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

Iván Rendo (TSE)

IO Workshop

Toulouse, 18 December 2023

Motivation

Motivation

- Increased interest in **online** hateful/extreme/**unsafe content**:
    - E.g. spread of jihadism, bullying…
    - Jiménez-Durán (2022) links online hate to **offline violence**
        - ➡ EU Commission mandates the **Digital Services Act** (DSA)

Motivation

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

    ➡ EU Commission mandates the **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

Today

Platforms' competition model to analyze the interaction between:

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

… for an ad-funded platform

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

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

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**
  - OM ultra likes fights vs PSG ultras, not cookies from *a* granny, and vice-versa.

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**
  - OM ultra likes fights vs PSG ultras, not cookies from *a* granny, and vice-versa.

2 Asymmetric **Platforms**:

- 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**

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**
  - OM ultra likes fights vs PSG ultras, not cookies from *a* granny, and vice-versa.

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

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**
  - OM ultra likes fights vs PSG ultras, not cookies from *a* granny, and vice-versa.

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

Key Driving Force: **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

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, w small network effects

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, w small network effects

2. **Policy:**
    - **Misalignment of incentives** between platform and regulator
    - Imposing a **minimal** content moderation intensity (policy):
        i. Large network effects: always **superfluous**
        ii. Mid to small network effects: can be **useful**

Roadmap

I. Theoretical Model
   - Characterization of the Equilibrium
   - Optimal Regulation

II. Empirical Evidence

THEORY

Model

Model

- A unit mass of **individuals**, heterogeneous in their most preferred unsafety of content: $\theta_i \sim U(0,1)$

Model

- A unit mass of **individuals**, heterogeneous in their most preferred unsafety of content: $\theta_i \sim U(0,1)$

- 2 **platforms** $j = 1,2$
  - with different K_j **max unsafety level allowed**

Model

- A unit mass of **individuals**, heterogeneous in their most preferred unsafety of content: $\theta_i \sim U(0,1)$

- 2 **platforms** $j = 1,2$
  - with different K_j **max unsafety level allowed**

- Each individual i in platform j **creates** 1 unit of content of unsafety

Model

- A unit mass of **individuals**, heterogeneous in their most preferred unsafety of content: $\theta_i \sim U(0,1)$

- 2 **platforms** $j = 1,2$
  - with different K_j **max unsafety level allowed**

- Each individual i in platform j **creates** 1 unit of content of unsafety

$$\min\{\theta_i, K_j\}$$

Model

- A unit mass of **individuals**, heterogeneous in their most preferred unsafety of content: $\theta_i \sim U(0,1)$

- 2 **platforms** $j = 1,2$
  - with different K_j **max unsafety level allowed**

- Each individual i in platform j **creates** 1 unit of content of unsafety

$$\min\{\theta_i, K_j\}$$

- Each individual i in platform j **reads** all the content, of avg unsafety

$$\bar{\theta}_j = \sum_{i \in j} \min\{\theta_i, K_j\}$$

Model

- A unit mass of **individuals**, heterogeneous in their most preferred unsafety of content: $\theta_i \sim U(0,1)$

- 2 **platforms** $j = 1,2$
  - with different K_j **max unsafety level allowed**

- Each individual i in platform j **creates** 1 unit of content of unsafety

$$\min\{\theta_i, K_j\}$$

- Each individual i in platform j **reads** all the content, of avg unsafety

$$\bar{\theta}_j = \sum_{i \in j} \min\{\theta_i, K_j\}$$

= 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:

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

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

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

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

Users in the Platform

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

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

- 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

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

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

- 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 Value of the Good Platform

- 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 Value of the Good Platform

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

- 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 Value of the Good Platform

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

Rk: I abstract of modelling the utility from creation of content

Advertisers

Advertisers

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

Advertisers

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

Are **averse** to unsafe content

Advertisers

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

Are **averse** to unsafe content

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

Advertisers

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

Are **averse** to unsafe content

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

Regulated Platform

Advertisers

Buy a fix 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** $\theta_i > K$

Advertisers

Buy a fix 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:

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

Advertisers

Buy a fix 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:

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

users in platform

Advertisers

Buy a fix 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:

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

users in platform

Price of ads

Advertisers

Buy a fix 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:

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 fix 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:

Advertisers aversion to unsafe content

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

Average content

users in platform

Price of ads

Advertisers

Buy a fix 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:

Advertisers aversion to unsafe content

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

Average content

users in platform

Price of ads

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

Timing

Timing

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

Timing

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

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

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

Timing

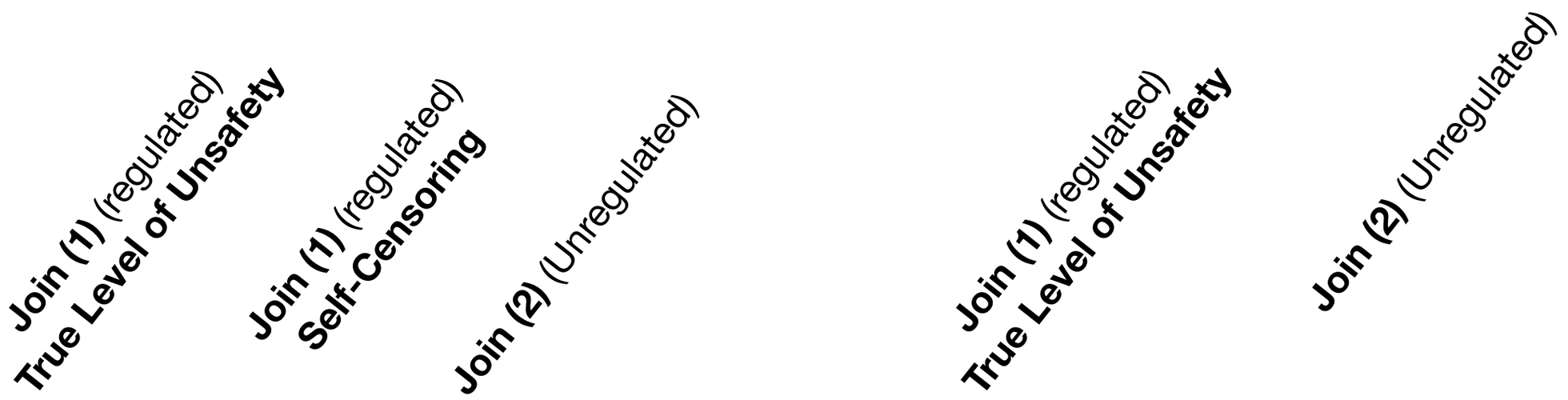
1. The regulated platform (1) chooses the content moderation policy K and commits to it

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

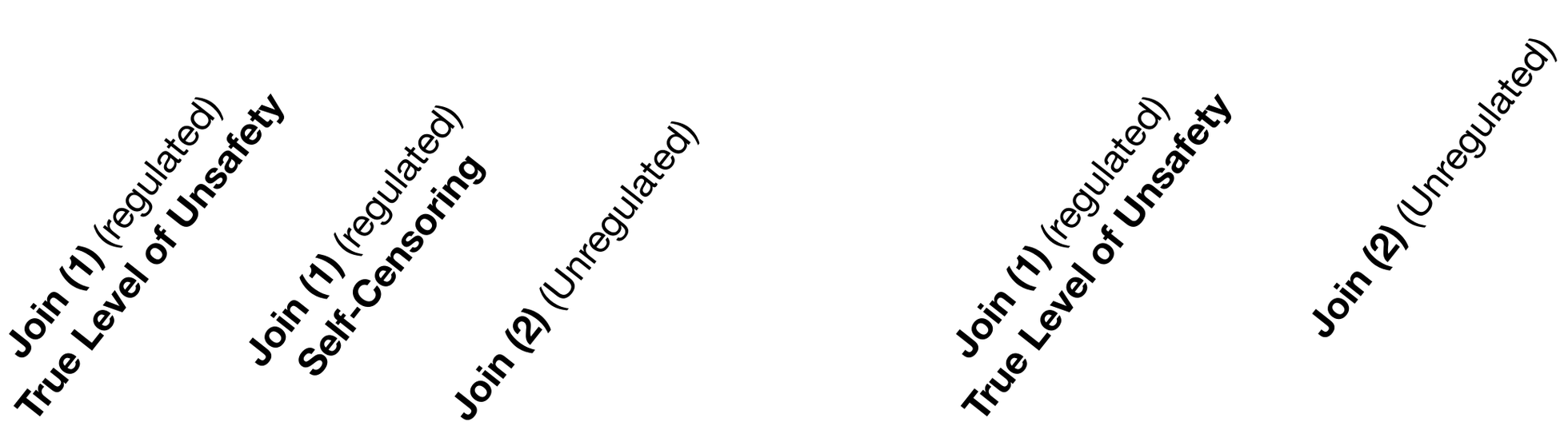
User i joins platform (1) iff $\theta_i < t^*$, otherwise, they join (2)



Threshold Equilibrium

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

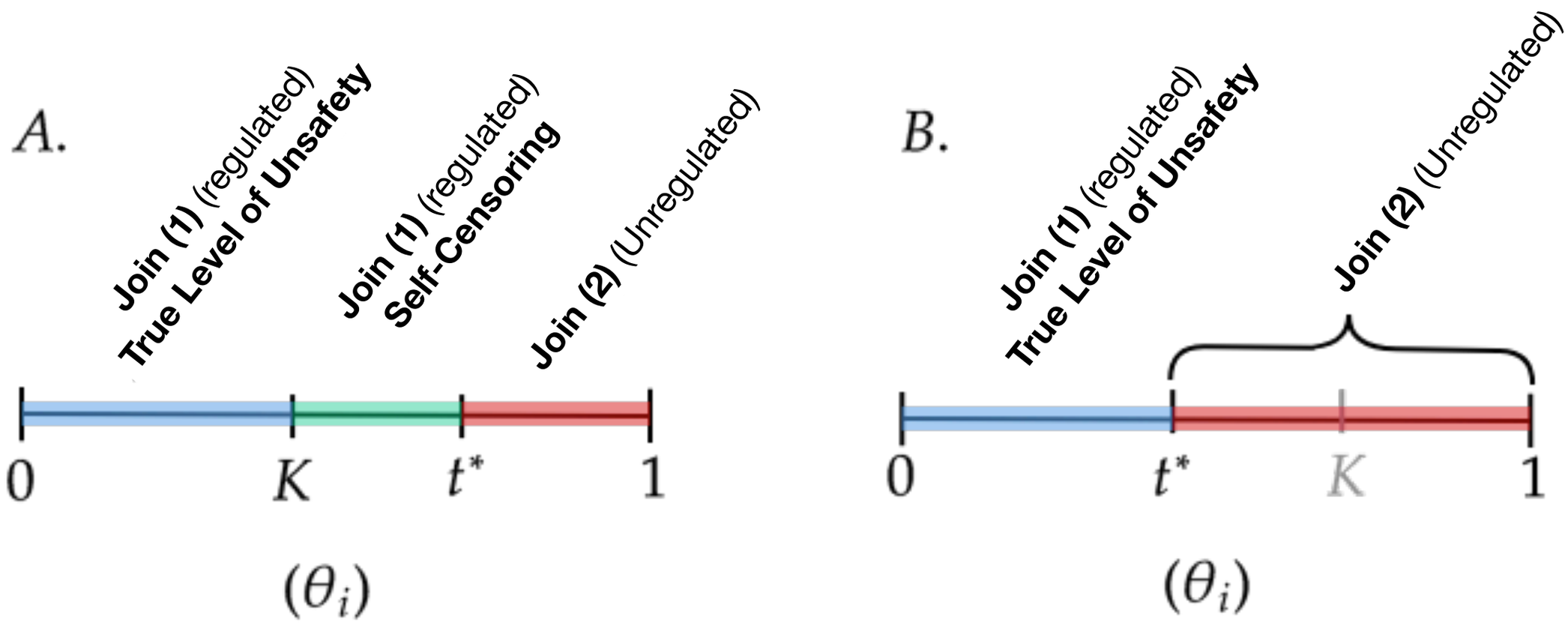
Under some conditions on α (not too low), for any K, there exist a **unique threshold equilibrium**, which takes one of these two forms:



