# Misbehavior and Account Suspension in an Online Financial Communication Platform

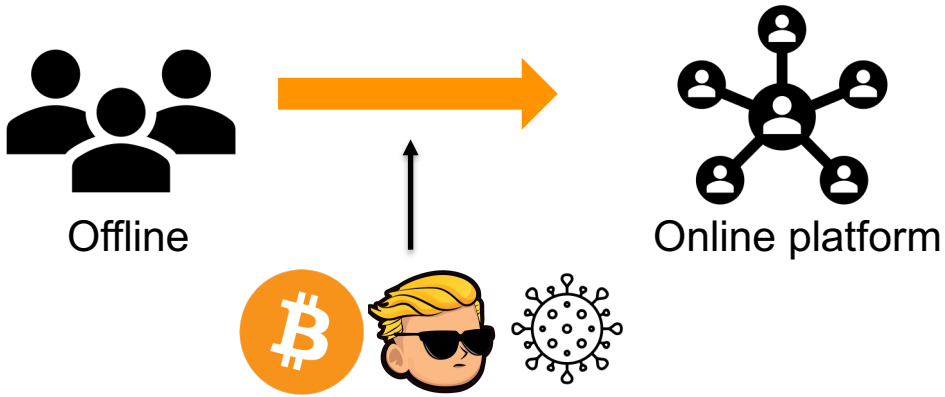
Taro Tsuchiya, Alejandro Cuevas, Thomas Magelinski, Nicolas Christin

*Supported by Carnegie Mellon CyLab's Secure Blockchain Initiative, Nakajima Foundation, and ONR (N00014-21-1-2229), CMU GSA/Provost Conference Funding*



Carnegie Mellon University
Security and Privacy Institute

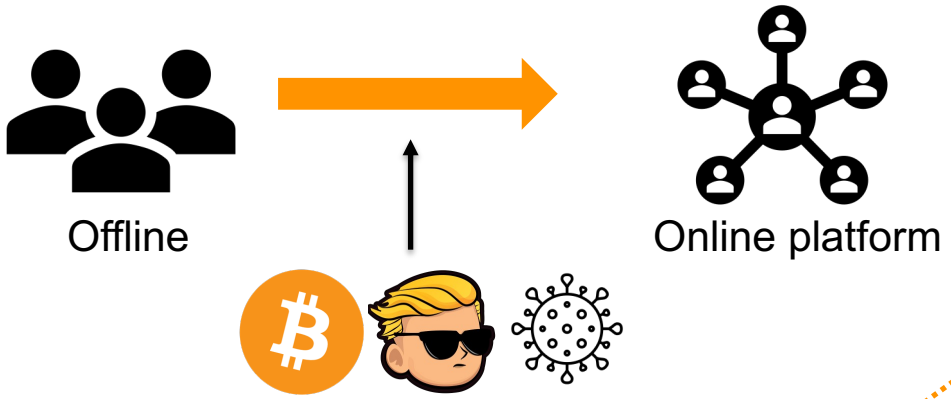
# Structural change in financial communities




the level/type of online misbehavior ↑

Many studies of online misbehavior, but in finance?

# Structural change in financial communities



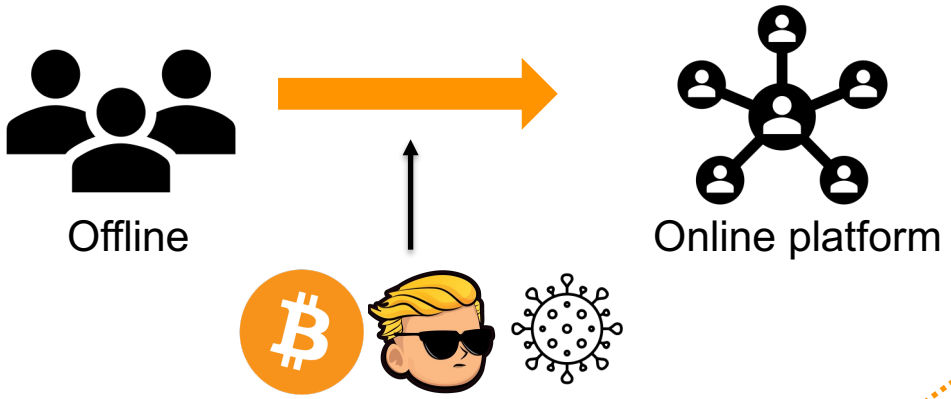

**Spam**

*"<…> Verify your address by sending 2 - 5 ETH to the address below and you will receive 20-50 ETH! This offer will last 2h!* ***Address ETH: 0x6aF562F7343DA3122a71C9350c6Cd6A0eA8107F5****"*

the level/type of online misbehavior ↑

Many studies of online misbehavior, but in finance?

# Structural change in financial communities




the level/type of online misbehavior ↑

**Spam**

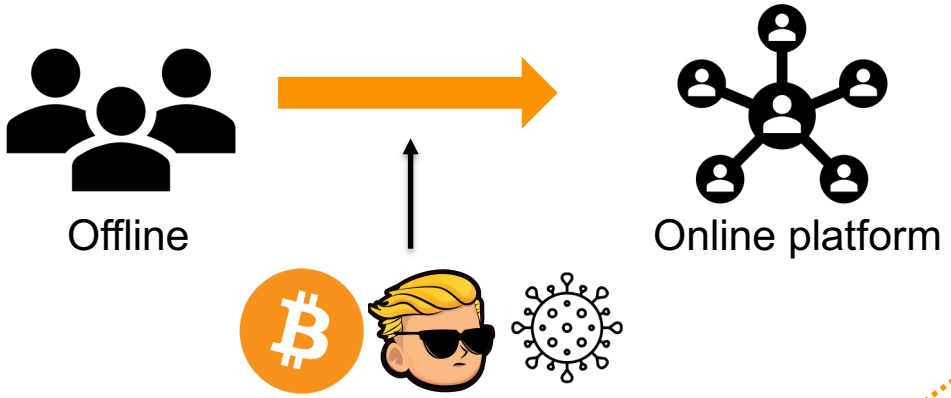
*"<…> Verify your address by sending 2 - 5 ETH to the address below and you will receive 20-50 ETH! This offer will last 2h!* ***Address ETH: 0x6aF562F7343DA3122a71C9350c6Cd6A0eA8107F5****"*

**Toxicity**

*"Stop the sharing!!! bullshit"*

Many studies of online misbehavior, but in finance?

