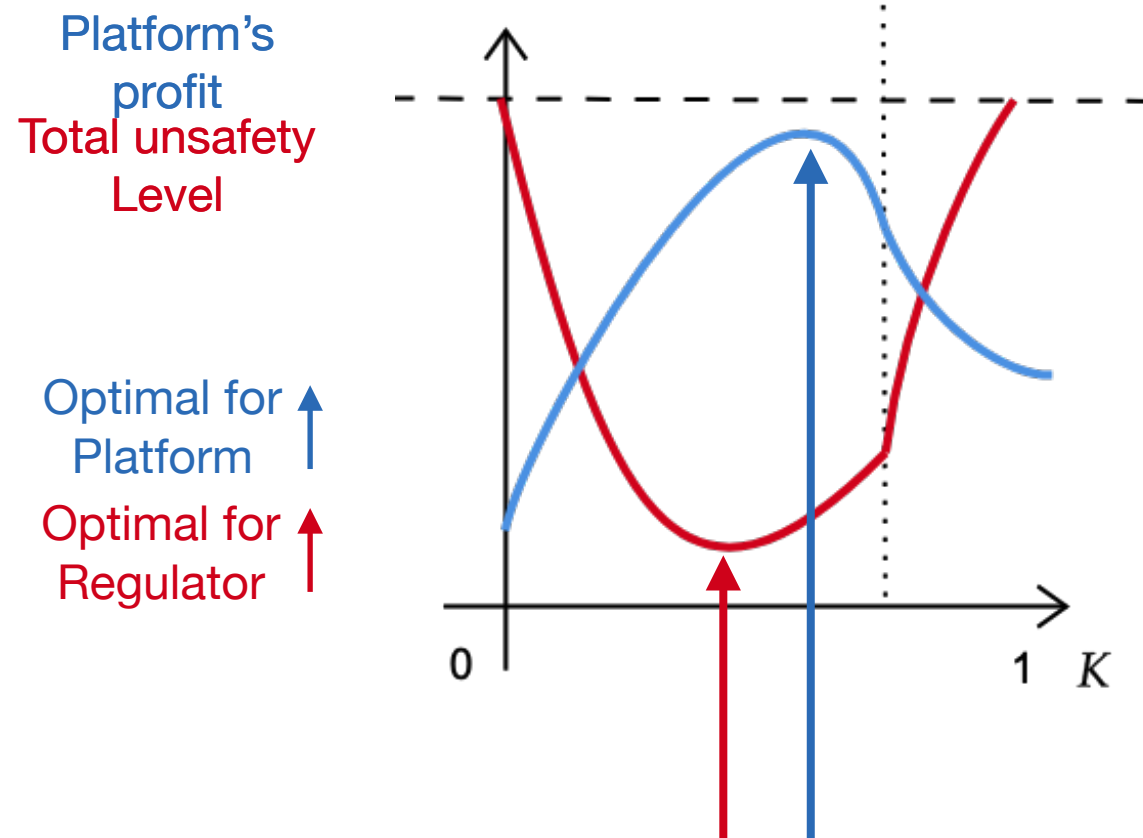
$$\uparrow \alpha, \uparrow \Delta$$

~ more attractive moderated platform (market could tip!)

$$\frac{d^2 N_1}{dK d\alpha} > 0$$

Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$



**In general, total unsafety non-monotonic!**

**Key:** mass of users willing to self-censor varies

**Incentives misalignment:**

**Regulator:** Unsafety (and participation) in both platforms

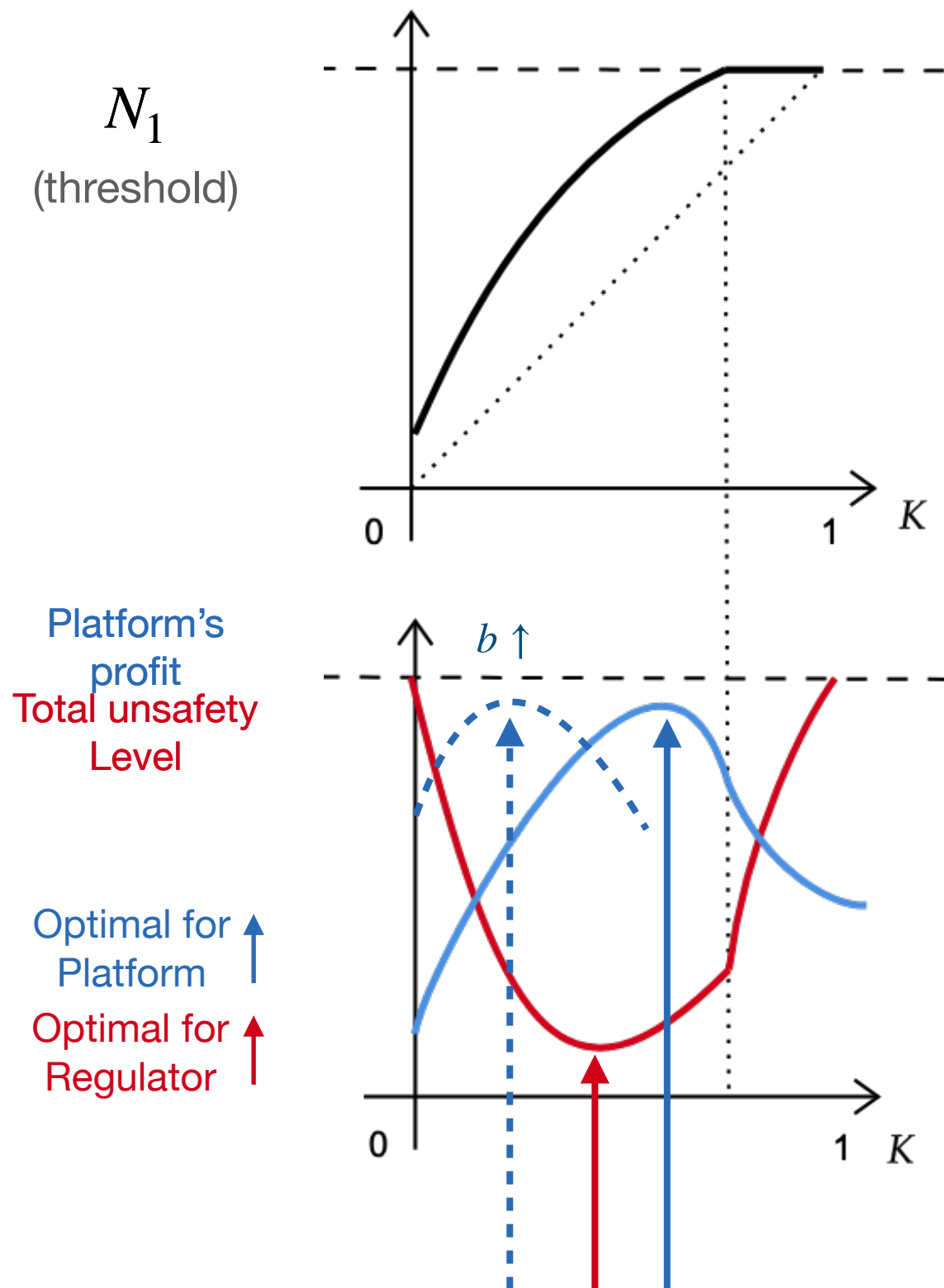
**Platform :** Unsafety and participation in his platform