Threshold Equilibrium

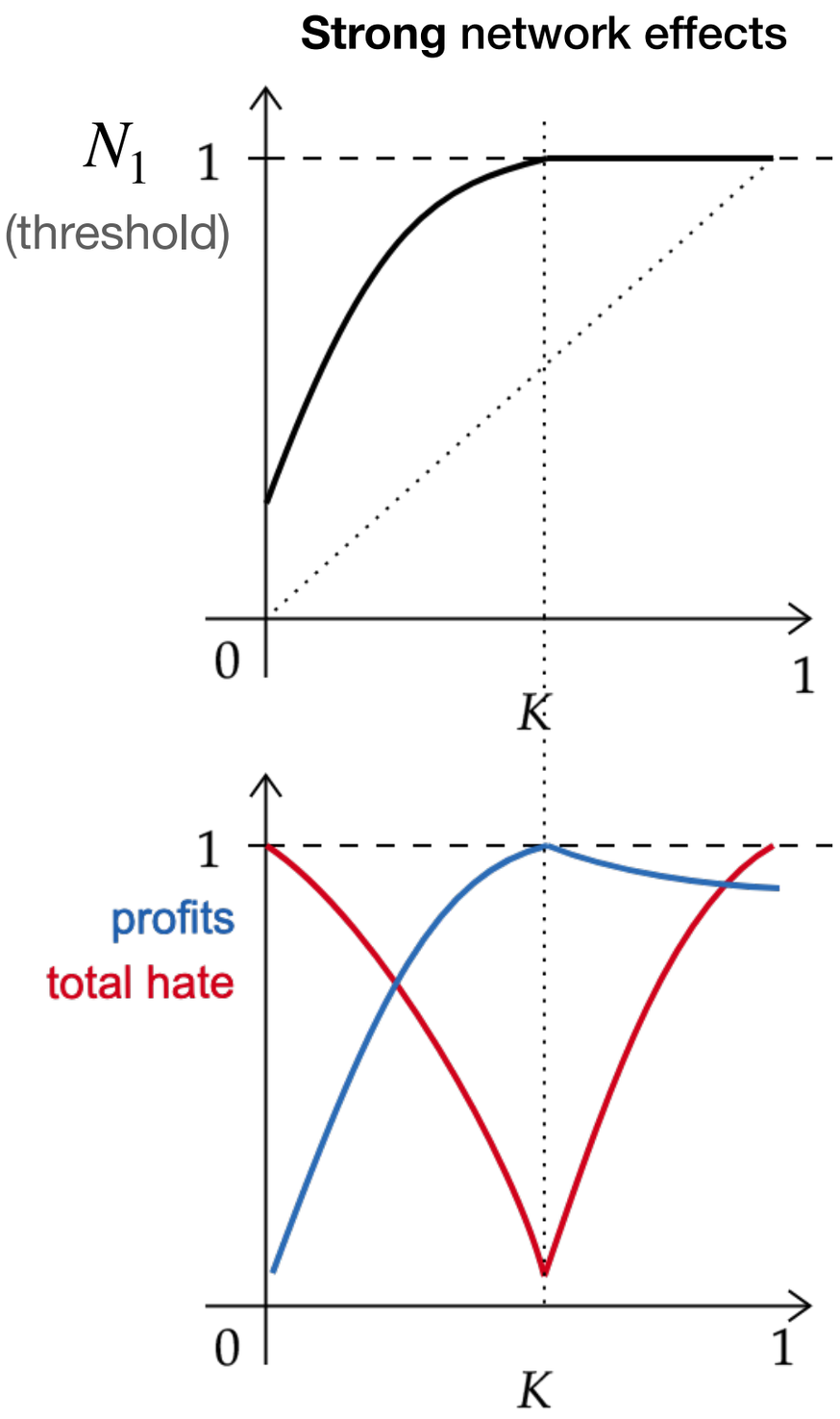
User i joins platform (1) iff $\theta_i < t^*$, otherwise, they join (2)

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



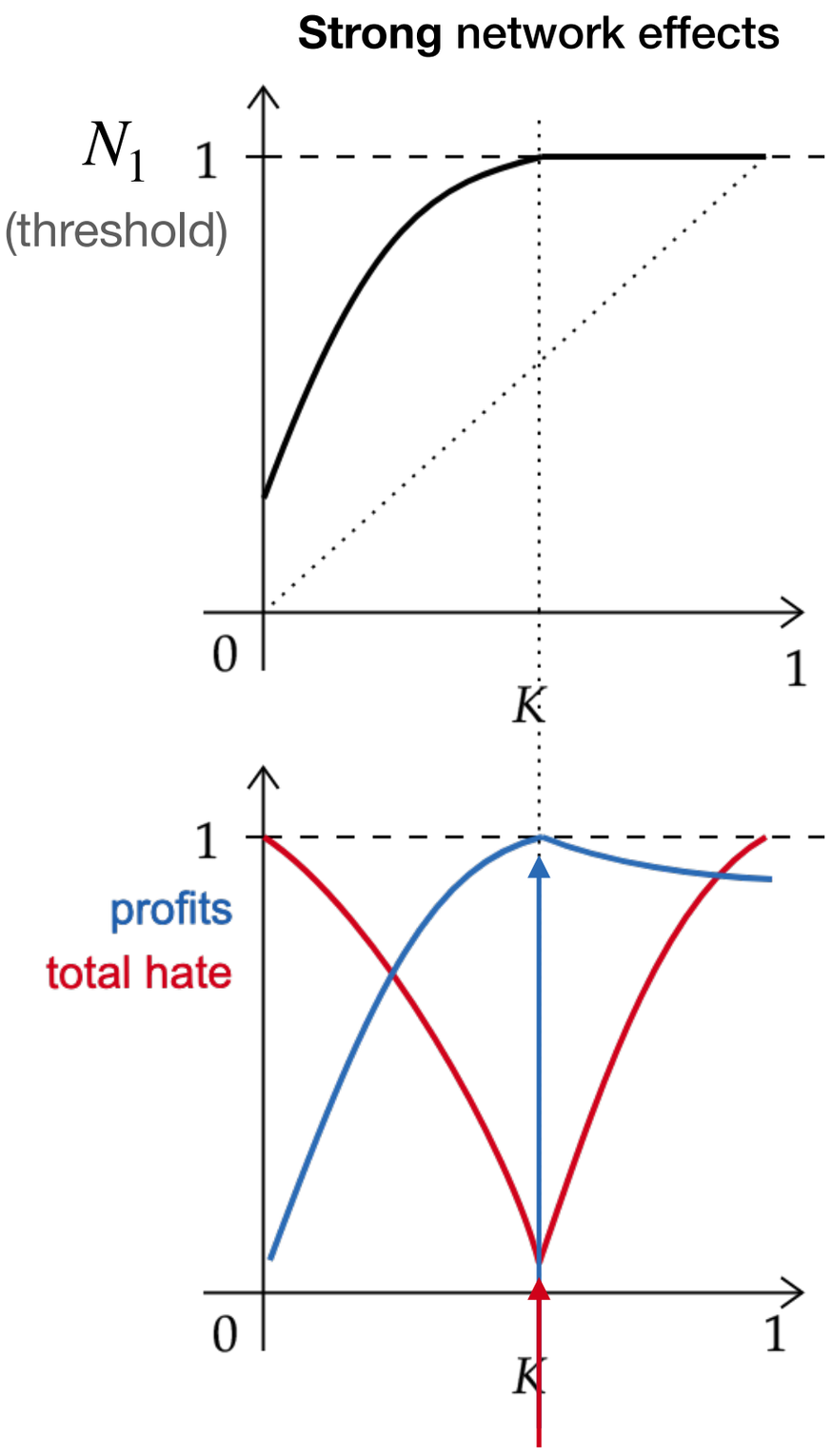
K=0, very strict content moderation
K=1, no content moderation

Characterization of the Equilibrium



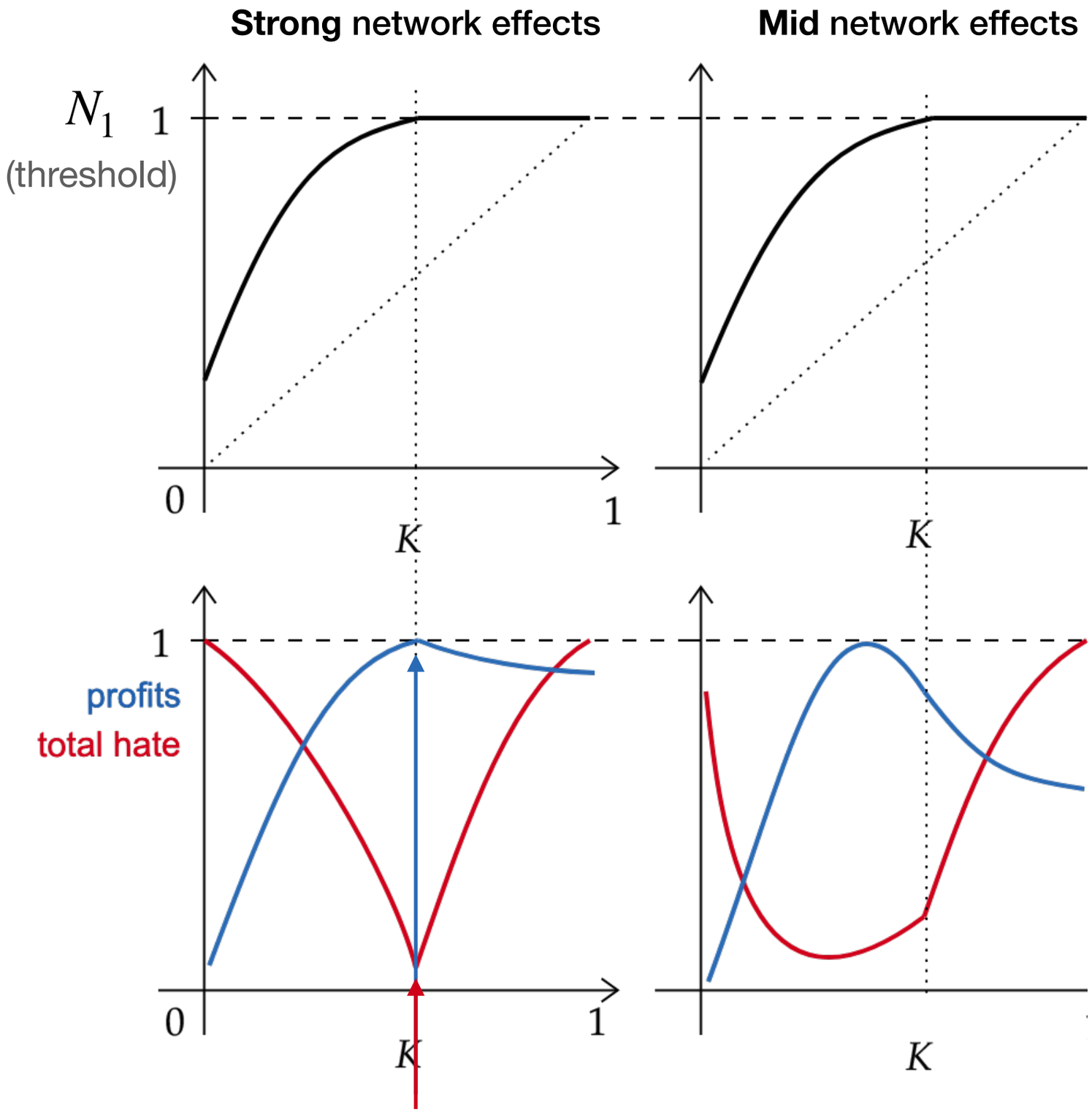
K=0, very strict content moderation
K=1, no content moderation