# Structural change in financial communities




the level/type of online misbehavior ↑

**Spam**

*"<…> Verify your address by sending 2 - 5 ETH to the address below and you will receive 20-50 ETH! This offer will last 2h!* ***Address ETH: 0x6aF562F7343DA3122a71C9350c6Cd6A0eA8107F5****"*

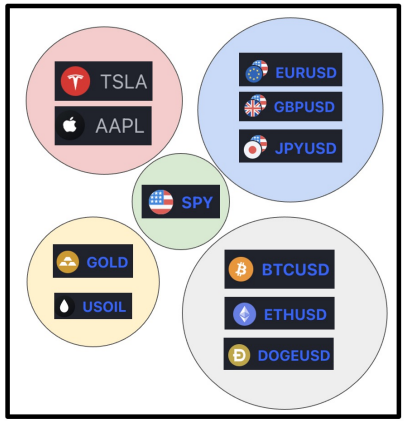
**Toxicity**

*"Stop the sharing!!! bullshit"*

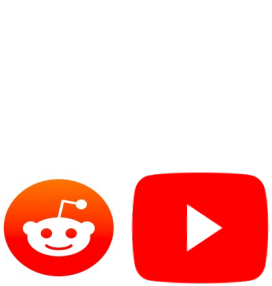
Many studies of online misbehavior, but in finance?

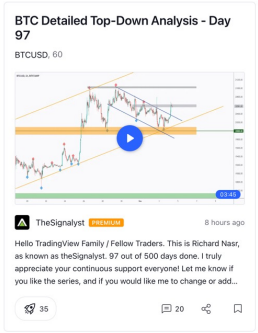
*Question: What are the characteristics of online investment forums and how can we make the platforms safer?*

# We conducted a case study of TradingView

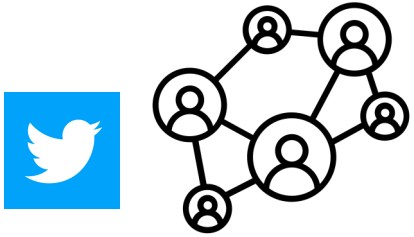



Data provider
(all classes of assets)



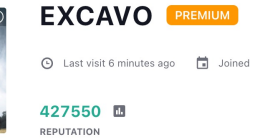



Post articles/scripts
(technical analysis)



Follow/Comment/like others

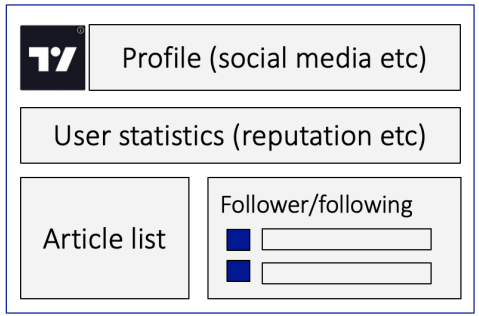
Account suspension
by moderators

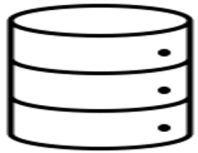



Paying users (“Pro”)

# Data collection thruough snowball sampling (Jul 20th – Aug 8th, 2022)

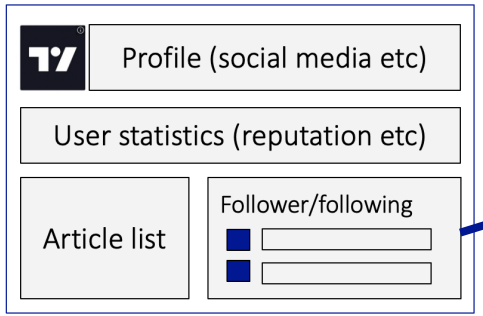
Robust with different initial seeds, # of hops, timeframe

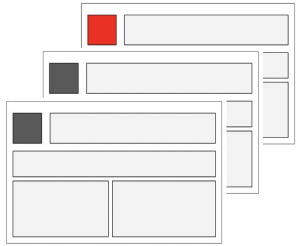


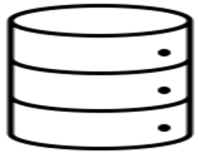
# Data collection thruough snowball sampling (Jul 20th – Aug 8th, 2022)

Robust with different initial seeds, # of hops, timeframe



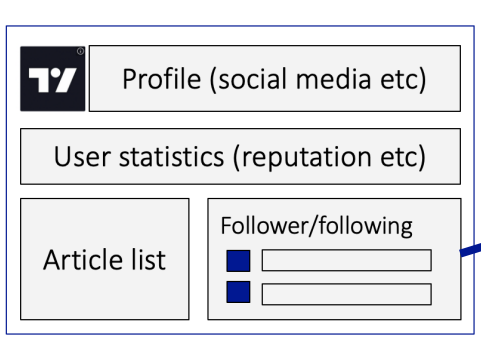
Snowball sampling 3 hops

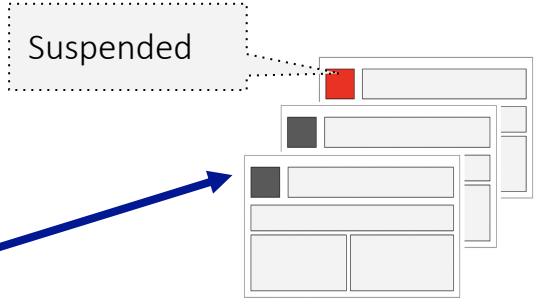


2.76M users' basic profile
200K users' social graph

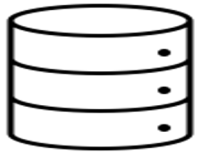
# Data collection thruough snowball sampling (Jul 20th – Aug 8th, 2022)