**High Network Effects: Stricter or Lenient Moderation?**

**Regulator:** Lenient! wants users to self-censor IN the moderated platform

**Platform:** Stricter! Prefers smaller size and clean content due to advertisers

# Characterization of the Equilibrium



$$\uparrow \alpha, \uparrow \Delta$$

~ more attractive moderated platform (market could tip!)

$$\frac{d^2 N_1}{dK d\alpha} > 0$$

Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$

**In general, total unsafety non-monotonic!**

**Key:** mass of users willing to self-censor varies

**Incentives misalignment:**

**Regulator:** Unsafety (and participation) in both platforms

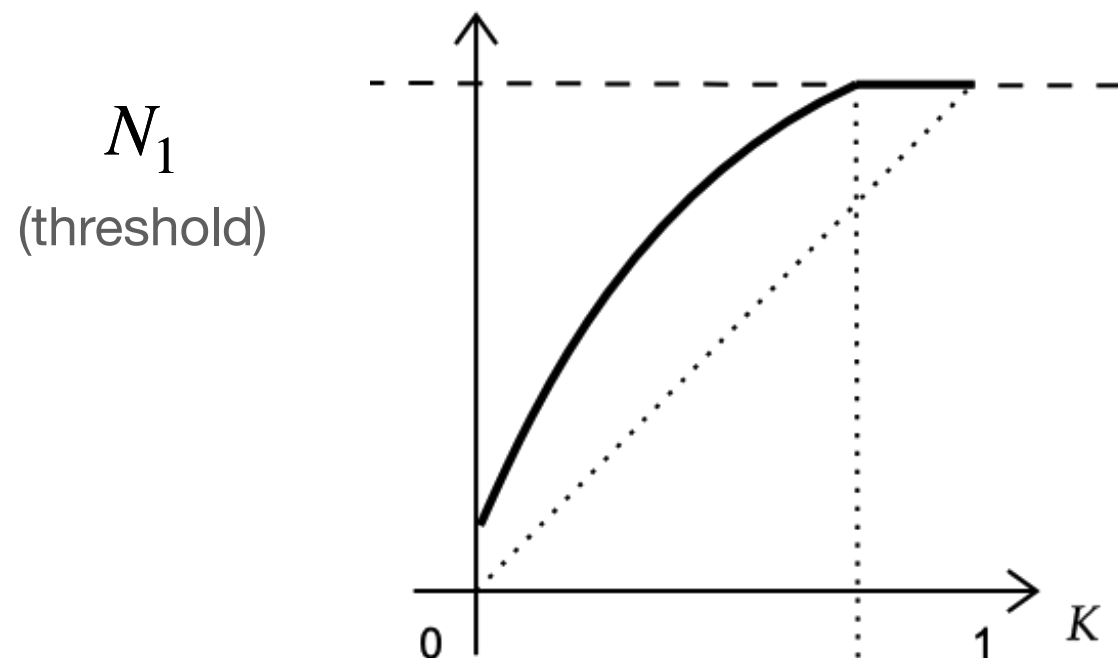
**Platform :** Unsafety and participation in his platform

**High Network Effects: Stricter or Lenient Moderation?**

**Regulator:** Lenient! wants users to self-censor IN the moderated platform

**Platform:** Stricter! Prefers smaller size and clean content due to advertisers

# Characterization of the Equilibrium



$$\uparrow \alpha, \uparrow \Delta$$

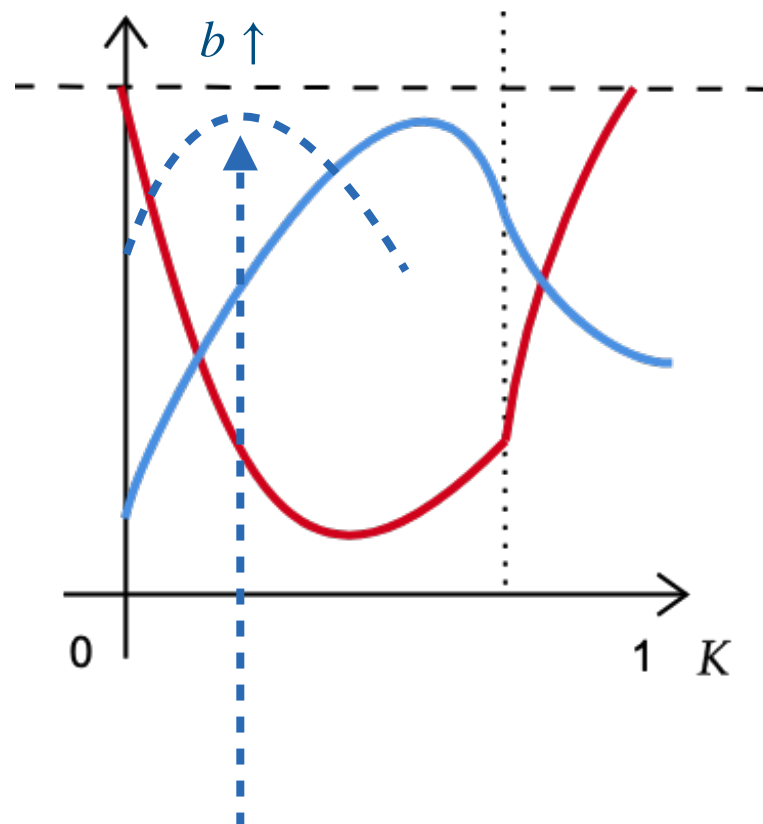
~ more attractive moderated platform (market could tip!)