Characterization of the Equilibrium



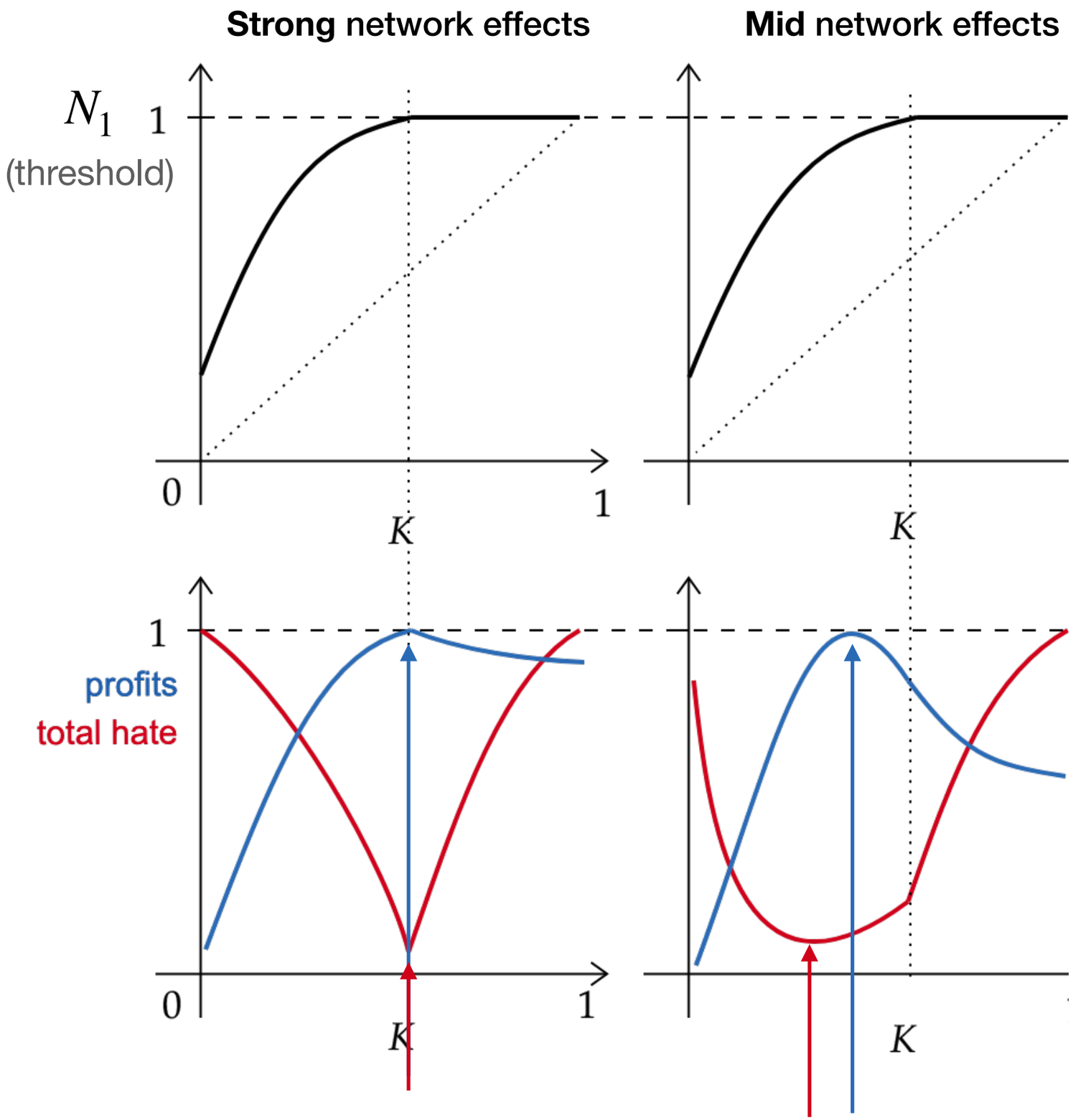
K=0, very strict content moderation
K=1, no content moderation

Characterization of the Equilibrium



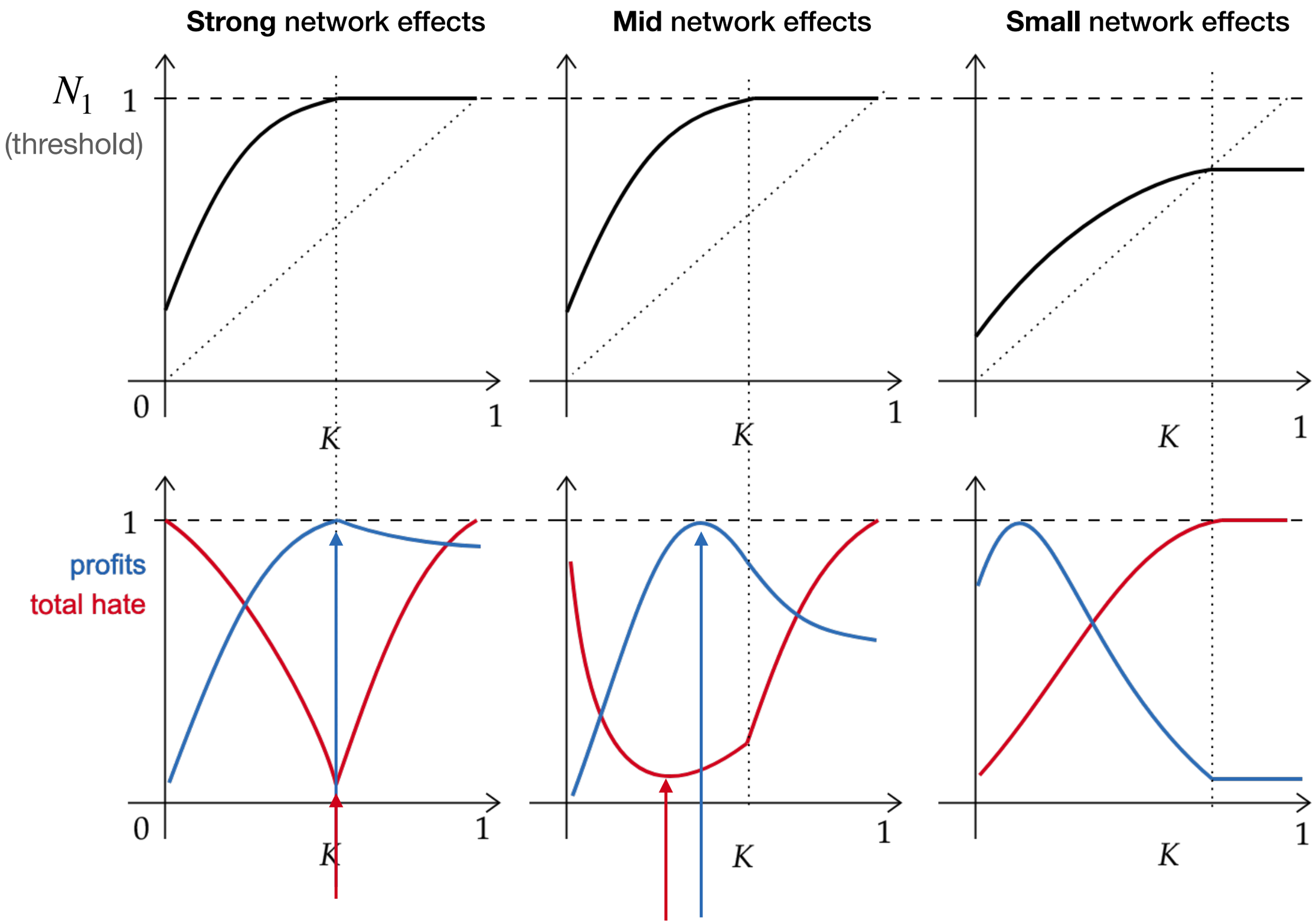
K=0, very strict content moderation
K=1, no content moderation

Characterization of the Equilibrium



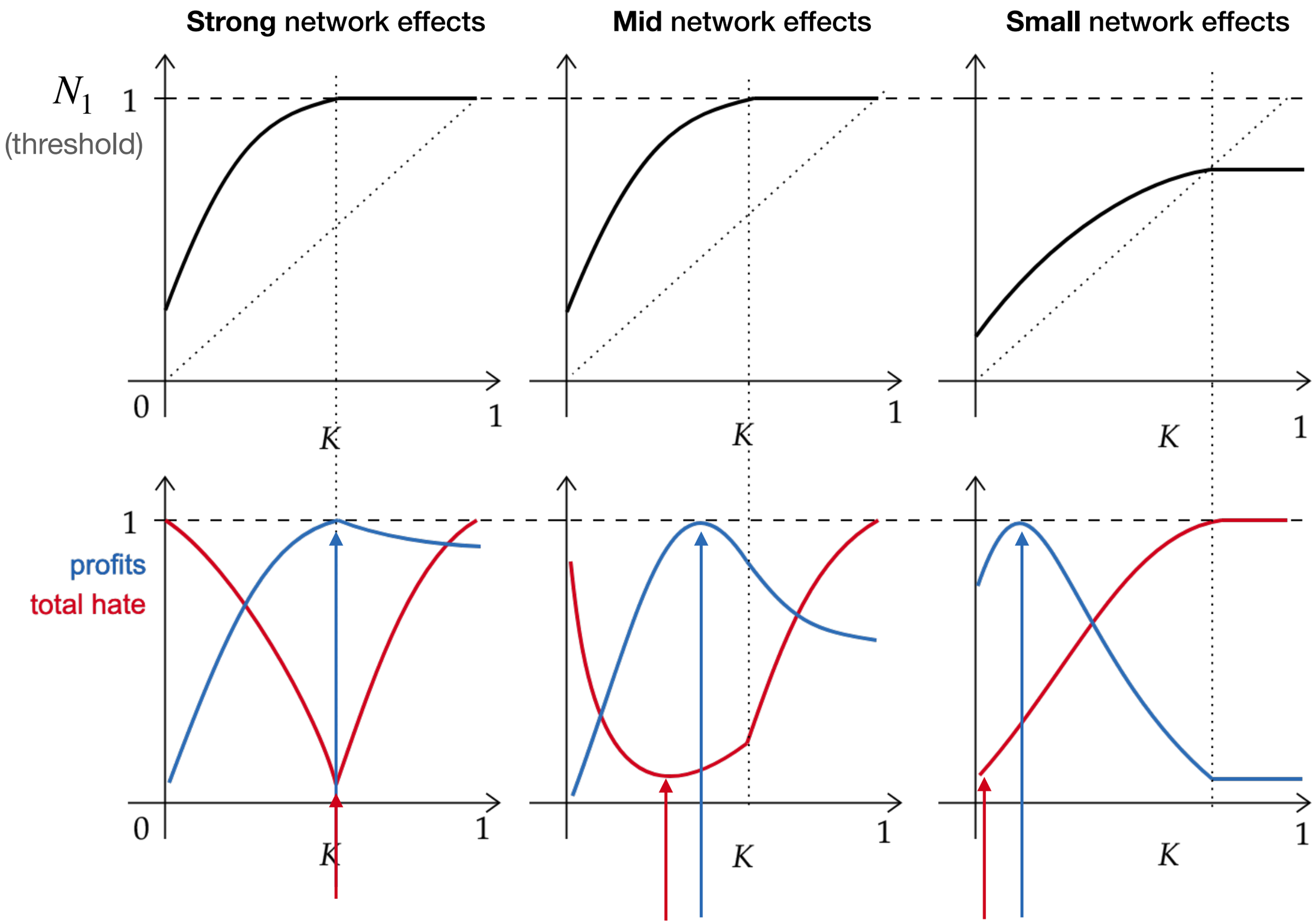
K=0, very strict content moderation
K=1, no content moderation

