Content Moderation in Presence of Fringe Platforms

Iván Rendo (TSE)

ENTER Jamboree

Brussels, 4 June 2024

Motivation

- Increased interest in **online** hateful/extreme/**unsafe content**:
  - e.g. Jiménez-Durán (2022) links online hate to **offline violence**
  - e.g. **25%** of terrorists radicalized **exclusively** online
  - e.g. bullying, food disorders…

➡ EU Response: **Digital Services Act**

- Duch-Brown’s view: **Constant unsafe content** across time

  ➡ BUT more **impact** today: everyone in the same large platforms

- Rizzi (2023)

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

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

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** 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

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
  - Moderated platform's trade-off: participation, unsafe content

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

- 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 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 unsafety $\bar{\theta}_j$

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

= average unsafety 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

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

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

- 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

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

Average content unsafety

Price of ads

# users in platform

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

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 θ_i

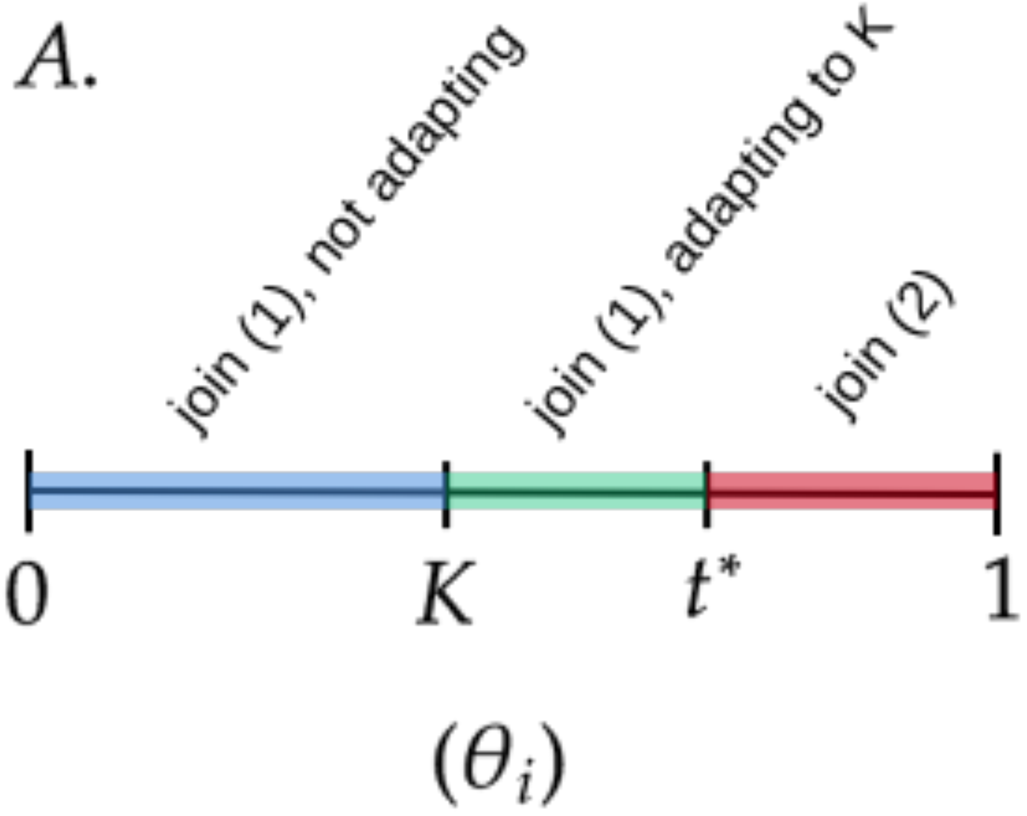
3. Agents derive the corresponding payoffs from the composition of the social network

Threshold Equilibrium (for given K)