$$\frac{d^2 N_1}{dK d\alpha} > 0$$

Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$

Platform's  
profit  
Total unsafety  
Level



**In general, total unsafety non-monotonic!**

**Key:** mass of users willing to self-censor varies

**Incentives misalignment:**

**Regulator:** Unsafety (and participation) in both platforms

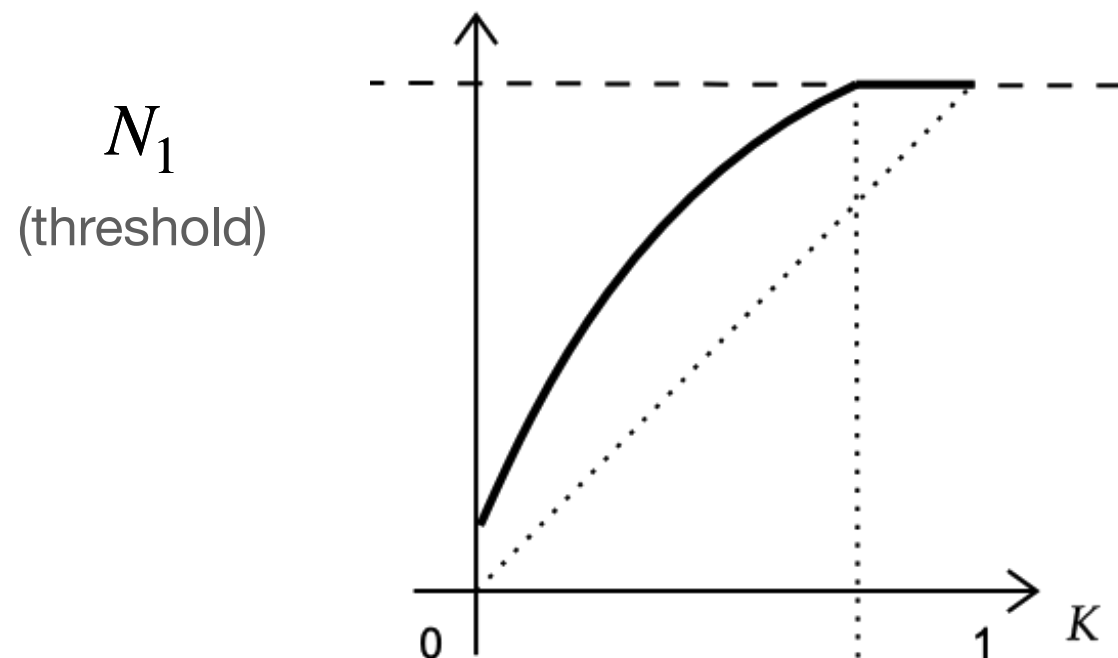
**Platform :** Unsafety and participation in his platform

**High Network Effects: Stricter or Lenient Moderation?**

**Regulator:** Lenient! wants users to self-censor IN the moderated platform

**Platform:** Stricter! Prefers smaller size and clean content due to advertisers

# Characterization of the Equilibrium



$$\uparrow \alpha, \uparrow \Delta$$

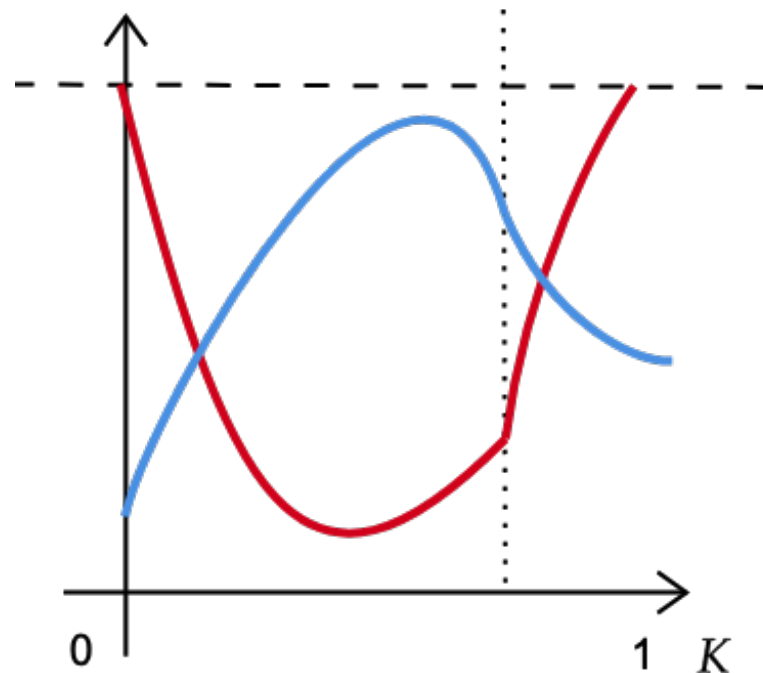
~ more attractive moderated platform (market could tip!)

