Content Moderation and Migration in Social Media: Evidence from Musk's Twitter Acquisition

Iván Rendo (TSE)

IO Workshop

Toulouse, 18 December 2023

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)

- Different complementary views on content moderation:
    - "Old Internet" - Duch-Brown's perspective:

      ➡ **Constant unsafe content** across time BUT today **good and bad people together**
    - Lefouili & Madio (2022): migration = ↓ impact and enforcement costs
    - Anti Defamation League (ADL) viral video: trading-off **moderation** in Twitter and **migration** to other (hateful, small) environments

Today

Platforms’ competition model to analyze the interaction between:

Content Moderation, Content (Un)safety, **Migration** (to other platforms)

… for an ad-funded platform

➡How **migration** is affected by content moderation **policies**

➡ How **unsafe content** is affected by **migration**

➡ What **incentives** do the platforms have to **self-regulate**

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

+ **Empirical evidence** through Musk’s acquisition of Twitter

Main Features of the Model

Users:

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

2 Asymmetric **Platforms**:

Twitter

- A **Regulated** one, higher quality platform: **moderates (bans) content**
  - Maximizes profits from **advertisers** (**averse to unsafe content** = pay less)
- An **Unregulated** one, lower quality platform: **no content moderation**

8Chan

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

Preview of the Main Results

1. **Prevalence of unsafe content:**

   i. **U-shaped** function of moderation intensity, w large network effects

   ii. **Decreasing** in moderation intensity in, w small network effects

2. **Policy:**

   - **Incentives misalignment** between platform & regulator (min unsafe content)
   - Imposing a **minimal** content moderation intensity (policy):

     i. W Large network effects: always **superfluous**

     ii. w Mid to small network effects: can be **useful**

Roadmap

I. Theoretical Model

- Characterization of the Equilibrium
- Optimal Regulation

II. Empirical Evidence

THEORY

Model

- A unit mass of **individuals**, 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** $(K_2 = 1)$

- Individual i in platform j **creates** 1 unit of content of unsafety θ_i^C

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

- Each individual i in platform j **reads** all the content, of avg unsafety $\bar{\theta}_j$

$$\bar{\theta}_j = \frac{1}{N_j} \sum_{i \in j} \theta_i^C$$

= average unsafety of content in platform j

- Platform 1, **regulated**, is intrinsically better than 2, **unregulated**

- Utilities of user i joining $j = 1,2$ are defined as:

# Users in the Platform

Average "Unsafety" of the Created Content

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

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

Intrinsic Quality of the Reg. Platform

Inverse 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 **regulated** platform (1)

Are **averse** to unsafe content

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

Regulated Platform

- The **regulated** 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:

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

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

Timing

1. The regulated 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

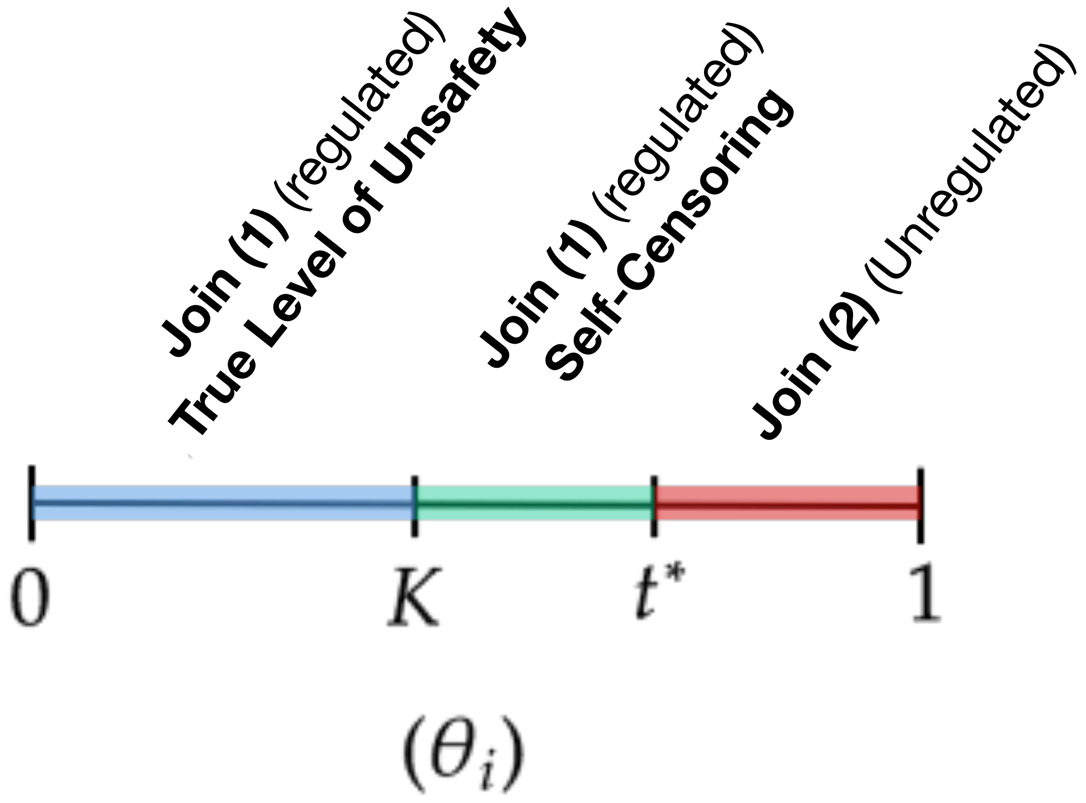
3. Agents derive the corresponding payoffs from the composition of the social network