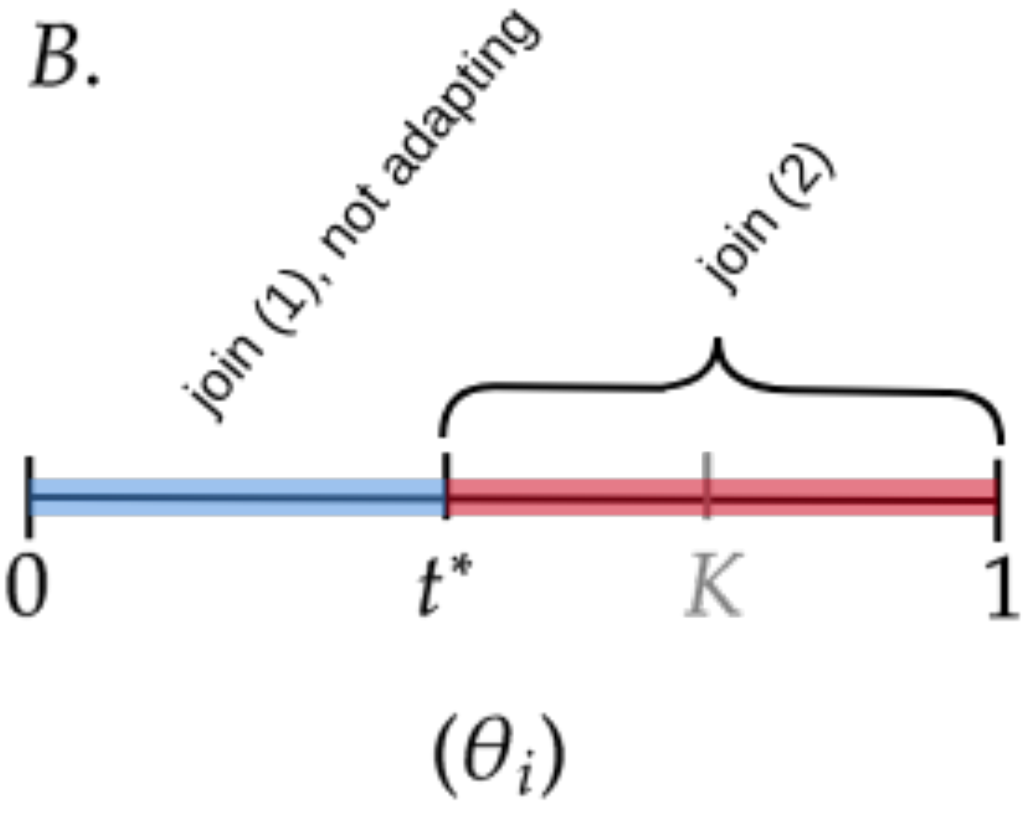
User i joins platform (1) iff $\theta_i < t^*$, otherwise, they join (2)

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

Low K (strict policy)

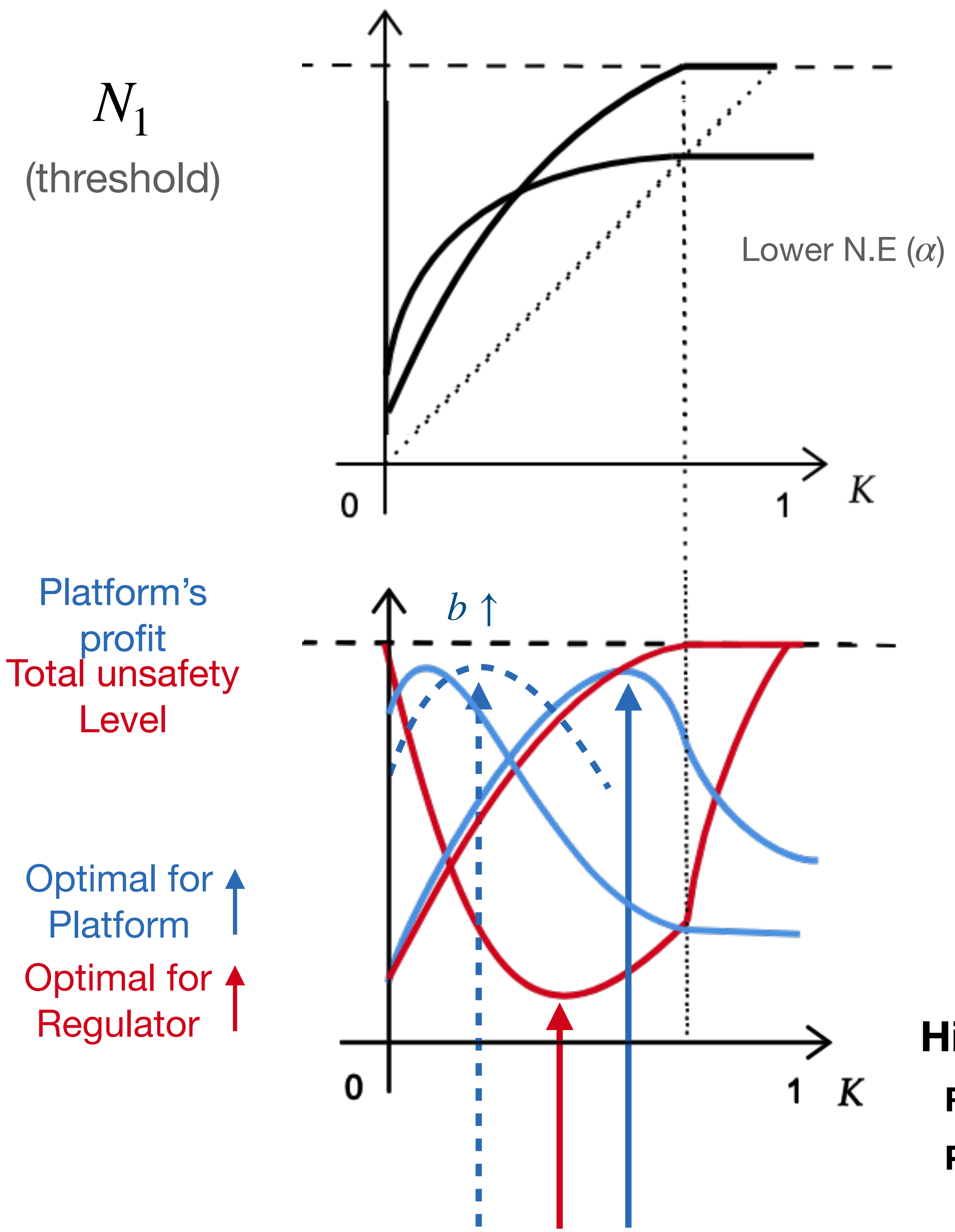


High K (lenient policy)



K=0, very strict content moderation
K=1, no content moderation

Characterization of the Equilibrium



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

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

Content moderation affects participation more with high network effects

In general, total unsafety non-monotonic!