$$\frac{d^2 N_1}{dK d\alpha} > 0$$

Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$

Platform's  
profit  
Total unsafety  
Level



**In general, total unsafety non-monotonic!**

**Key:** mass of users willing to self-censor varies

**Incentives misalignment:**

**Regulator:** Unsafety (and participation) in both platforms

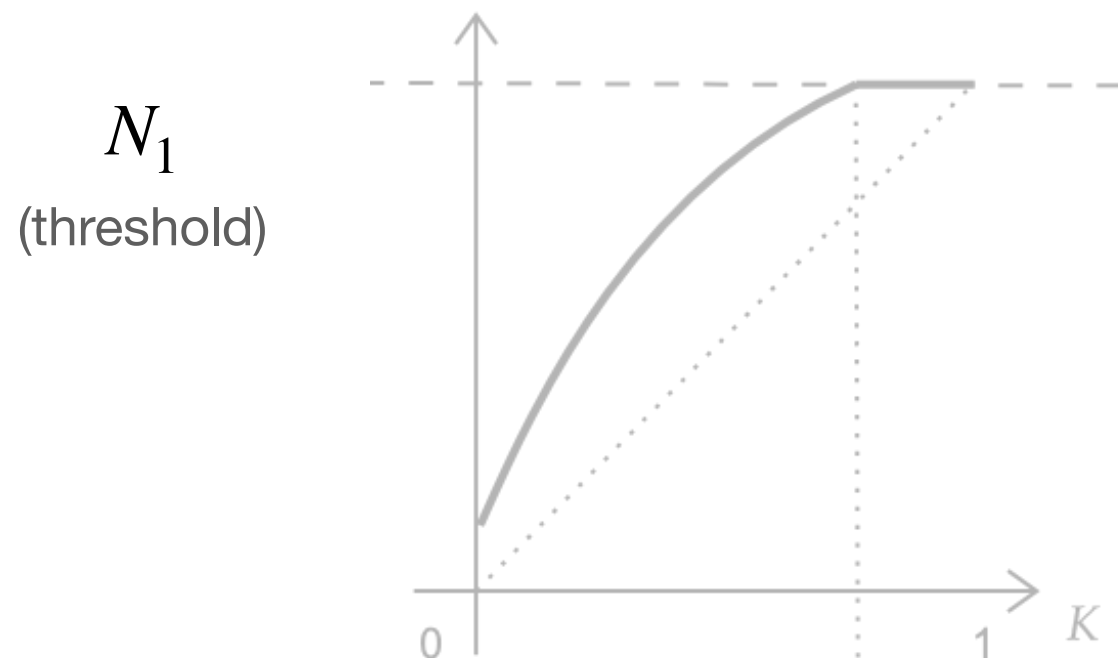
**Platform :** Unsafety and participation in his platform

**High Network Effects: Stricter or Lenient Moderation?**

**Regulator:** Lenient! wants users to self-censor IN the moderated platform

**Platform:** Stricter! Prefers smaller size and clean content due to advertisers

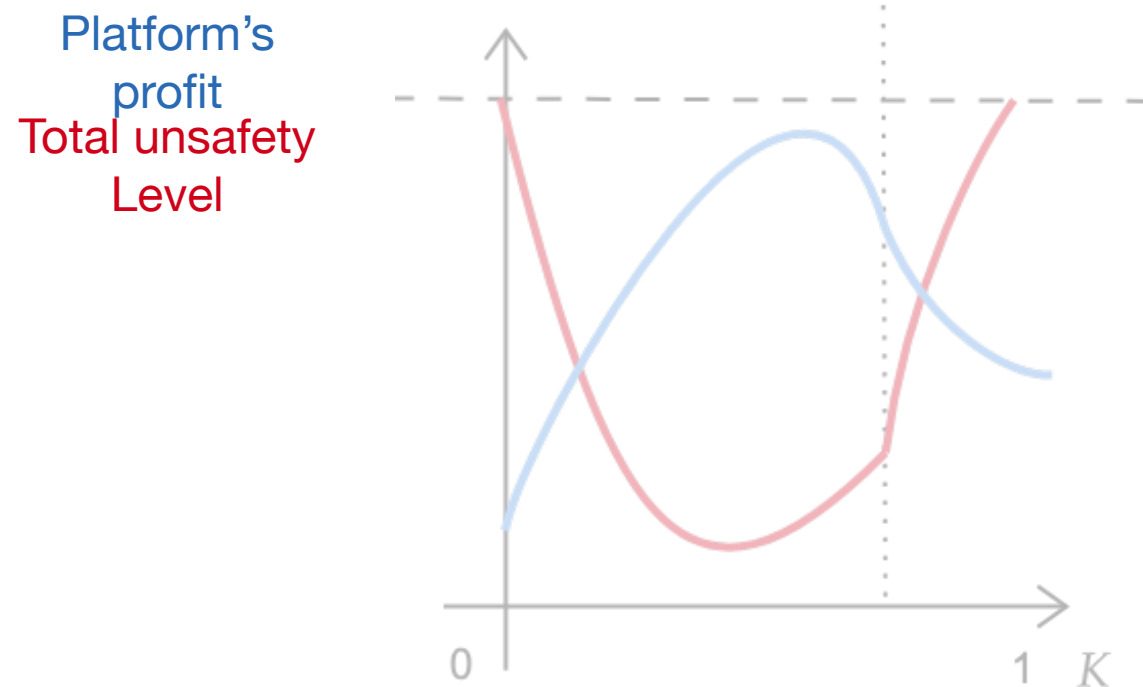
# Characterization of the Equilibrium



$\uparrow \alpha, \uparrow \Delta$  ~ more attractive moderated platform (market could tip!)