Characterization of the Equilibrium



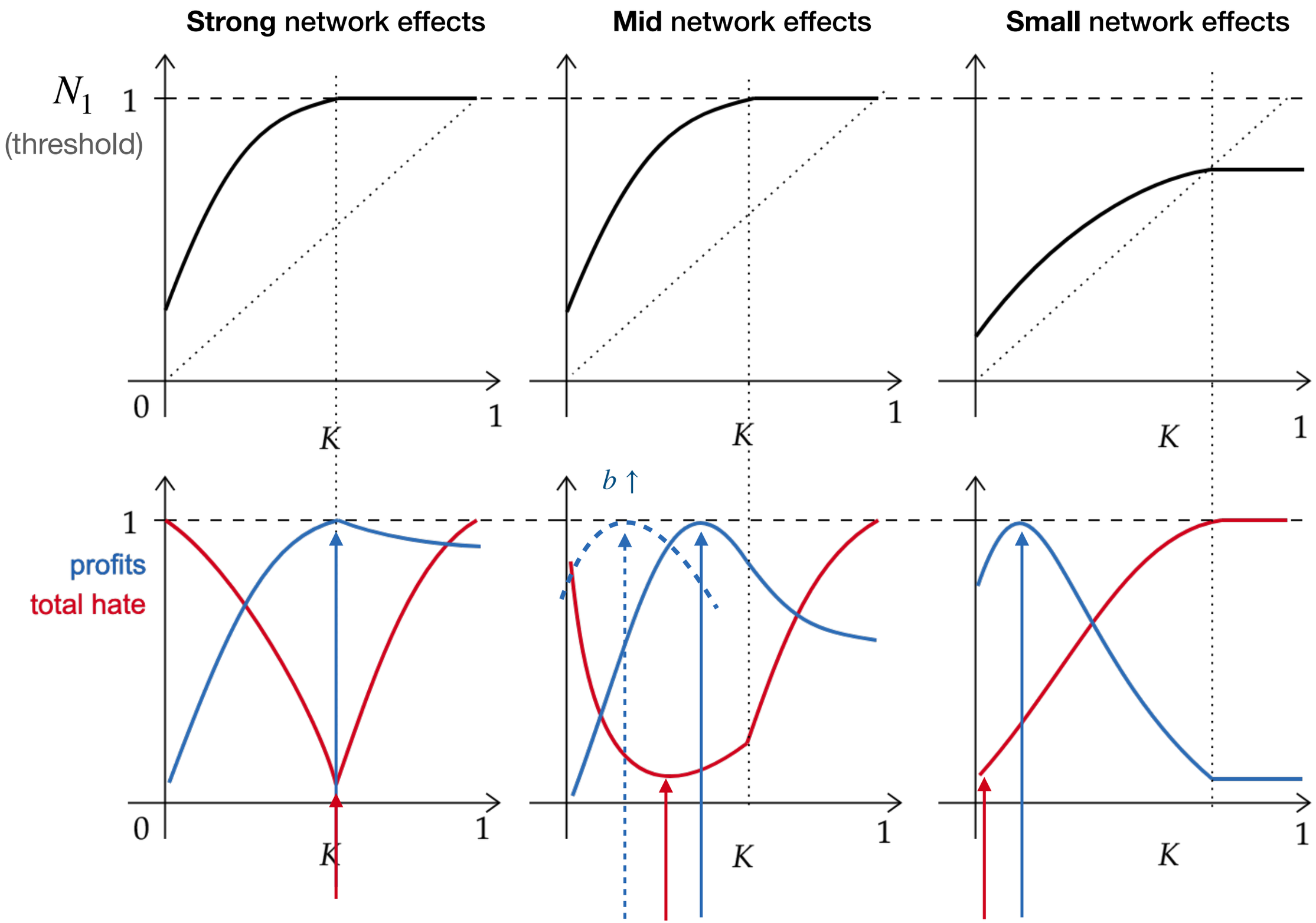
K=0, very strict content moderation
K=1, no content moderation

Characterization of the Equilibrium



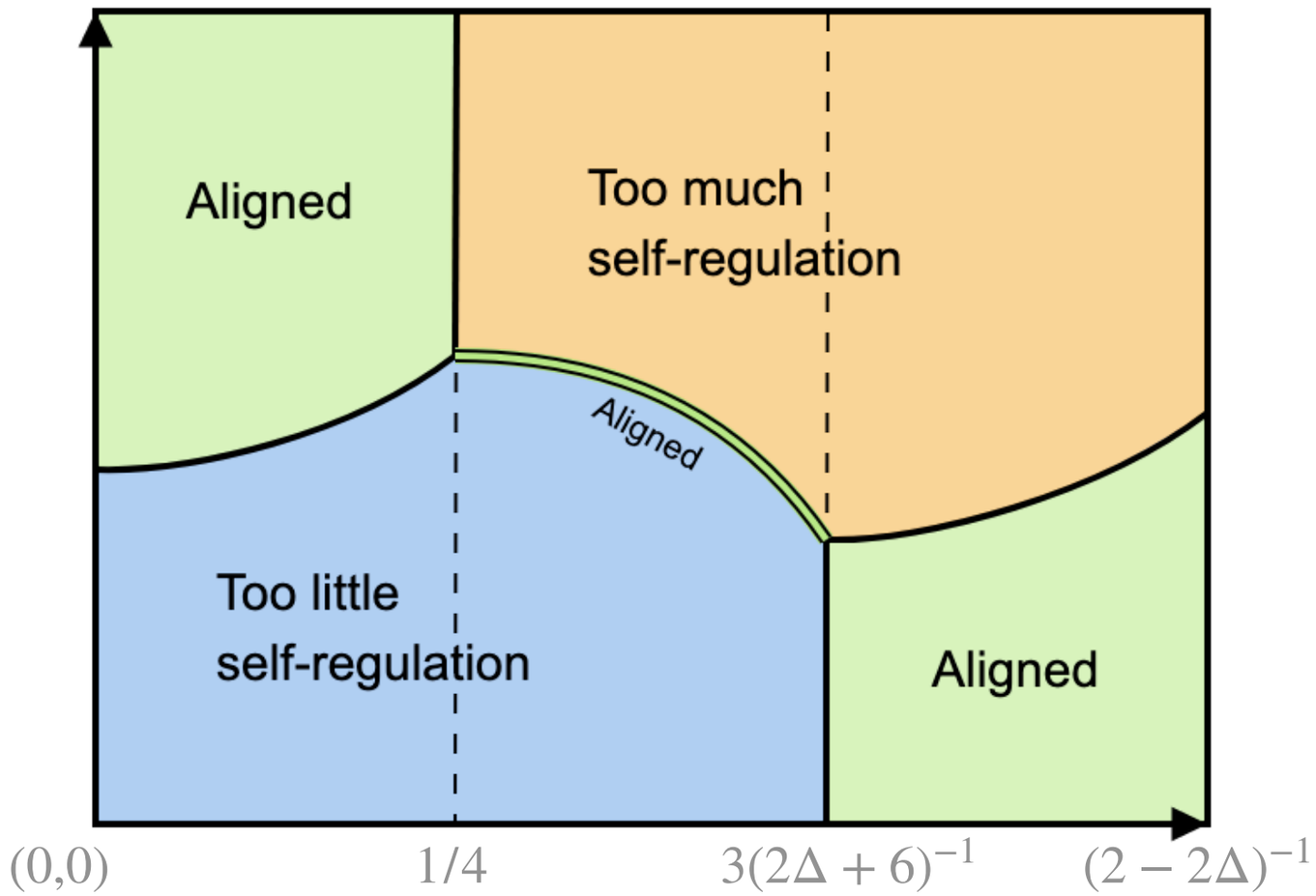
K=0, very strict content moderation
K=1, no content moderation

Characterization of the Equilibrium



Policy (to min unsafe content)

Advertisers aversion to unsafe content (b)

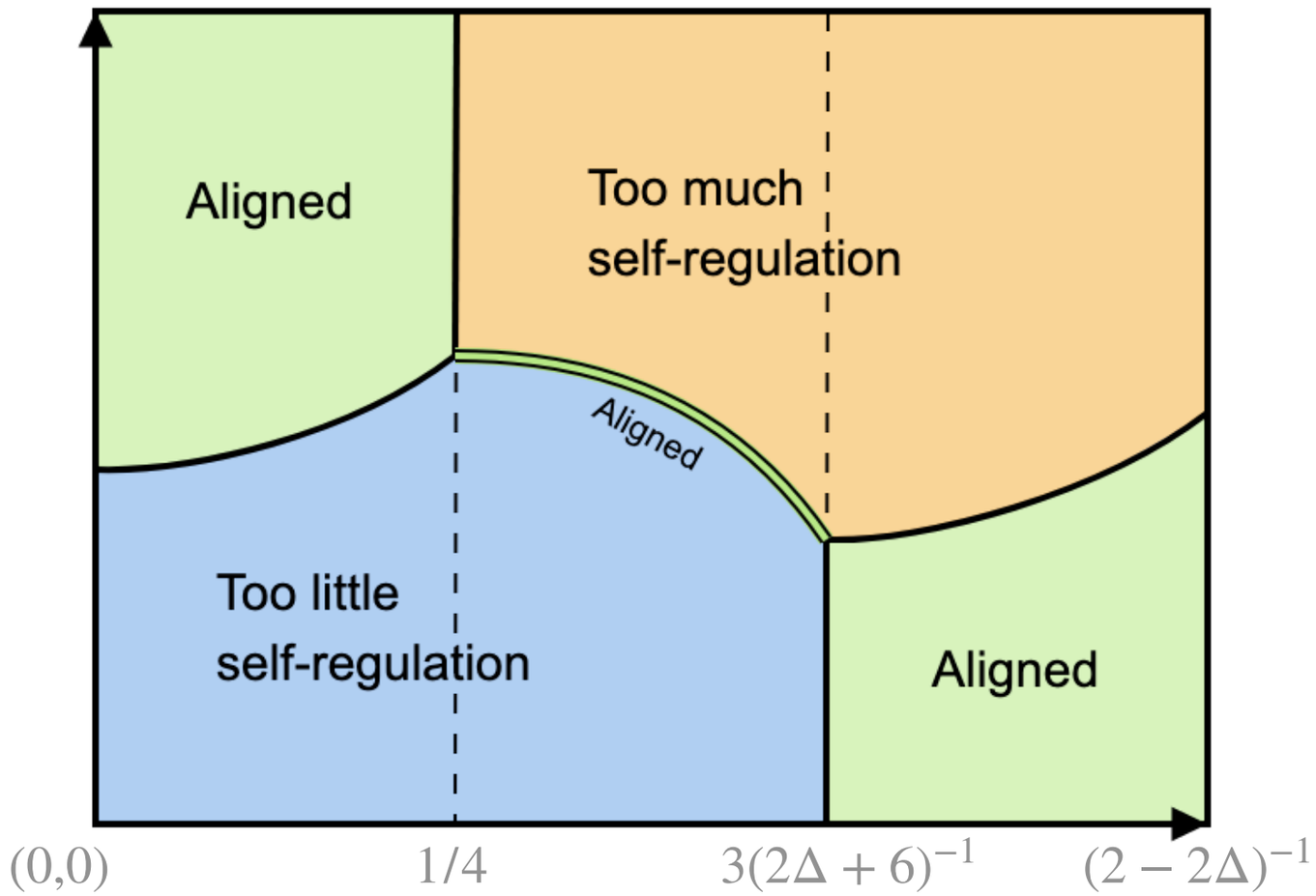


Network Effects (α^{-1})

Policy (to min unsafe content)

Green Area: Nothing to do!