Robust with different initial seeds, # of hops, timeframe




Snowball sampling 3 hops



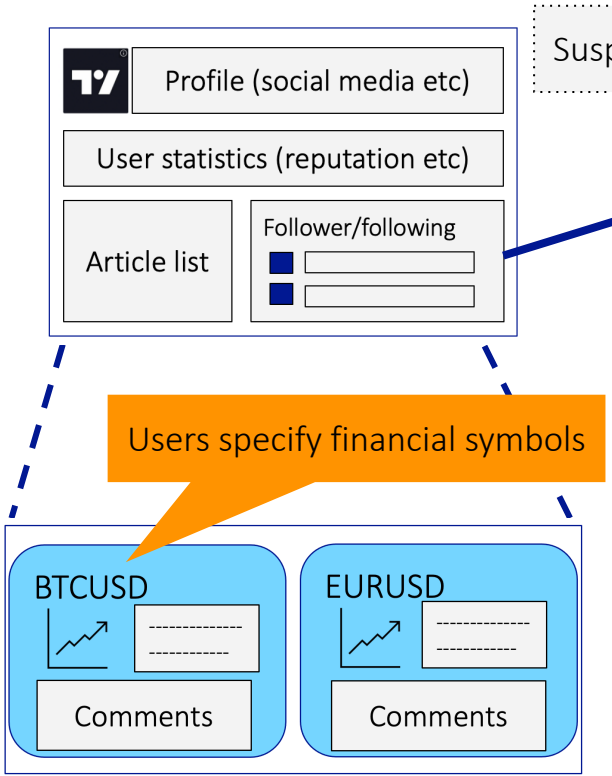
**2.76M** users' basic profile

**200K** users' social graph

- 4K (2%) suspension

# Data collection thruough snowball sampling (Jul 20th – Aug 8th, 2022)

Robust with different initial seeds, # of hops, timeframe

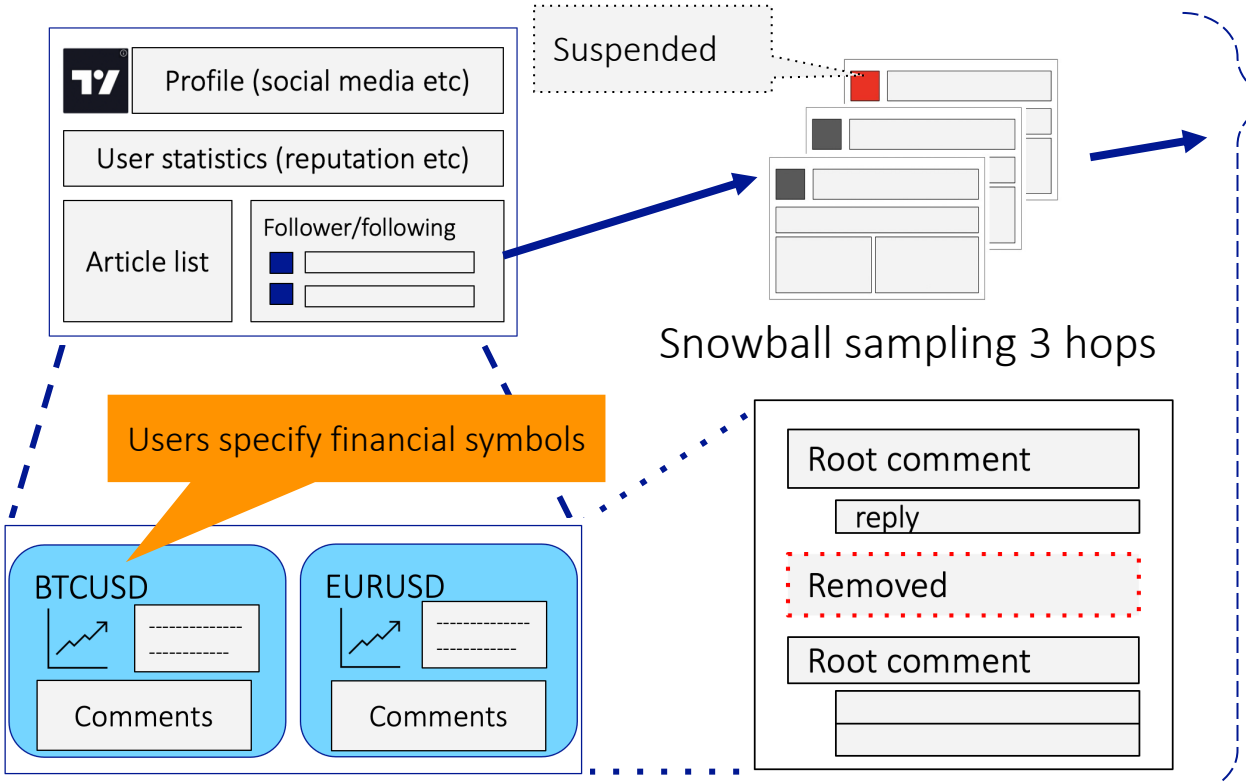



2.76M users' basic profile