Threshold Equilibrium

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

Under some conditions on α (not too low), for any K, there exist a **unique threshold equilibrium**, which takes one of these two forms:

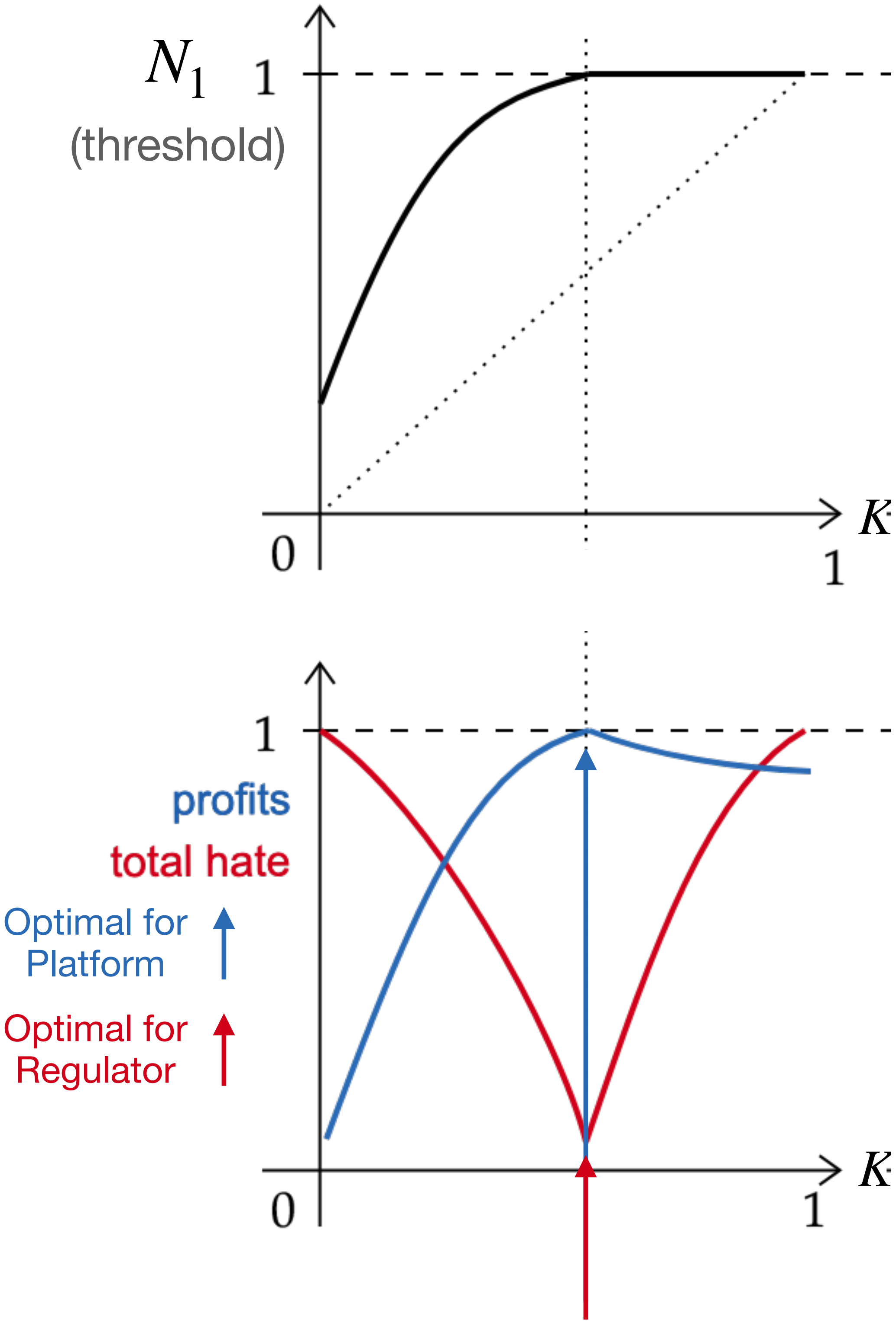
$K < t*$



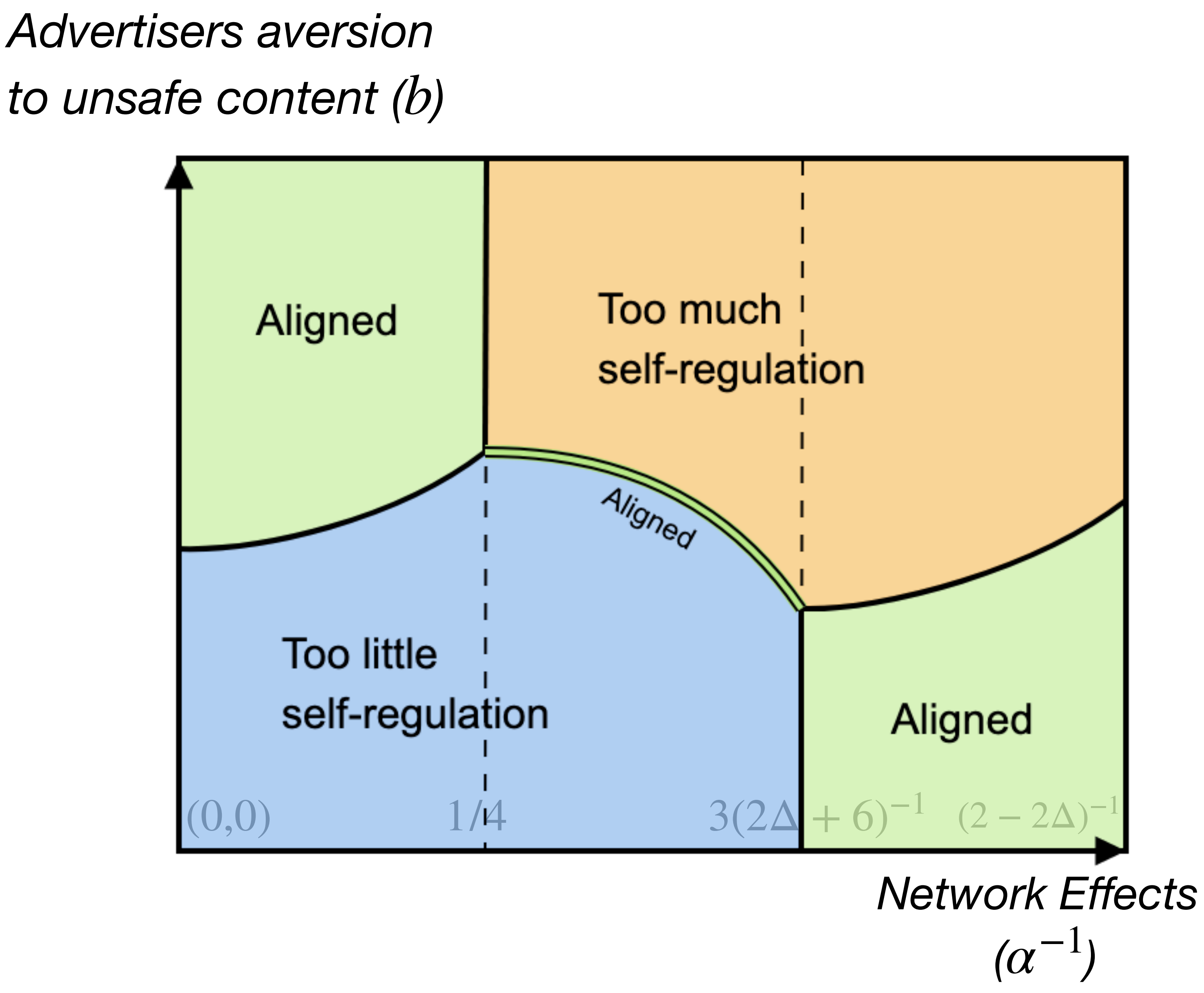
K=0, very strict content moderation
K=1, no content moderation

Characterization of the Equilibrium

Strong network effects



Policy (to min unsafe content)


Advertisers aversion to unsafe content (b)
Aligned
Too much self-regulation
Aligned
Too little self-regulation
Aligned
(0,0)
1/4
$3(2\Delta + 6)^{-1}$
$(2 - 2\Delta)^{-1}$
Network Effects (α^{-1})

Green Area: Nothing to do!