Key: mass of users willing to self-censor varies

Incentives misalignement:

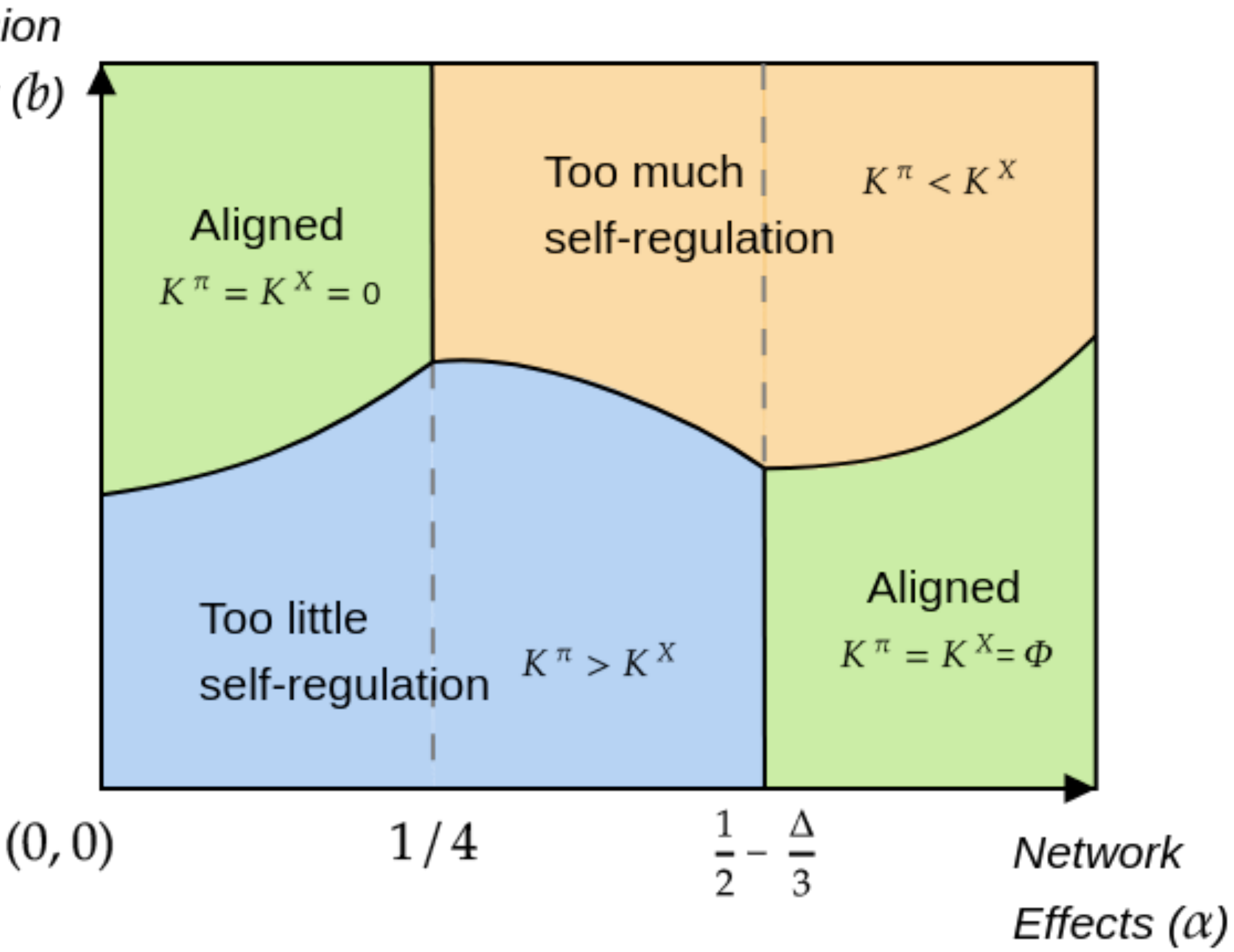
Regulator: Unsafety (and participation) in both platforms
Platform : Unsafety and participation in its platform

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)



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 : $U_1^i + U_2^i$

In equilibrium:

Multihoming users ↑ with strictness of content moderation and ↓ with NE

Essentially users that otherwise singlehome in the moderated platform

General Result in the Literature:

Multihoming = Soften Network Effects = Increase Competition

Here… same!

Multihoming = Soften Network Effects = Increase Competition =

↓ 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

- 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:

- 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 suggestions: ivan.rendo@tse-fr.eu