200K users' social graph
- 4K (2%) suspension

4.2M articles

# Data collection thruough snowball sampling (Jul 20th – Aug 8th, 2022)

Robust with different initial seeds, # of hops, timeframe




2.76M users' basic profile

200K users' social graph

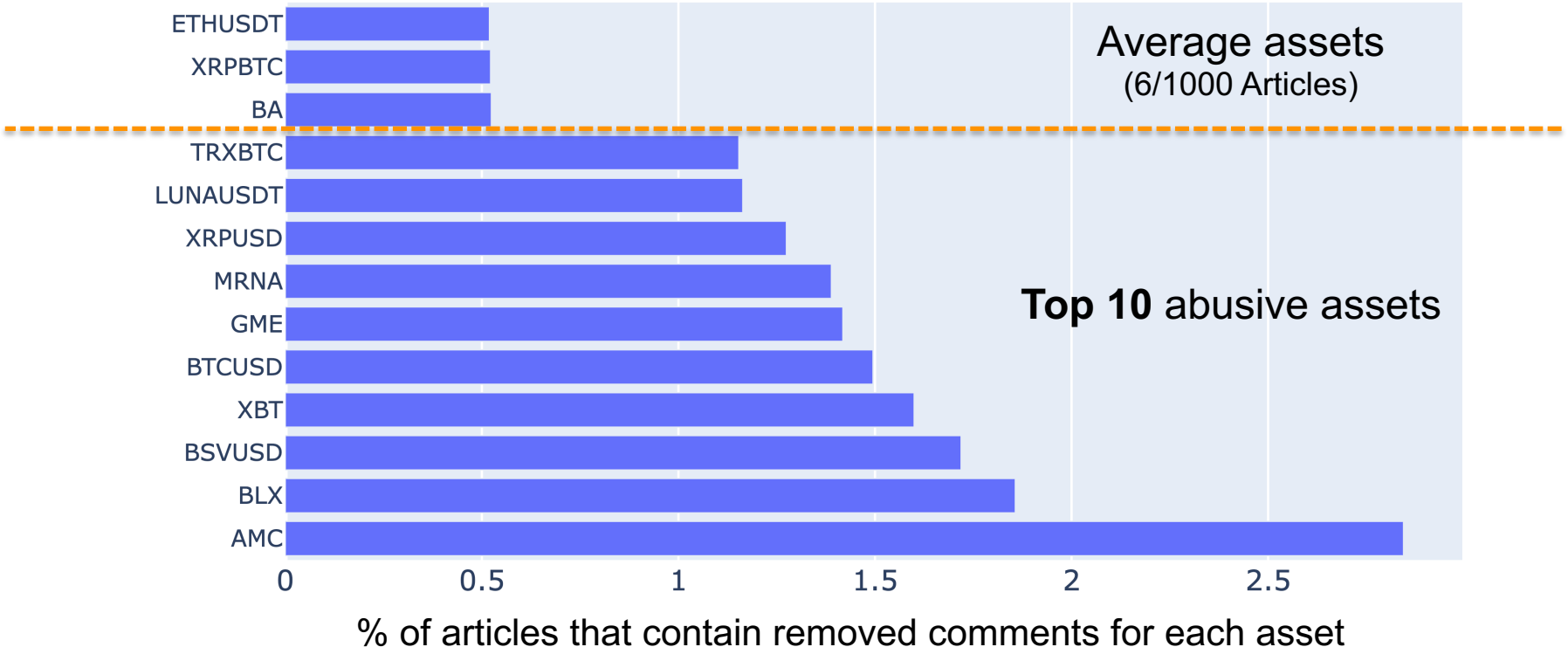
- 4K (2%) suspension

4.2M articles

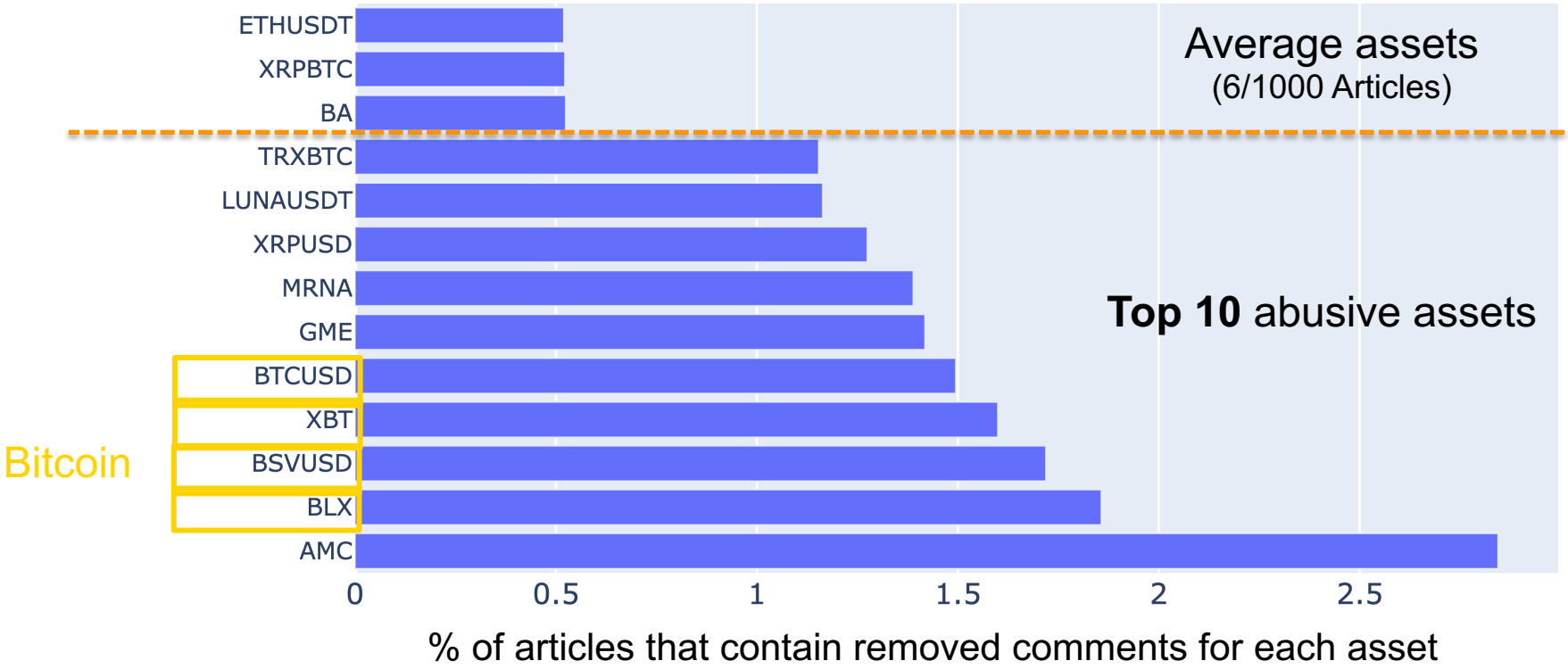
5.3M comments

- 17K (0.3%) removed
- Spam (35.8%), Toxicity (30.8%)