$\frac{d^2 N_1}{dK d\alpha} > 0$  Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$



**In general, total unsafety non-monotonic!**

**Key:** mass of users willing to self-censor varies

**Incentives misalignment:**

**Regulator:** Unsafety (and participation) in both platforms

**Platform :** Unsafety and participation in his platform

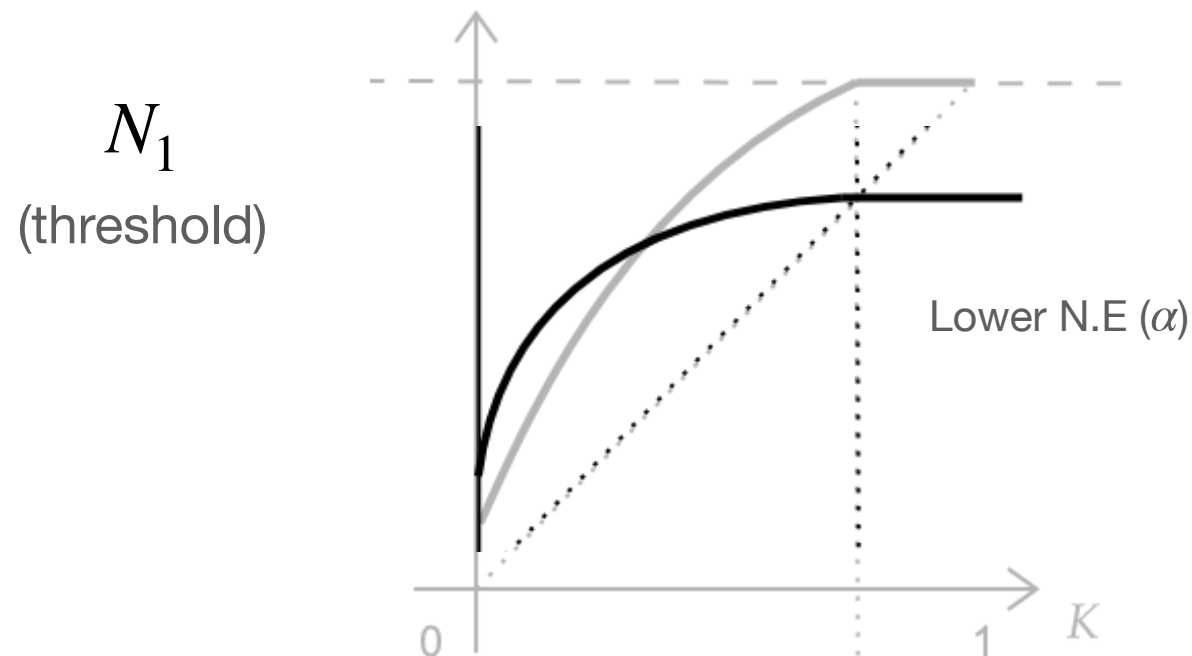
**High Network Effects: Stricter or Lenient Moderation?**

**Regulator:** Lenient! wants users to self-censor IN the moderated platform

**Platform:** Stricter! Prefers smaller size and clean content due to advertisers



# Characterization of the Equilibrium



$\uparrow \alpha, \uparrow \Delta$

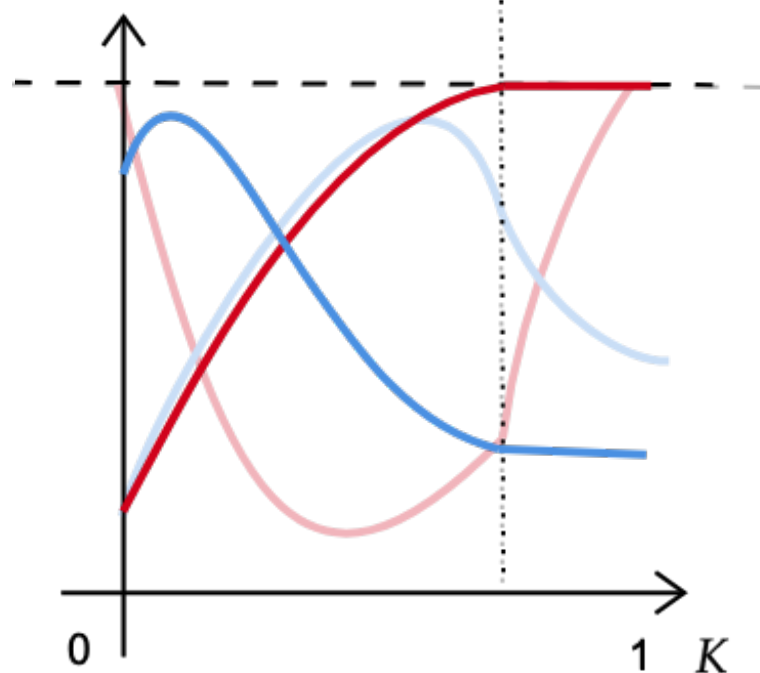
~ more attractive moderated platform (market could tip!)