Blue Area:
The regulator can impose a minimum content moderation level, and it would be beneficial: **there won't be too much migration**

Orange Area: the policy wouldn't bind as the minimum content imposed is higher than the optimal for the platform

(We saw this in the DSA)

EMPIRICS

Event: Musk buys Twitter: *exogenous* $\uparrow K$

Hypothesis to take to the Data (from the model)

1. More unsafe content in Twitter. *Hickey et al.* (2023)
2. More ‘hate’ from ‘hateful users’. *Hickey et al.* (2022)
3. “Migration” from Telegram to TW from creators of unsafe content:
   1. Hateful for Twitter standards
   2. Decrease of unsafe content in Telegram from these users

Today

(4). Total unsafe content increases or decreases?

Review of the Data I Have:

12 million tweets around the invasion of Ukraine

- Checked if created by a “**Telegram User**”
- Computed “**toxicity**” levels of a sample of >100k of them using a *extremely* good Google API (*Perspective*)

Example

In terms of *toxicity*:

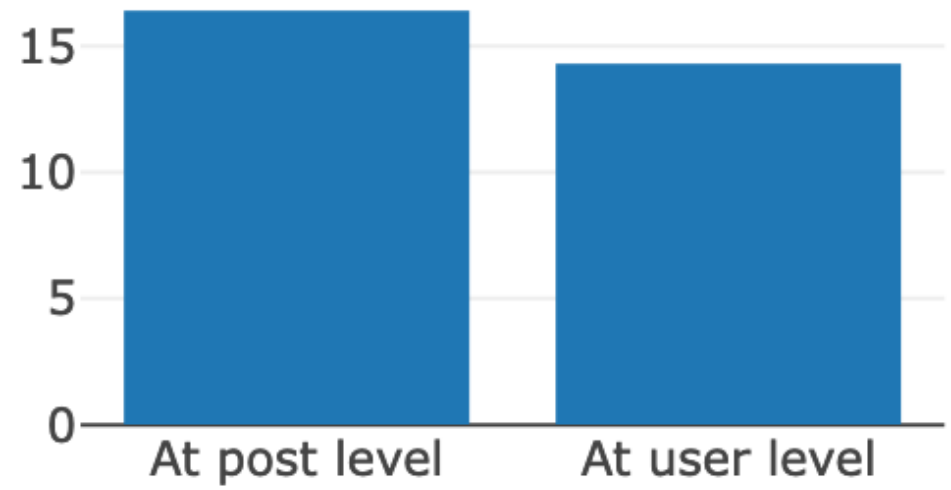
“You are great hahaha” > “You are great”

“Son of a bitch” > “Son of a bitch hahaha”

Review of the “Evidence” I Got:

“Diff-in-diff” 1 month before and after Musk’s acquisition

∇ Toxicity Telegram users - ∇Toxicity Non-Telegram users



Telegram users’ unsafe content descends less after Musk’s acquisition

Observations:

- Downwards trend of toxicity (natural for an invasion?)
- Robust to the temporal window chosen
- **Activity**
  - … a lot of Telegram-based bots/heavy users
  - Telegram users in both highest and lowest percentiles of unsafe content

(Lot of) Next Steps…

Theoretically:

Difficult model to extend (low analytical tractability)

Empirically:

Make a proper empirical model (structural, with a stochastic part)

Migration of Activity ≠ Migration?

+ *Fancy* things to try:

- Find bots? (It used to be possible before Musk)
- **Match (some) users from Telegram to Twitter**

Main takeaway

- **A policy (e.g. a stronger version of the DSA) can have unintended effects due to migration to non-regulated platforms**

  ➡ greatly depends on the network effects, advertisers' aversion to unsafe content, and quality of the outside platform

Not shown today: Monopolist model

- If a monopoly faces entry
  - ↓ strictness of moderation just enough to **deter entry**
  - min (unsafe content) = max (profits) at that point
  - **There is no need of regulation**

Most Importantly: **Merry Christmas !**

Appendix

Literature

- *Closest Paper:* **Madio & Quinn (2023)**.
    - Rich ads model, but exogenous creation of content.
    - Focuses in the monopolist + pricing of ads.

- **Liu et al (2021)** focuses on the (imperfect) technology

Empirical Side

- Jiménez Durán (2022), Jiménez Durán, Müller & Schwarz (2022)
- *Some CS Literature:* Schmitz, Muric, et al. (2022 and 2023)

Remarks

- Only in terms of total hate, leaving aside CS (the analysis is less neat, but possible)

- The regulator might care more about the hate experienced by low-hate people:
    - there is a rational for stricter policy if this is the case
    - but could end up “throwing to the lions” to “median” users