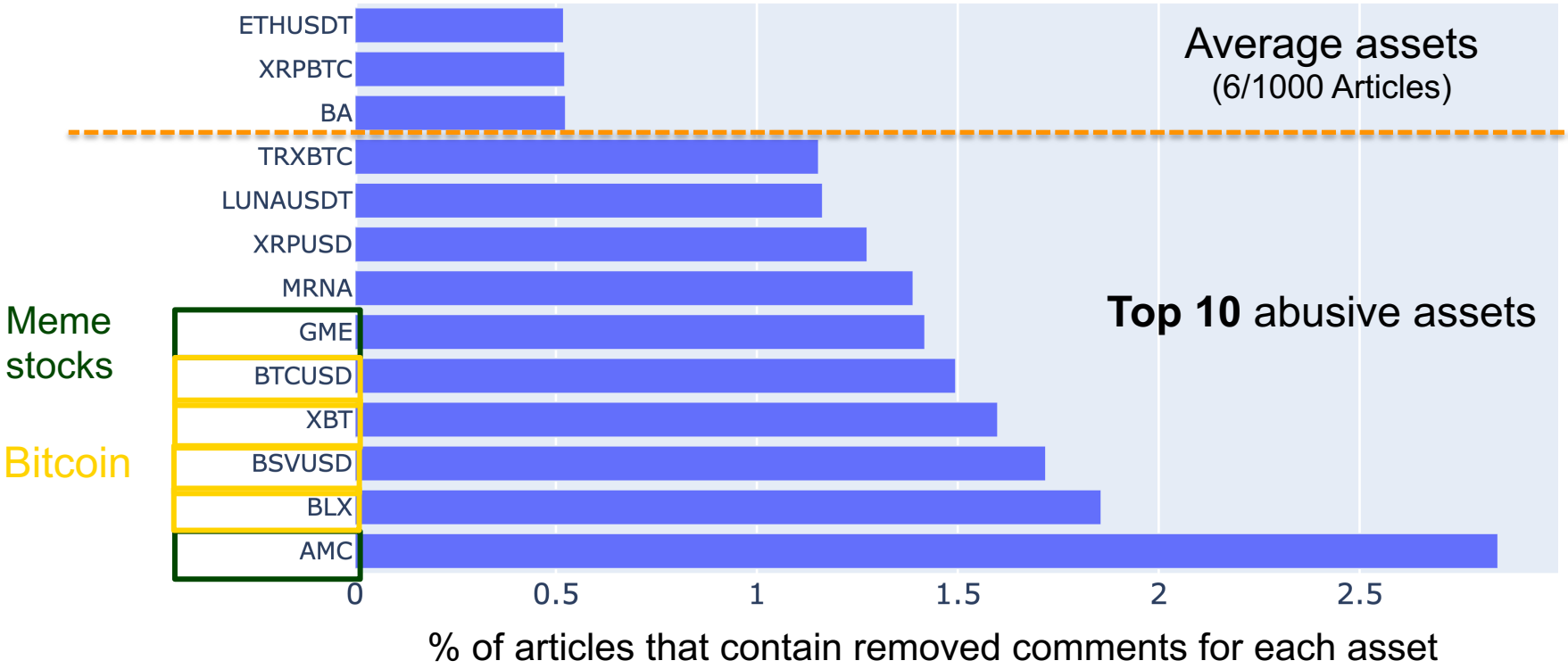
# 1. There is a connection between market and online abuse



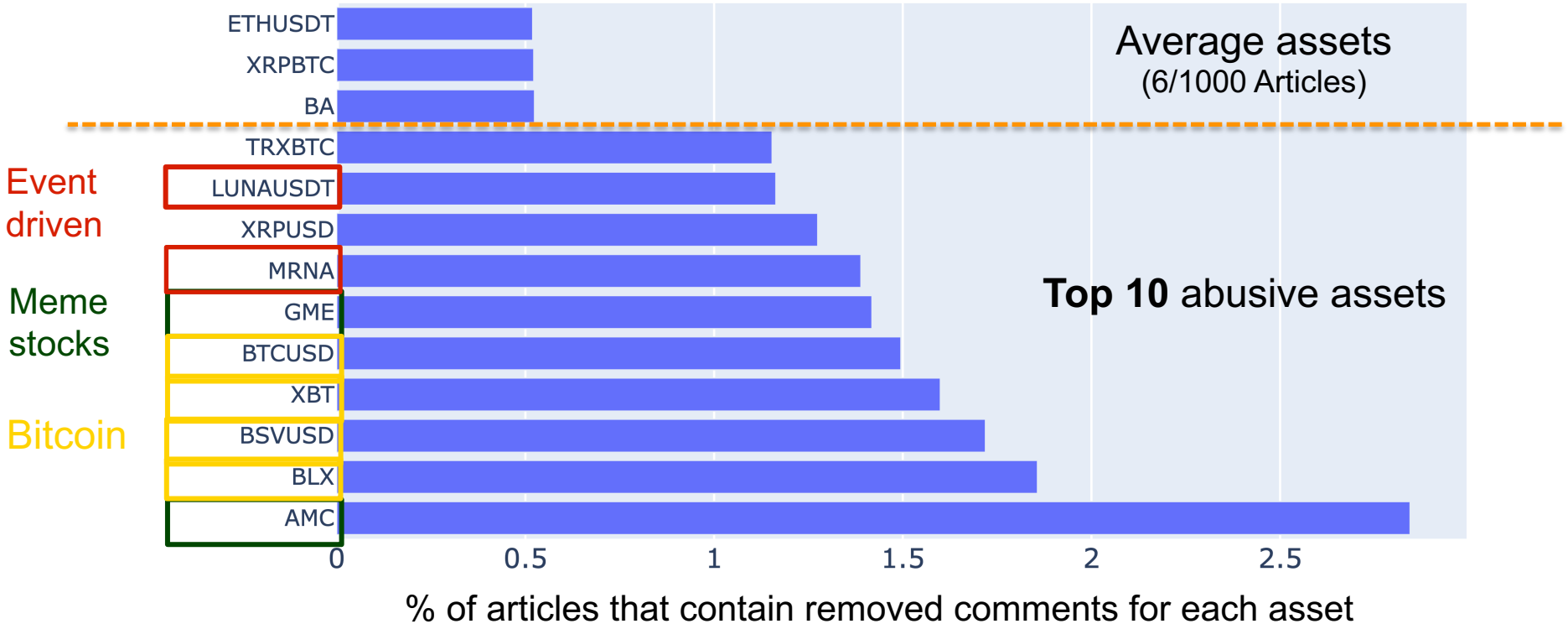
# 1. There is a connection between market and online abuse