$$\frac{d^2 N_1}{dK d\alpha} > 0$$

Content moderation affects participation more with high network effects

**Low network effects:** users do want to self-differentiate, irrespective of  $K$

Platform's  
profit  
Total unsafety  
Level



**In general, total unsafety non-monotonic!**

**Key:** mass of users willing to self-censor varies

**Incentives misalignment:**

**Regulator:** Unsafety (and participation) in both platforms

**Platform :** Unsafety and participation in his platform

**High Network Effects: Stricter or Lenient Moderation?**

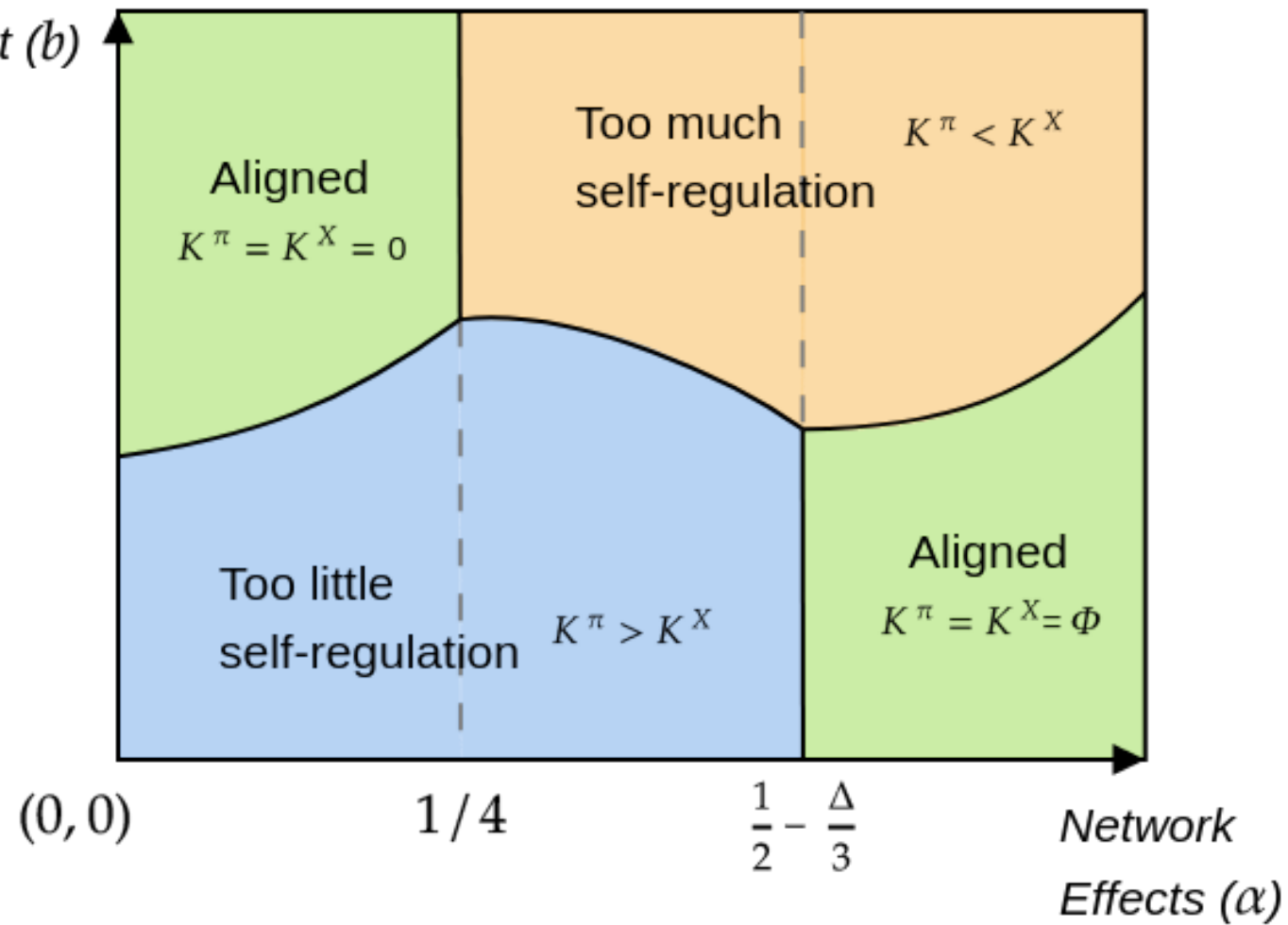
**Regulator:** Lenient! wants users to self-censor IN the moderated platform

**Platform:** Stricter! Prefers smaller size and clean content due to advertisers



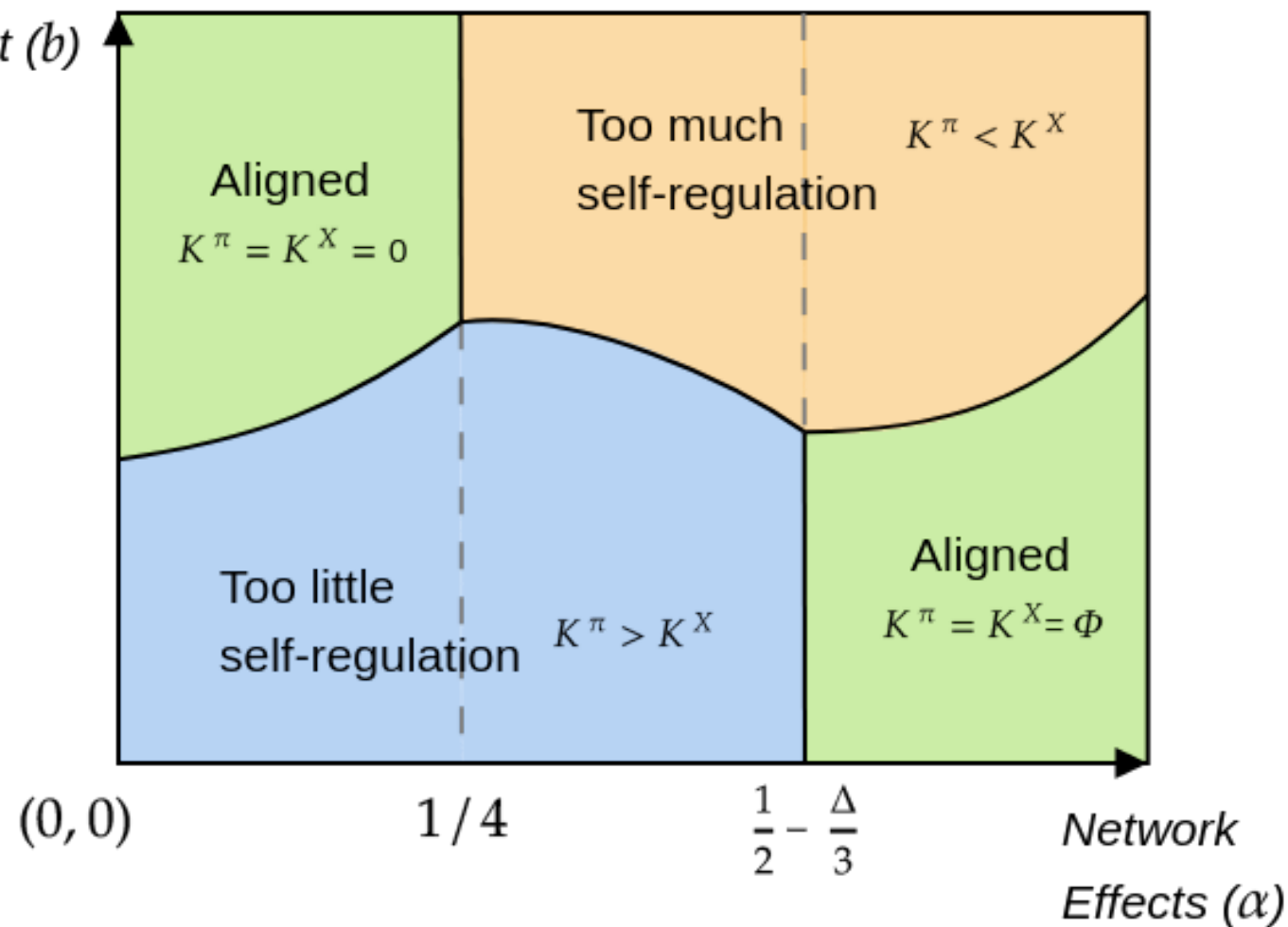
# Policy (to min unsafe content)