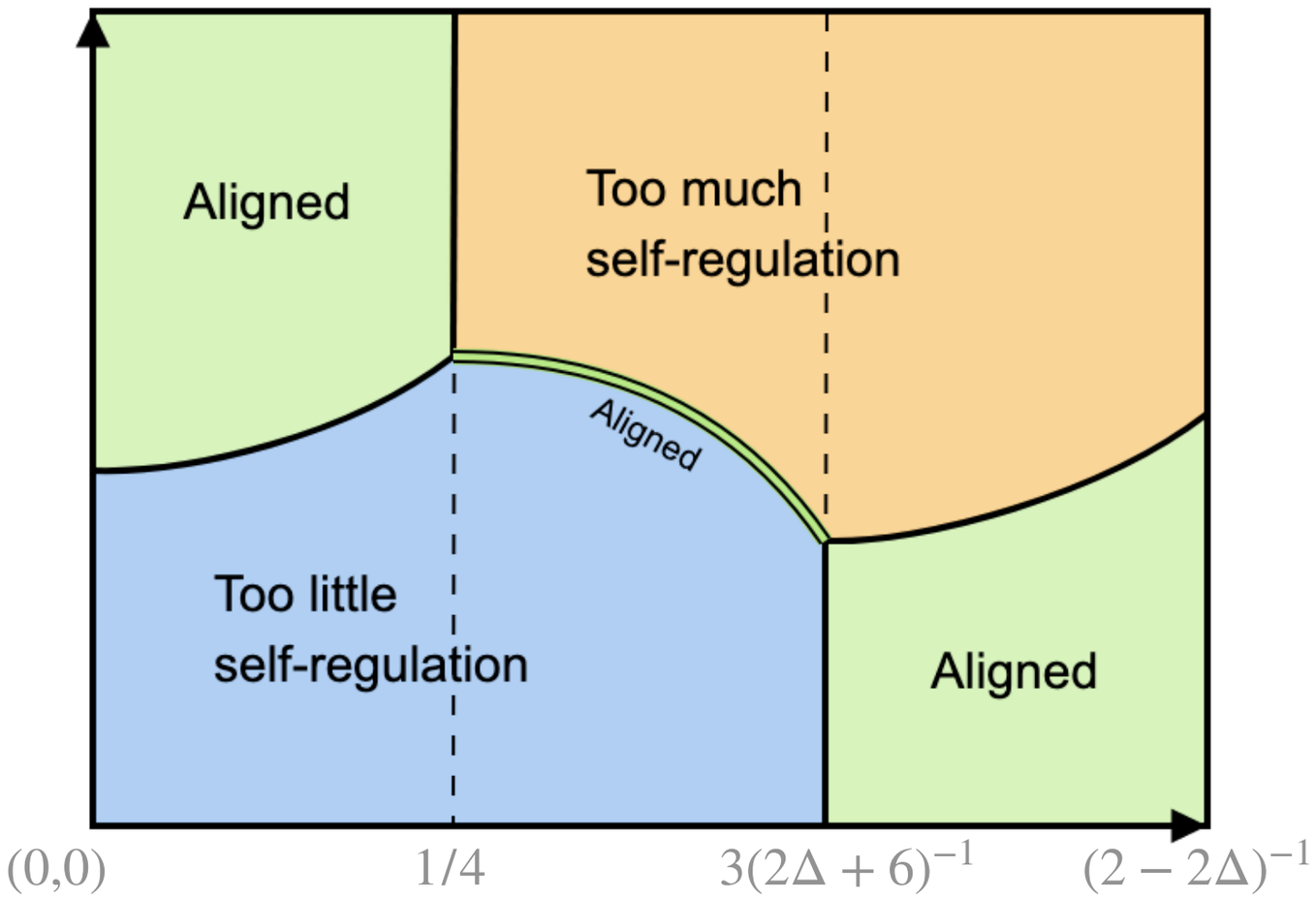
Advertisers aversion to unsafe content (b)



Network Effects (α^{-1})

Policy (to min unsafe content)

Advertisers aversion to unsafe content (b)



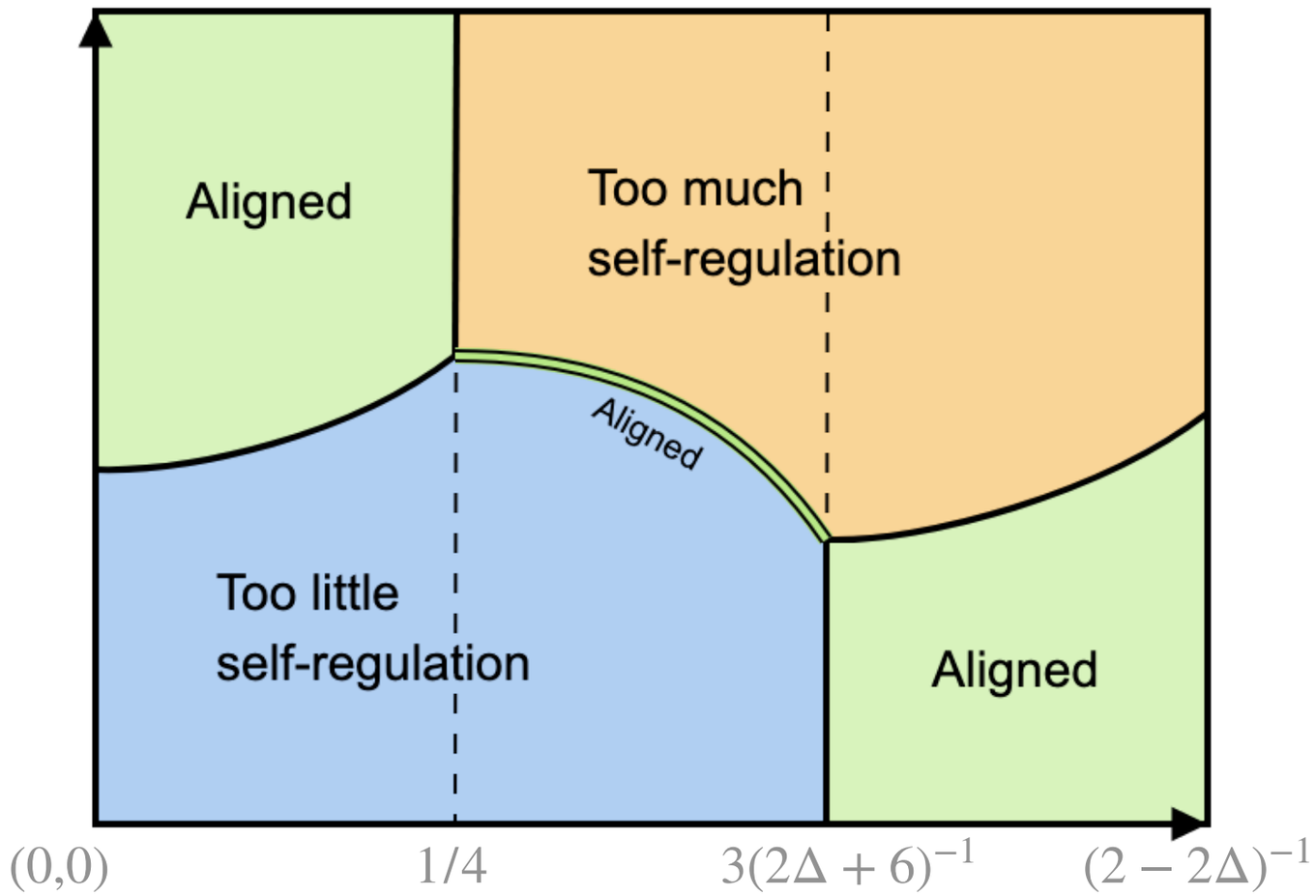
Network Effects (α^{-1})

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**

Policy (to min unsafe content)

Advertisers aversion to unsafe content (b)



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

Predictions to take to the Data, using Musk's Event

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Some Predictions:

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Some Predictions:

1. More unsafe content in general. *Hickey et al.* (2023)

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Some Predictions:

1. More unsafe content in general. *Hickey et al.* (2023)
2. More 'hate' from 'hateful users'. *Hickey et al.* (2022)

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Some Predictions:

1. More unsafe content in general. *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:

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Some Predictions:

1. More unsafe content in general. *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:
    i. Hateful for Twitter standards

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Some Predictions:

1. More unsafe content in general. *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:
    i. Hateful for Twitter standards
    ii. Users on Twitter for whom the policy was binding, increase the unsafe content

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Some Predictions:

1. More unsafe content in general. *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:
    i. Hateful for Twitter standards
    ii. Users on Twitter for whom the policy was binding, increase the unsafe content
    iii. Decrease of unsafe content in Telegram from these users

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Some Predictions:

1. More unsafe content in general. *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:
    i. Hateful for Twitter standards
    ii. Users on Twitter for whom the policy was binding, increase the unsafe content
    iii. Decrease of unsafe content in Telegram from these users

(4). What happens to total unsafe content?

Predictions to take to the Data, using Musk's Event

Context: Musk comes to Twitter: *exogenous* $\uparrow K$

Some Predictions:

1. More unsafe content in general. *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:

Today

i. Hateful for Twitter standards

ii. Users on Twitter for whom the policy was binding, increase the unsafe content

iii. Decrease of unsafe content in Telegram from these users

(4). What happens to total unsafe content?

Review of the Data I Have:

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*)

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

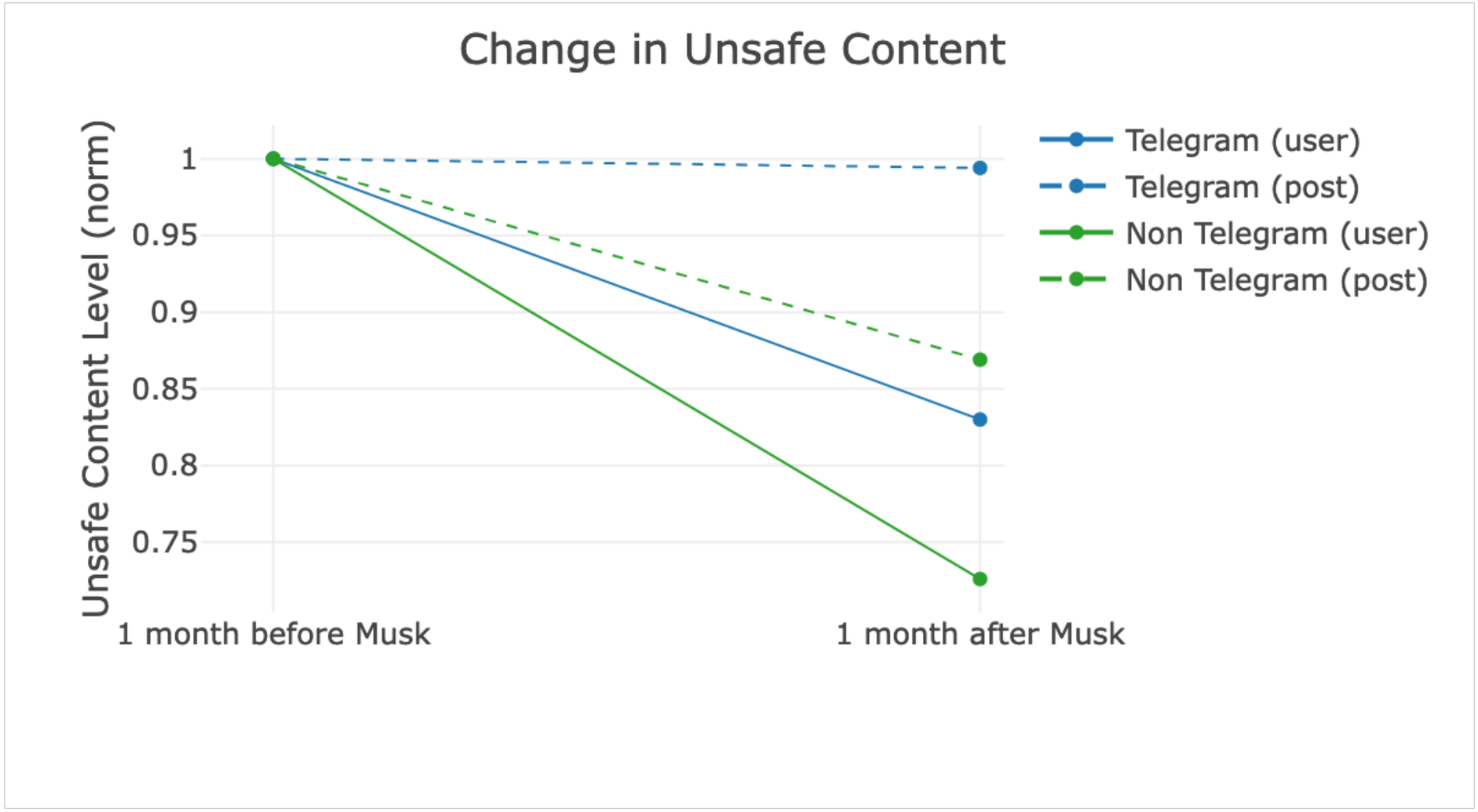
In terms of *toxicity*:

"You are great hahaha" > "You are great"

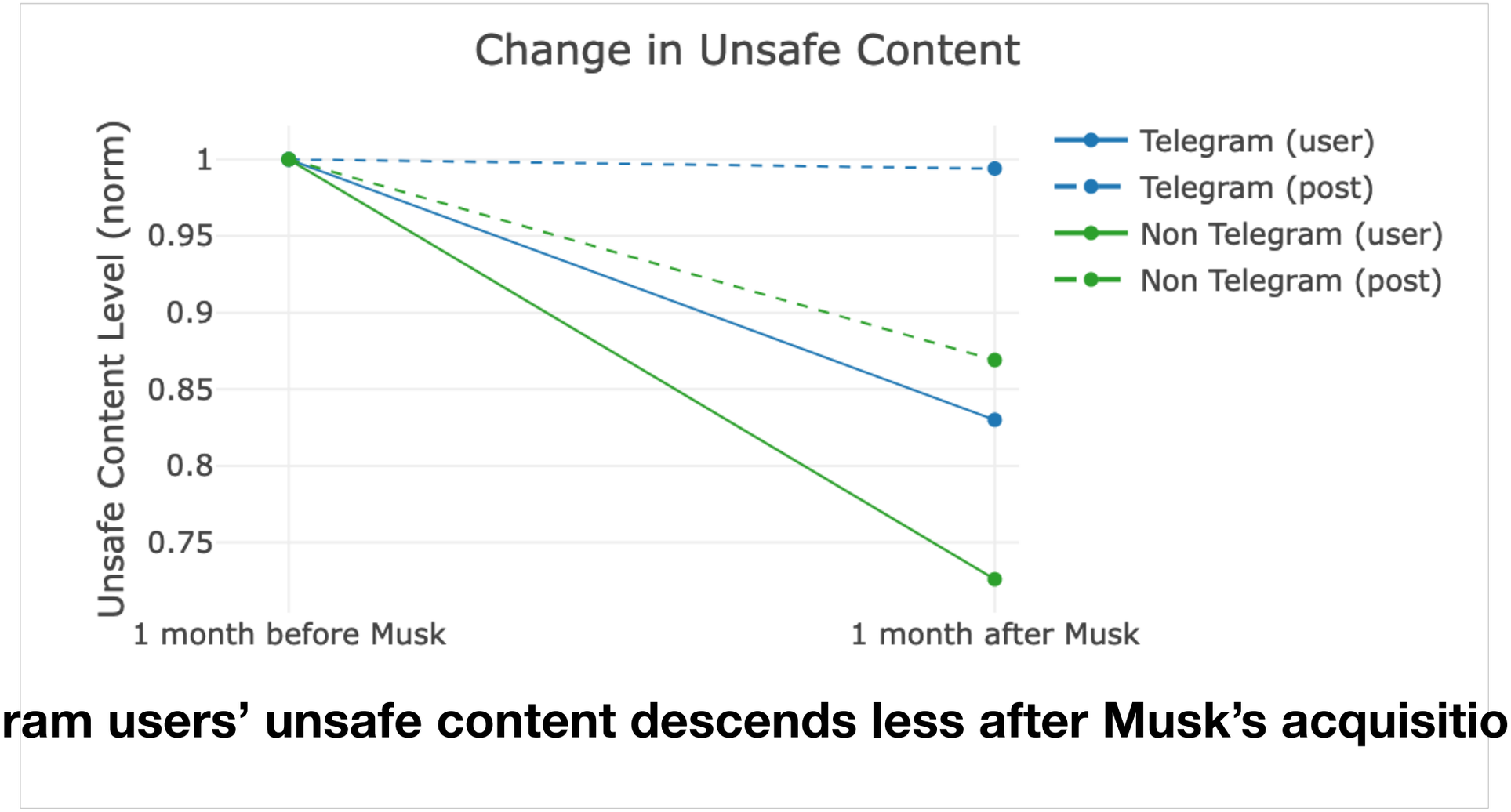
"Son of a bitch" > "Son of a bitch hahaha"

Review of the “Evidence” I Got:

Review of the “Evidence” I Got:

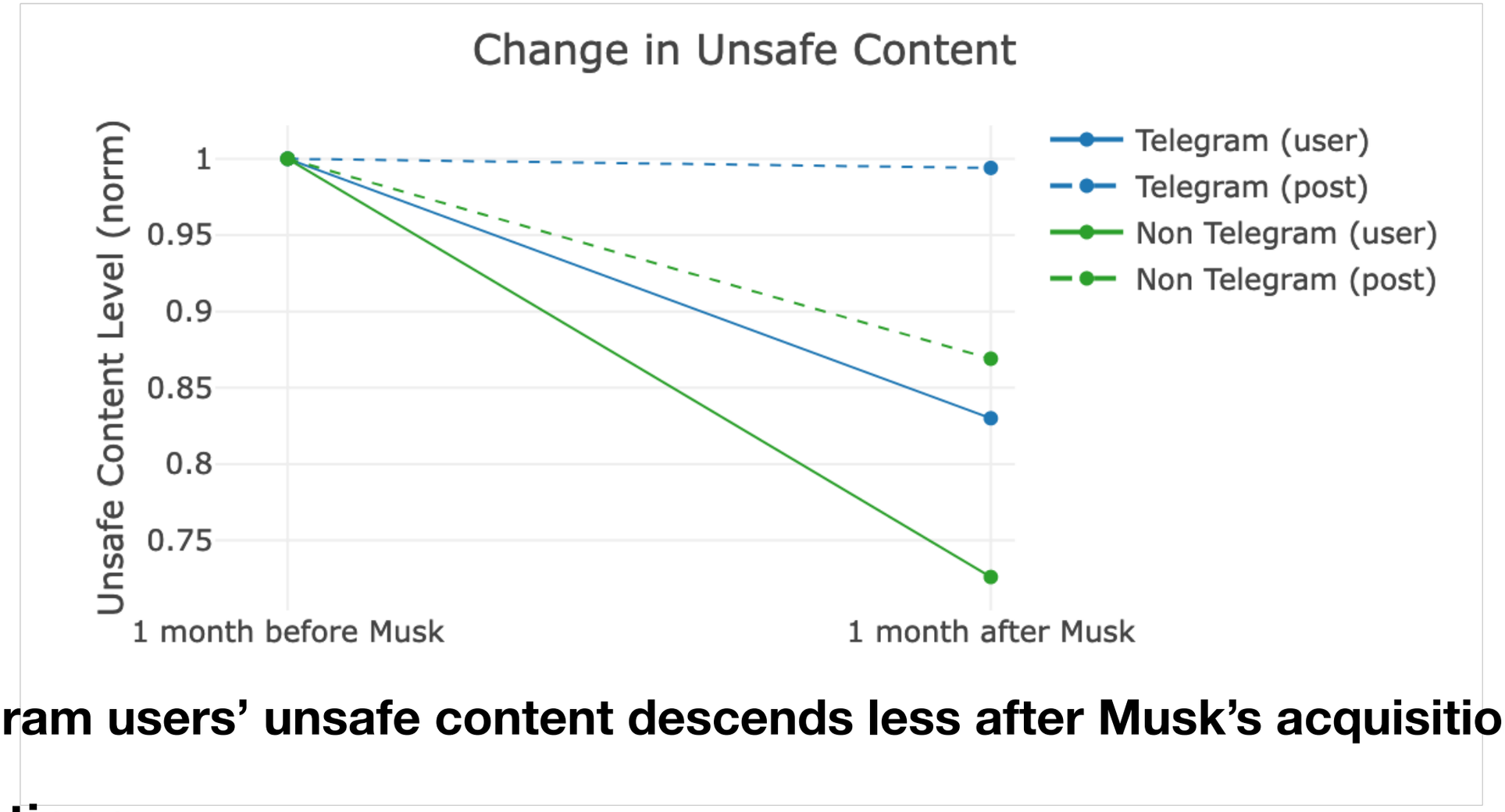


Review of the “Evidence” I Got:



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

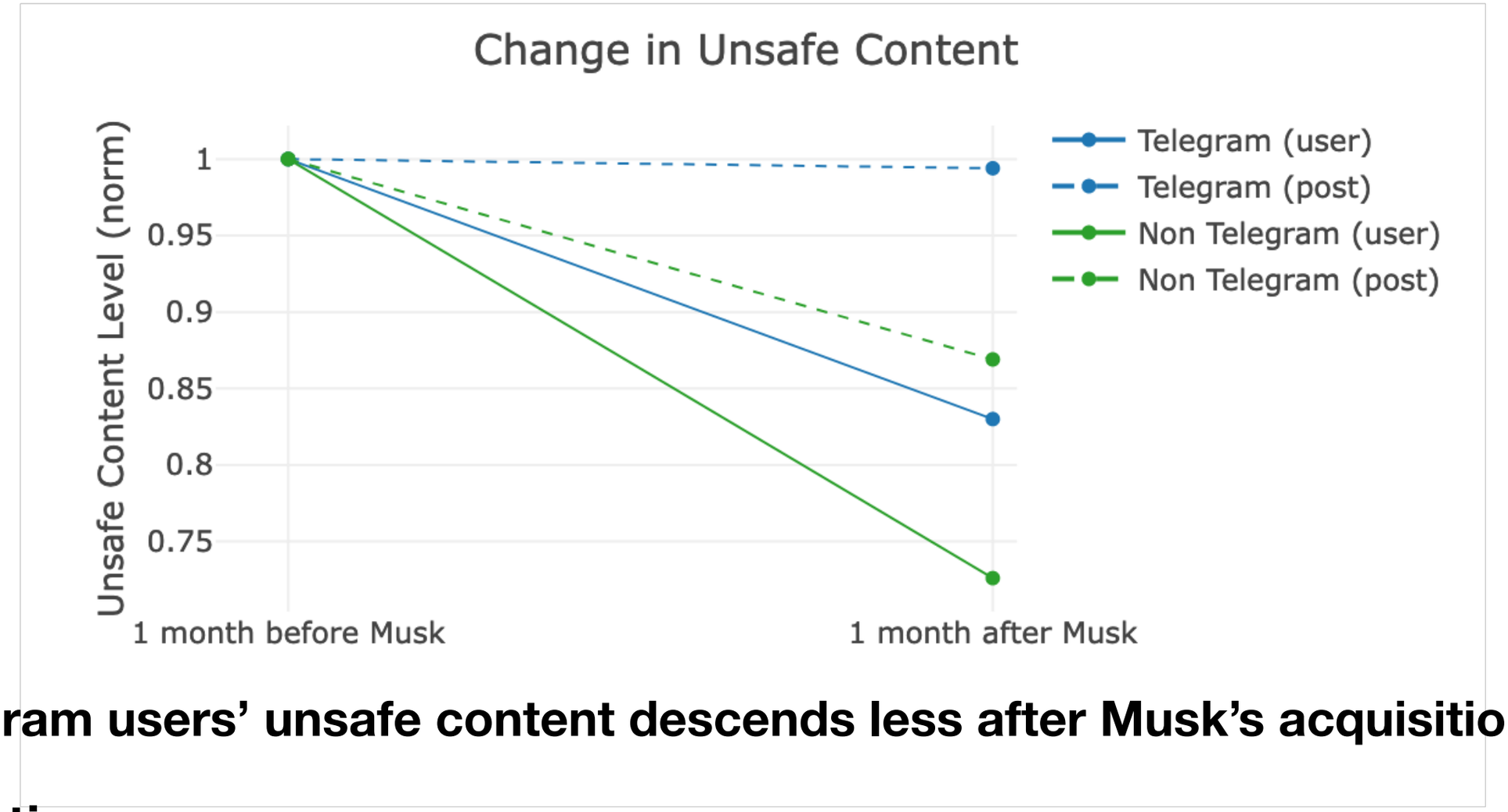
Review of the “Evidence” I Got:



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

Observations:

Review of the “Evidence” I Got:

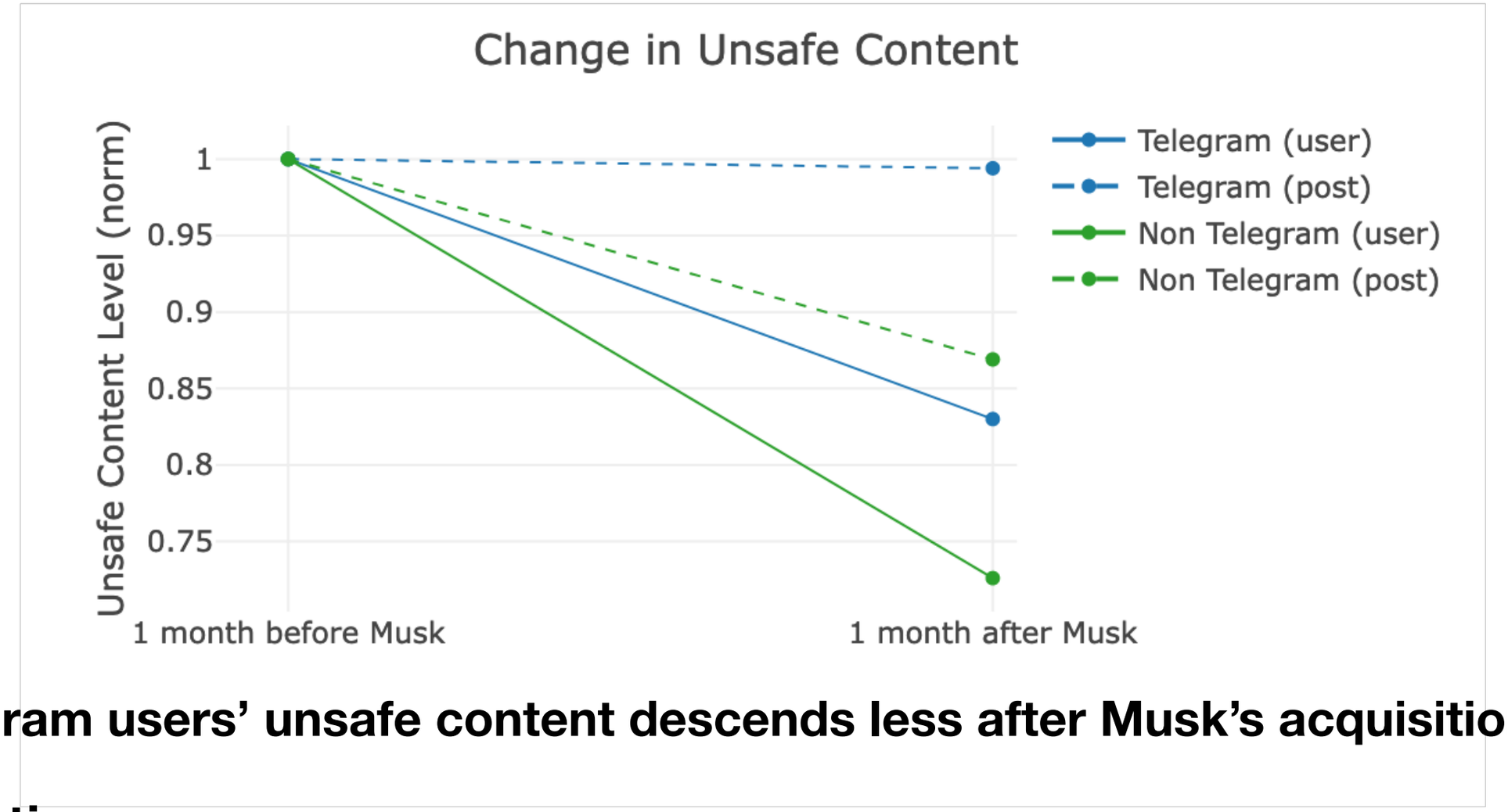


Telegram users’ unsafe content descends less after Musk’s acquisition

Observations:

- Downwards (natural?) trend of the invasion

Review of the “Evidence” I Got:

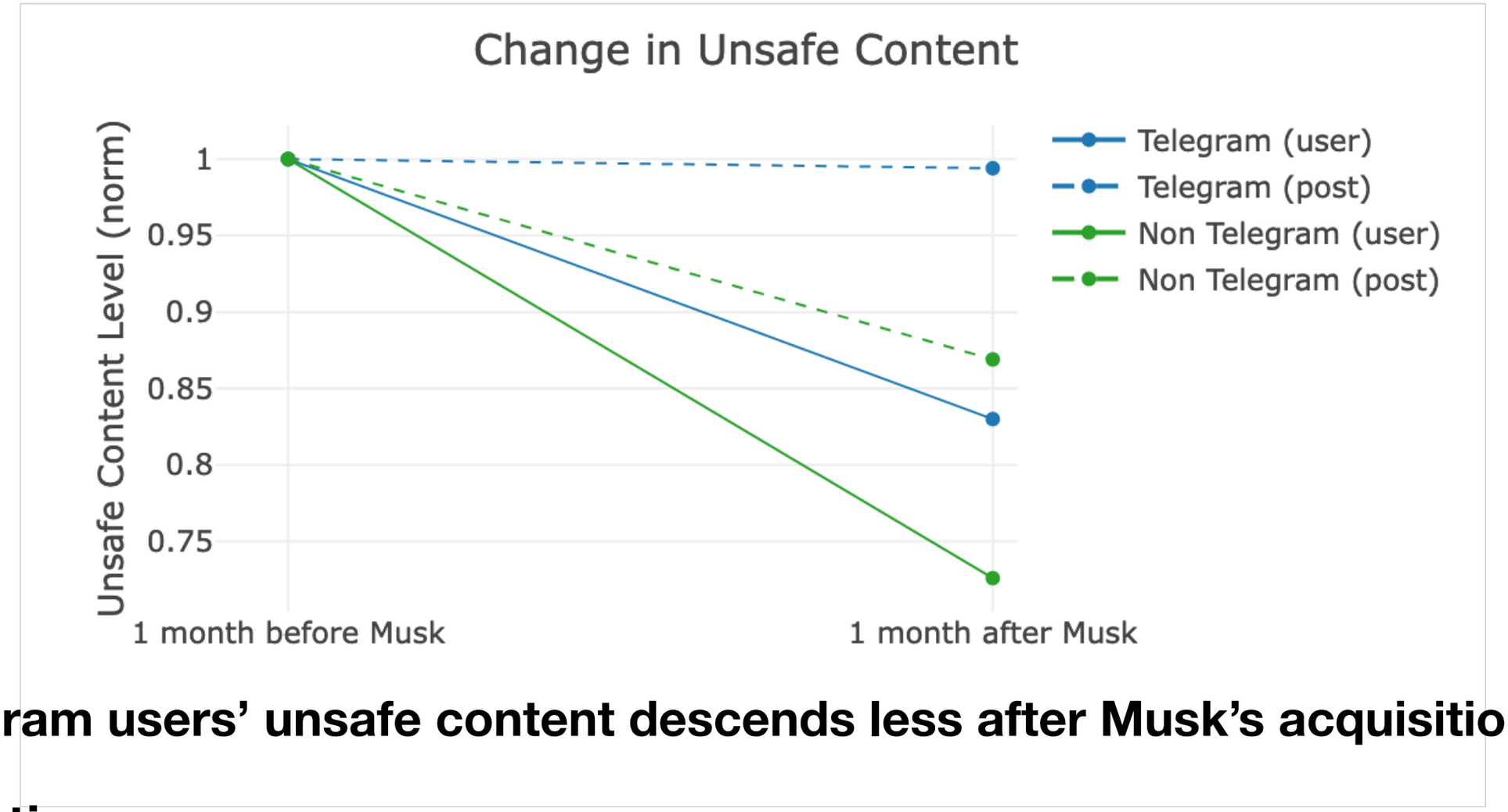


Telegram users’ unsafe content descends less after Musk’s acquisition

Observations:

- Downwards (natural?) trend of the invasion
- Robust to the temporal window chosen

Review of the “Evidence” I Got:

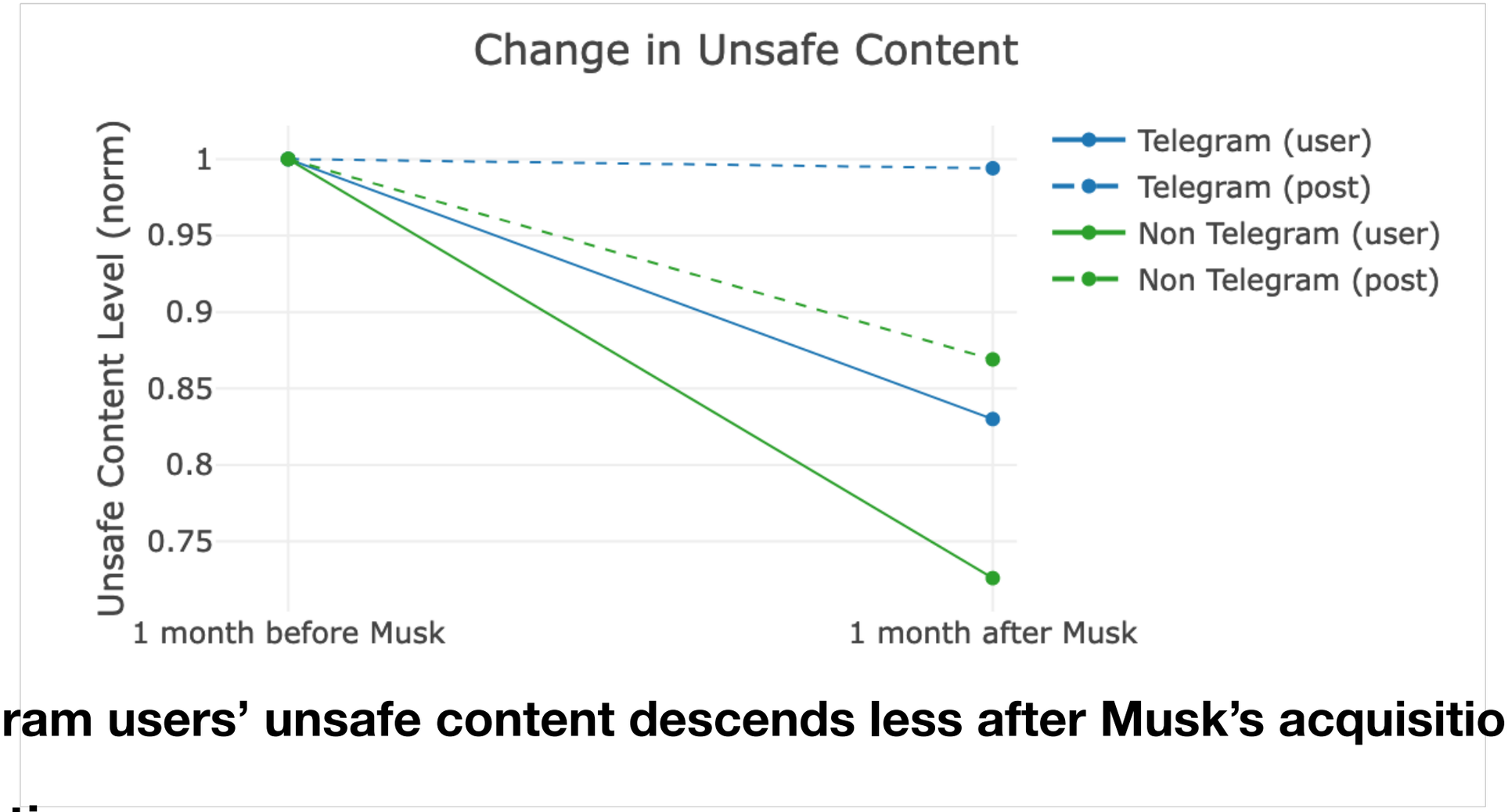


Telegram users’ unsafe content descends less after Musk’s acquisition

Observations:

- Downwards (natural?) trend of the invasion
- Robust to the temporal window chosen
- Discrepancy: mainly due to **activity**

Review of the “Evidence” I Got:

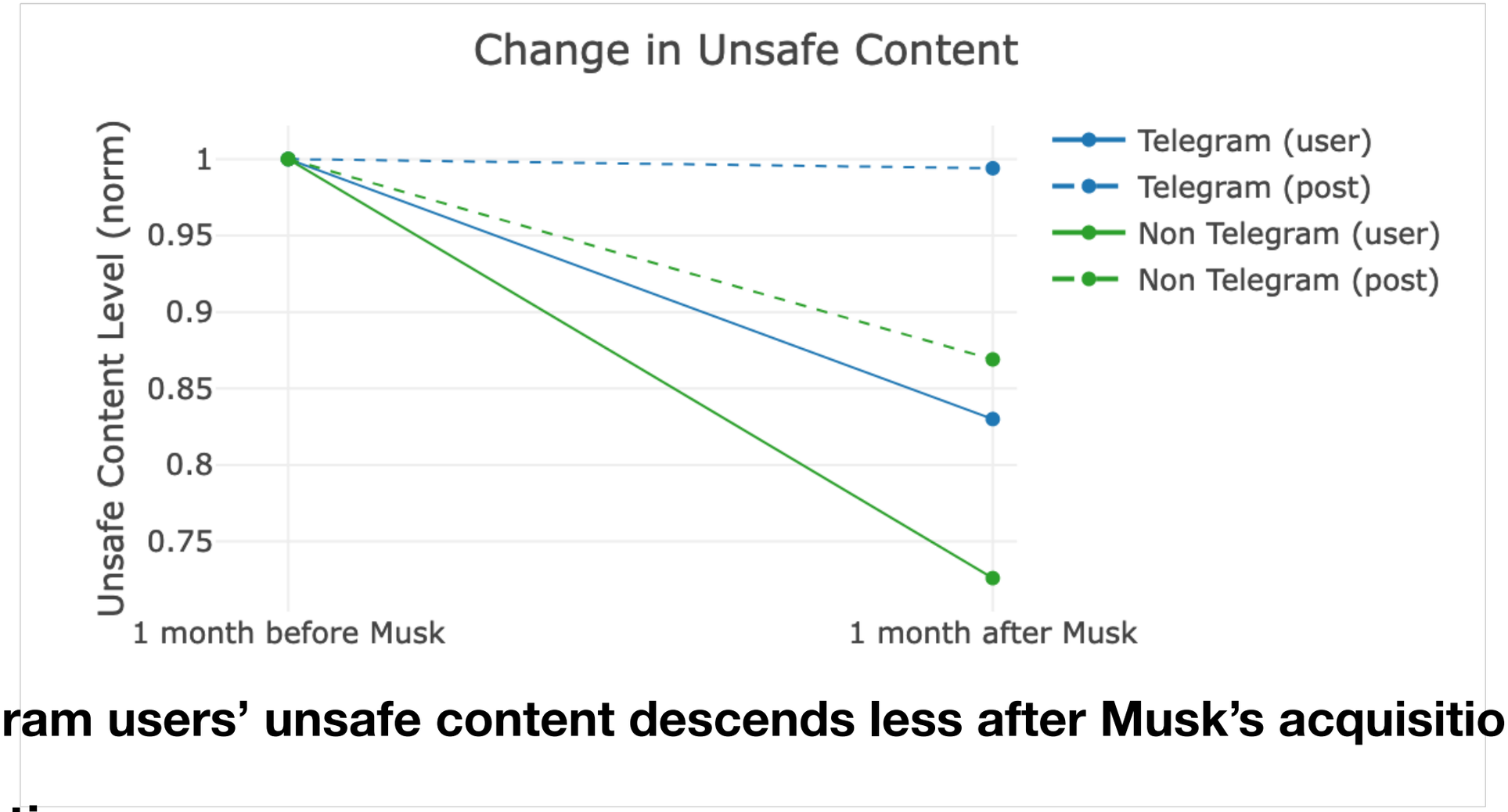


Telegram users’ unsafe content descends less after Musk’s acquisition

Observations:

- Downwards (natural?) trend of the invasion
- Robust to the temporal window chosen
- Discrepancy: mainly due to **activity**
  - … a lot of Telegram-based bots/heavy users

Review of the “Evidence” I Got:

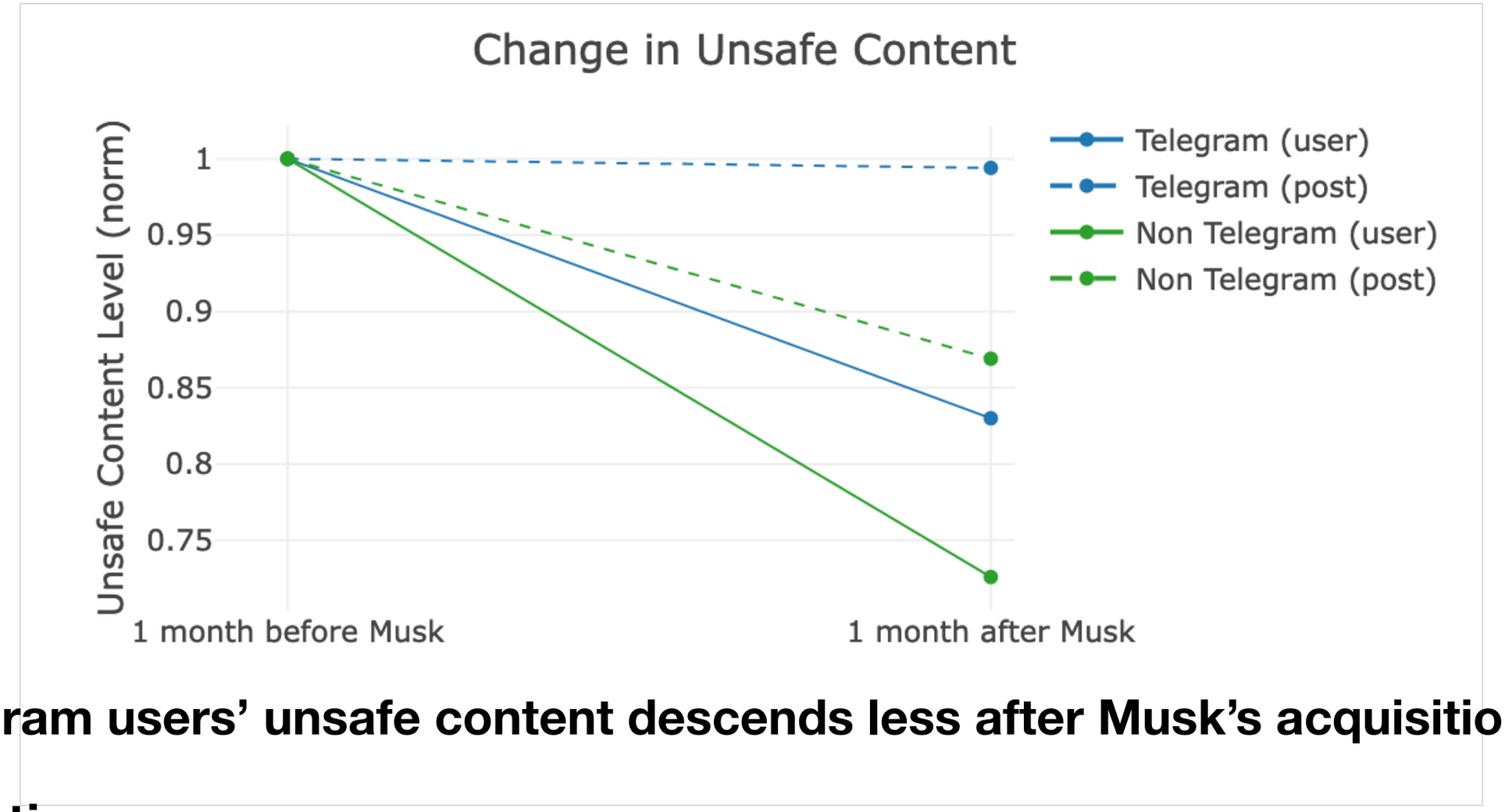


Telegram users’ unsafe content descends less after Musk’s acquisition

Observations:

- Downwards (natural?) trend of the invasion
- Robust to the temporal window chosen
- Discrepancy: mainly due to **activity**
  - … a lot of Telegram-based bots/heavy users
  - Telegram users in both highest and lowest percentiles of unsafe content

Review of the “Evidence” I Got:



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

Observations:

- Downwards (natural?) trend of the invasion
- Robust to the temporal window chosen
- Discrepancy: mainly due to **activity**
  - … a lot of Telegram-based bots/heavy users
  - Telegram users in both highest and lowest percentiles of unsafe content
- Large anomalous activity some specific days for non-TG users

(Lot of) Next Steps…

(Lot of) Next Steps…

Theoretically:

Difficult model to extend (low analytical tractability)

(Lot of) Next Steps…

Theoretically:

Difficult model to extend (low analytical tractability)

Empirically:

Make a more suitable model (structural, with a stochastic part)

Activity =\= Participation?

\+ *Fancy* things to try:

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

Main takeaway

Main takeaway

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

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

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:

- 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**

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:

- 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**

More Important:

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:

- 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**

More Important: **Merry Christmas !**

Appendix

Literature

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

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

Remarks

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

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:

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

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