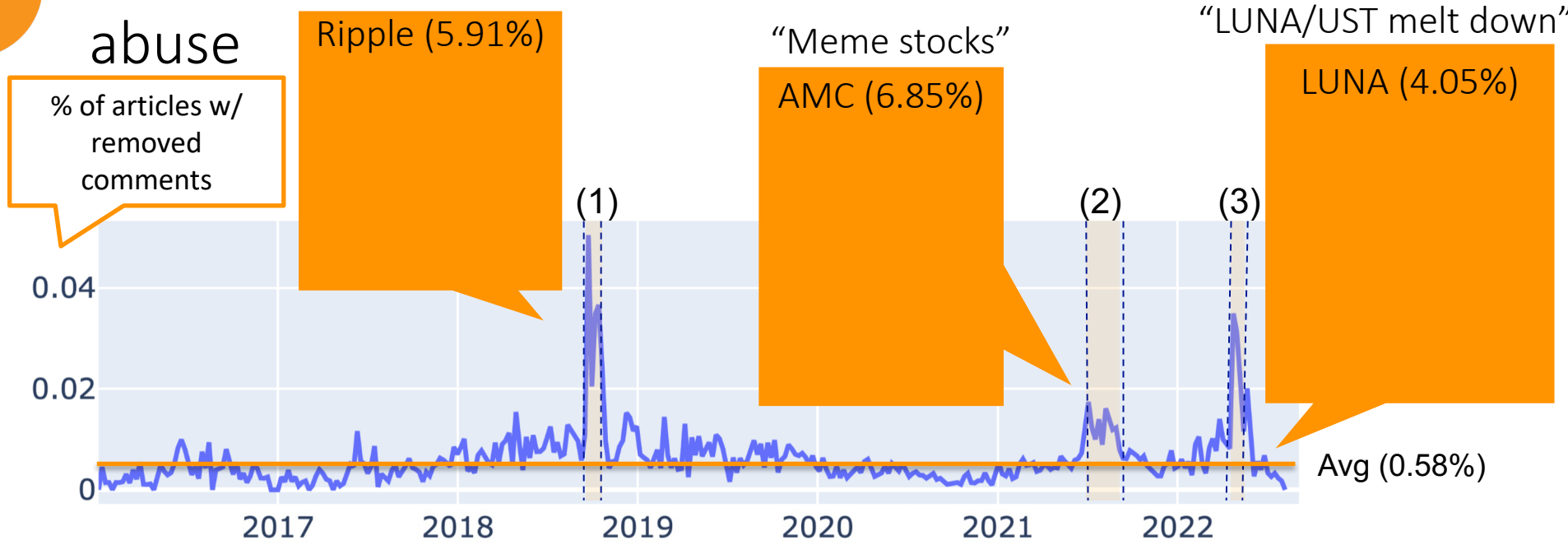
# 1. There is a connection between market and online abuse