Advertisers aversion  
to unsafe content ( $b$ )



# Policy (to min unsafe content)

Advertisers aversion  
to unsafe content ( $b$ )

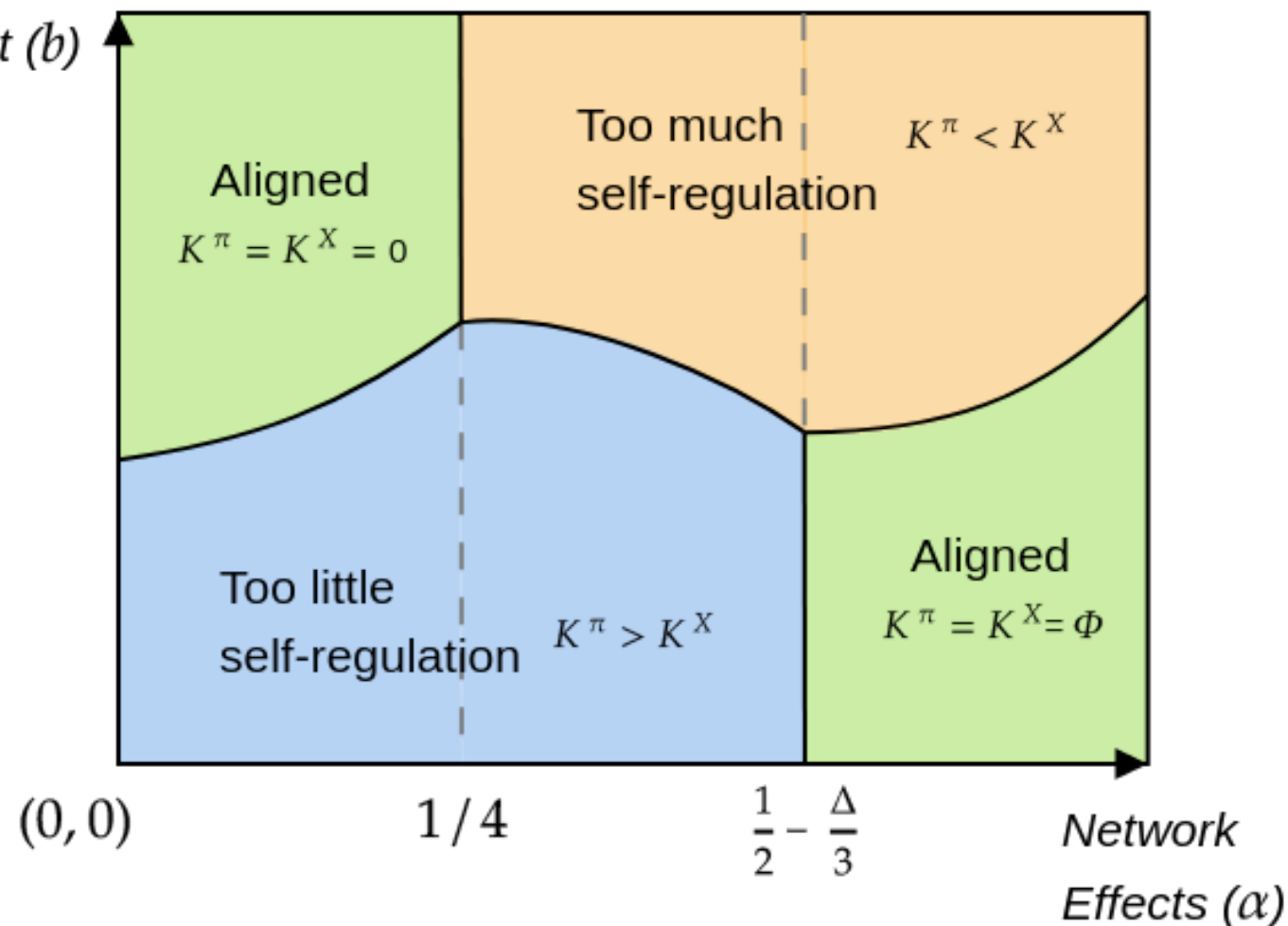


## Blue Area:

Beneficial for the regulator to  
impose a minimal  
moderation policy

# Policy (to min unsafe content)

Advertisers aversion  
to unsafe content ( $b$ )



## 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.

**(We saw this in the DSA)**

**EXTENSIONS**

# Multihoming

Allow for multihoming. User  $i$  utilities are:

- If single homing in  $j \in \{1,2\}$ :  $U_j^i$
- If multihoming :  $U_1^i + U_2^i$

# Multihoming

Allow for multihoming. User  $i$  utilities are:

- If single homing in  $j \in \{1,2\}$ :  $U_j^i$
- If multihoming :  $U_1^i + U_2^i$

**In equilibrium:**

Multihoming users  $\uparrow$  with strictness of content moderation and  $\downarrow$  with NE

Essentially users that otherwise singlehome in the moderated platform

# Multihoming

Allow for multihoming. User  $i$  utilities are:

- If single homing in  $j \in \{1,2\}$ :  $U_j^i$
- If multihoming:  $U_1^i + U_2^i$

## In equilibrium:

Multihoming users  $\uparrow$  with strictness of content moderation and  $\downarrow$  with NE