# 1. There is a connection between market and online abuse

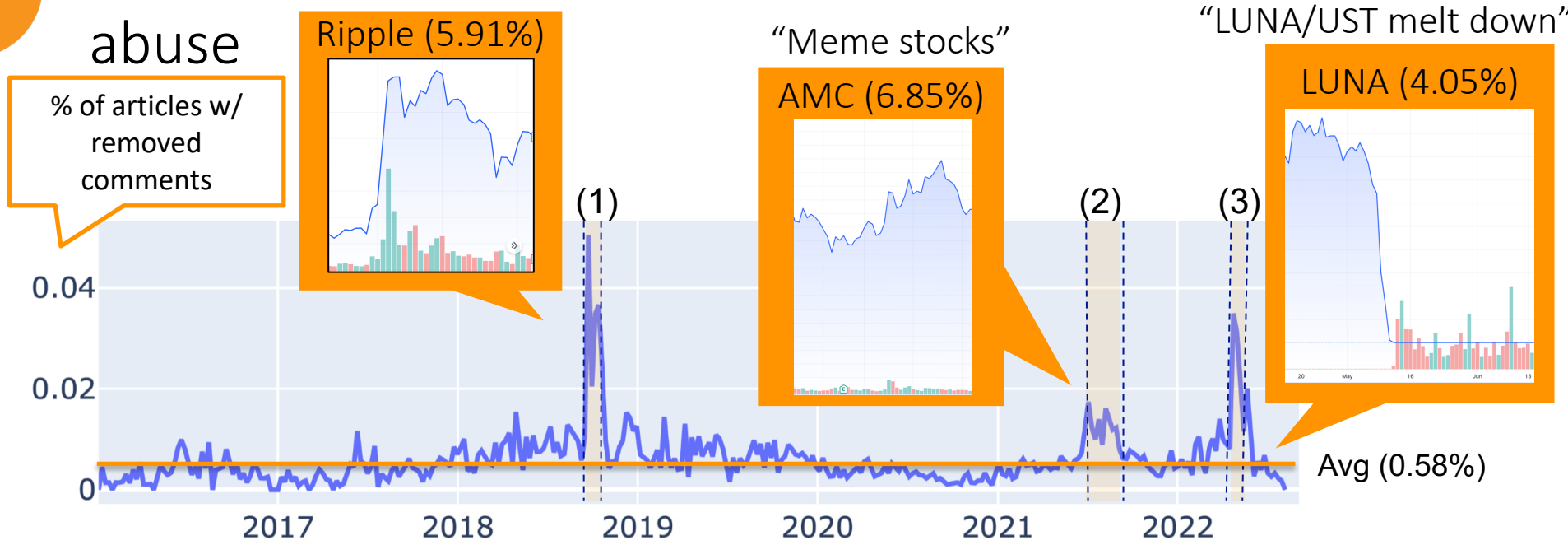


# 1. There is a connection between market and online abuse




Type of assets & Price fluctuation & Level of misbehavior

# 1. There is a connection between market and online abuse

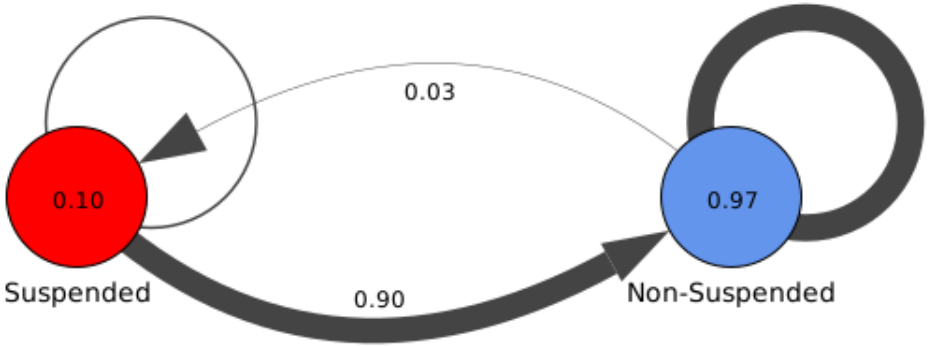



Type of assets & Price fluctuation & Level of misbehavior

# 2. Suspended users tend to form more closely-knit communities

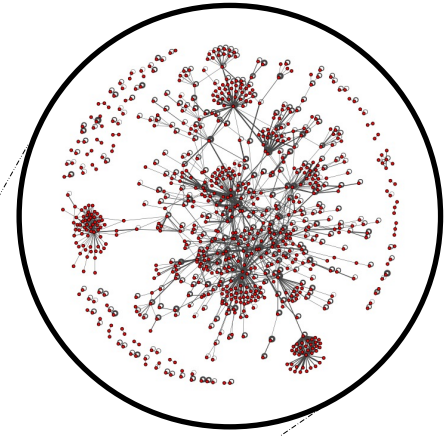
- Suspended accounts are **3 times** more likely to comment on other suspended accounts
- Same results from network metrics (**low following quality**, **high ego density**)

Consistent with the other studies on account suspension on online platforms
(Cao et al. 2014, Yang et al. 2012, Le et al. 2019)

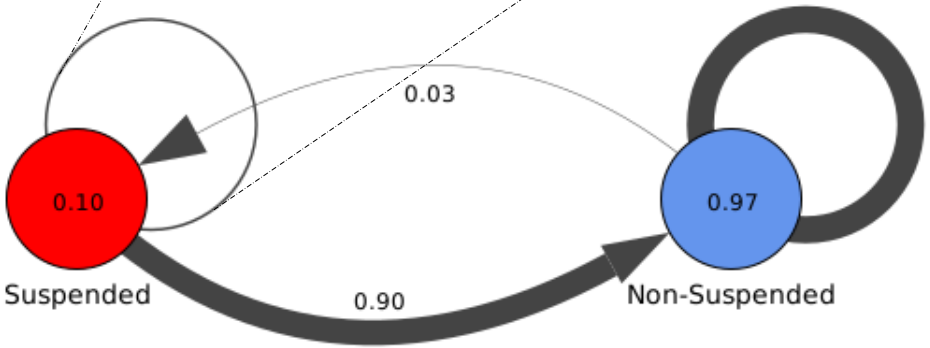


# 2. Suspended users tend to form more closely-knit communities

**Node**: Only suspended accounts (2%)
**Edge**: Comments b/w them



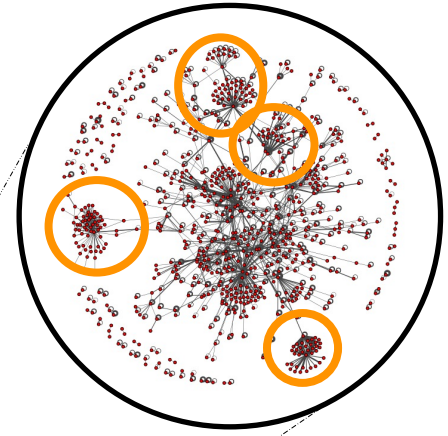
- Suspended accounts are **3 times** more likely to comment on other suspended accounts
- Same results from network metrics (**low following quality**, **high ego density**)



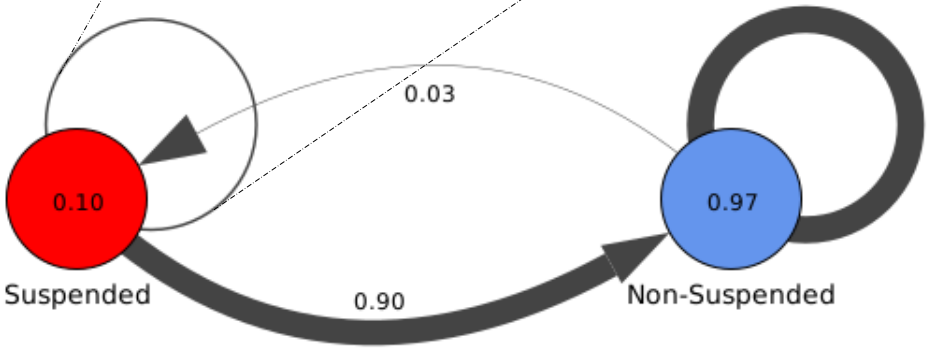

Consistent with the other studies on account suspension on online platforms
(Cao et al. 2014, Yang et al. 2012, Le et al. 2019)