Essentially users that otherwise singlehome in the moderated platform

## General Result in the Literature:

Multihoming = Soften Network Effects = Increase Competition (here  $\sim \downarrow \Delta$ )

# Multihoming

Allow for multihoming. User  $i$  utilities are:

- If single homing in  $j \in \{1,2\}$ :  $U_j^i$
- If multihoming:  $U_1^i + U_2^i$

## In equilibrium:

Multihoming users  $\uparrow$  with strictness of content moderation and  $\downarrow$  with NE  
Essentially users that otherwise singlehome in the moderated platform

## General Result in the Literature:

Multihoming = Soften Network Effects = Increase Competition (here  $\sim \downarrow \Delta$ )

## Here... same!

Multihoming = Soften Network Effects = Increase Competition =  
 $\downarrow$  Incentives of the platform to moderate content



# Multihoming

Allow for multihoming. User  $i$  utilities are:

- If single homing in  $j \in \{1,2\}$ :  $U_j^i$
- If multihoming:  $U_1^i + U_2^i$

## In equilibrium:

Multihoming users  $\uparrow$  with strictness of content moderation and  $\downarrow$  with NE  
Essentially users that otherwise singlehome in the moderated platform

## General Result in the Literature:

Multihoming = Soften Network Effects = Increase Competition (here  $\sim \downarrow \Delta$ )

## Here... same!

Multihoming = Soften Network Effects = Increase Competition =  
 $\downarrow$  Incentives of the platform to moderate content

## Also here

**High NE:** unsafety min with a **more lenient** moderation wrt single-homing  
**Low NE:** unsafety min with a **stricter** moderation wrt single-homing

# Radicalization and Offline Violence

- Online content unsafety is considered bad *per se*
- Its **offline consequences** are a first order concern for regulators

# Radicalization and Offline Violence

- Online content unsafety is considered bad *per se*
- Its **offline consequences** are a first order concern for regulators
- This extension adds two periods to the model seen:

# Radicalization and Offline Violence

- Online content unsafety is considered bad *per se*
- Its **offline consequences** are a first order concern for regulators
- This extension adds two periods to the model seen:
  - i. In  $t=3$ , users' preferences either:

# Radicalization and Offline Violence

- Online content unsafety is considered bad *per se*
- Its **offline consequences** are a first order concern for regulators
- This extension adds two periods to the model seen:
  - i. In  $t=3$ , users' preferences either:
    - **Converge** to the unsafety of the content they read, OR

# Radicalization and Offline Violence

- Online content unsafety is considered bad *per se*
- Its **offline consequences** are a first order concern for regulators
- This extension adds two periods to the model seen:
  - i. In  $t=3$ , users' preferences either:
    - **Converge** to the unsafety of the content they read, OR
    - **Diverge** from it

# Radicalization and Offline Violence

- Online content unsafety is considered bad *per se*
- Its **offline consequences** are a first order concern for regulators
- This extension adds two periods to the model seen:
  - i. In  $t=3$ , users' preferences either:
    - **Converge** to the unsafety of the content they read, OR
    - **Diverge** from it
  - ii) In  $t=4$ , users perpetrate a unit of violence with a probability

# Radicalization and Offline Violence

- Online content unsafety is considered bad *per se*
- Its **offline consequences** are a first order concern for regulators
- This extension adds two periods to the model seen:
  - i. In  $t=3$ , users' preferences either:
    - **Converge** to the unsafety of the content they read, OR
    - **Diverge** from it
  - ii) In  $t=4$ , users perpetrate a unit of violence with a probability
    - **Increasing** in their taste for unsafety



# Radicalization and Offline Violence

- Online content unsafety is considered bad *per se*
- Its **offline consequences** are a first order concern for regulators
- This extension adds two periods to the model seen:
  - i. In  $t=3$ , users' preferences either:
    - **Converge** to the unsafety of the content they read, OR
    - **Diverge** from it
  - ii) In  $t=4$ , users perpetrate a unit of violence with a probability
    - **Increasing** in their taste for unsafety
    - **Decreasing** in their taste for unsafety (i.e. substitutes, *video games*)

# Radicalization and Offline Violence

- Online content unsafety is considered bad *per se*
- Its **offline consequences** are a first order concern for regulators
- This extension adds two periods to the model seen:
  - i. In  $t=3$ , users' preferences either:
    - **Converge** to the unsafety of the content they read, OR
    - **Diverge** from it
  - ii) In  $t=4$ , users perpetrate a unit of violence with a probability
    - **Increasing** in their taste for unsafety
    - **Decreasing** in their taste for unsafety (i.e. substitutes, *video games*)

## Main Result:

- With **converging** preferences + and violence **increasing** in unsafety, then:
  - **Intermediate levels of moderation are preferable to min violence**
  - **Why?** We can attract users to the moderated platform to read content safer than it would be without content moderation

# CONCLUSION

# Conclusion

**Today:**

**Future:**

# Conclusion

## Today:

- Simple model, simple intuition, policy oriented:
  - ➔ DSA may have some unintended consequences
- It gets worse with more competition (DMA!)
- Non very tractable but still analytical
- Consumer surplus, another platform...  
in the paper

## Future:

# Conclusion

## Today:

- Simple model, simple intuition, policy oriented:
  - ➔ DSA may have some unintended consequences
- It gets worse with more competition (DMA!)
- Non very tractable but still analytical
- Consumer surplus, another platform...  
in the paper

## Future:

- Empirics!
- Is the model right. If yes, where are we?
- I do have data and a draft of a structural model
- Not my field, (will never be?)

# Conclusion

## Today:

- Simple model, simple intuition, policy oriented:
  - ➔ DSA may have some unintended consequences
- It gets worse with more competition (DMA!)
- Non very tractable but still analytical
- Consumer surplus, another platform... in the paper

## Future:

- Empirics!
- Is the model right. If yes, where are we?
- I do have data and a draft of a structural model
- Not my field, (will never be?)
- Open to collaborations:  
[ivan.rendo@tse-fr.eu](mailto:ivan.rendo@tse-fr.eu)