# 2. Suspended users tend to form more closely-knit communities

**Node**: Only suspended accounts (2%)
**Edge**: Comments b/w them



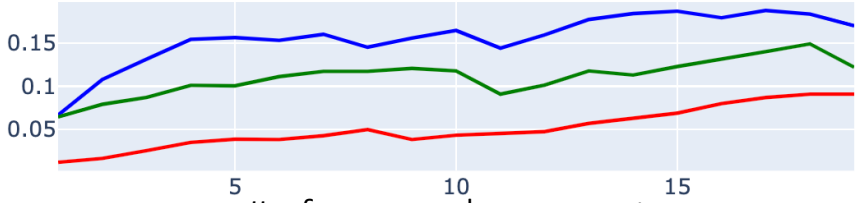
- Suspended accounts are **3 times** more likely to comment on other suspended accounts
- Same results from network metrics (**low following quality**, **high ego density**)



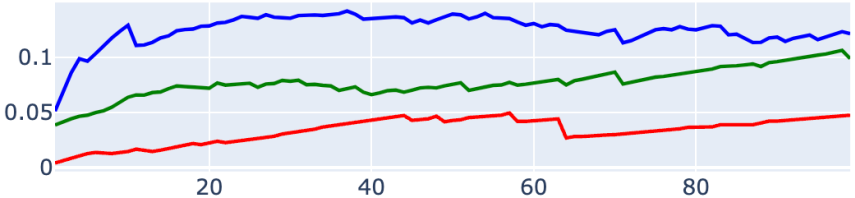

Consistent with the other studies on account suspension on online platforms
(Cao et al. 2014, Yang et al. 2012, Le et al. 2019)

# 3. Paying/reputable accounts are less likely to be suspended even if they have the same amount of violations

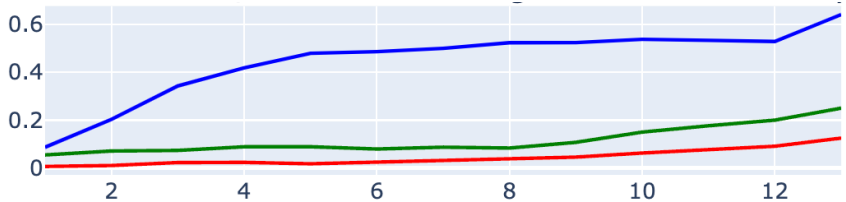
Y-axis: Likelihood (ratio) of suspension for those with more than $x$ violations




# of removed comments




# of moderated articles



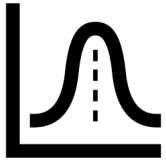

(proxy) # of fake accounts created

--- All the accounts
--- Top 5% of reputation score
(internally calculated by TradingView)
--- All pro accounts (pro/pro+/premium)

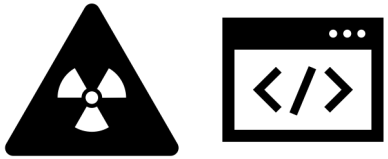
- Statistical difference w/ regression
- Fairness of content moderation !?

# Other contributions

Platform characteristics
(Sec. 4)



Suspension prediction --Stats/ML
(Sec. 6 & Appx. 3)



Removed comments
--Toxicity, URL (Sec 5.1)

Defense mechanism (Sec 7)

# Conclusion

First in-depth research on online misbehavior on the financial forum

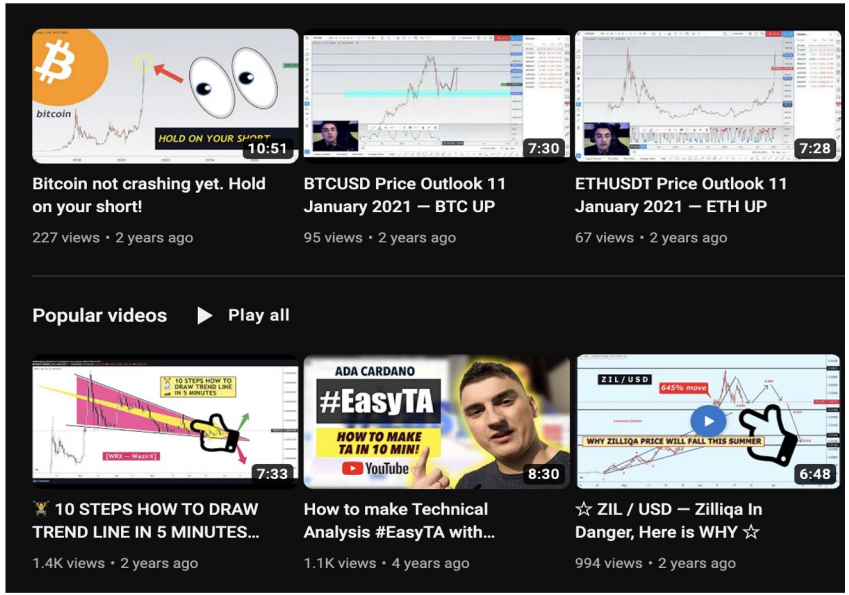
1. There is some association between the level of abuse and the market → new
2. Suspended accounts form a closely knit communities → consistent
3. Paying users are less likely to be suspended → new

--> Adopt current content moderation process and make the platform safer

Email: ttsuchiy@cs.cmu.edu

# Appendix

# Online communication




[13:38] BullSoldier: part*

[13:39] wassa_wassa_bitconee: BullSoldier: fooking universe alignment bro. That's why only I noticed. These low IQ bitches will get fooked and hand over their all money to us.

[13:39] BullSoldier: wassa_wassa_bitconee: hahaha exactly ❤
lets hope so

[13:40] NasQk: rought world

[13:41] wassa_wassa_bitconee: BullSoldier: I have nothing to lose. I am anyway a 9-5 slave. If we are wrong then nothing changes. But if it works then fook the world. 😂

[13:42] NasQk: do a job that you enjoy to do bro
world can't work with everyone on vacation

[13:43] BullSoldier: wassa_wassa_bitconee: exactly me too hahaha if it works in may i will tell my boss to gently caressoff
9-5 killing me

[13:44] Fuun: does it still hurt, kids?
its hard to be useless, right?

[13:44] NasQk: yes

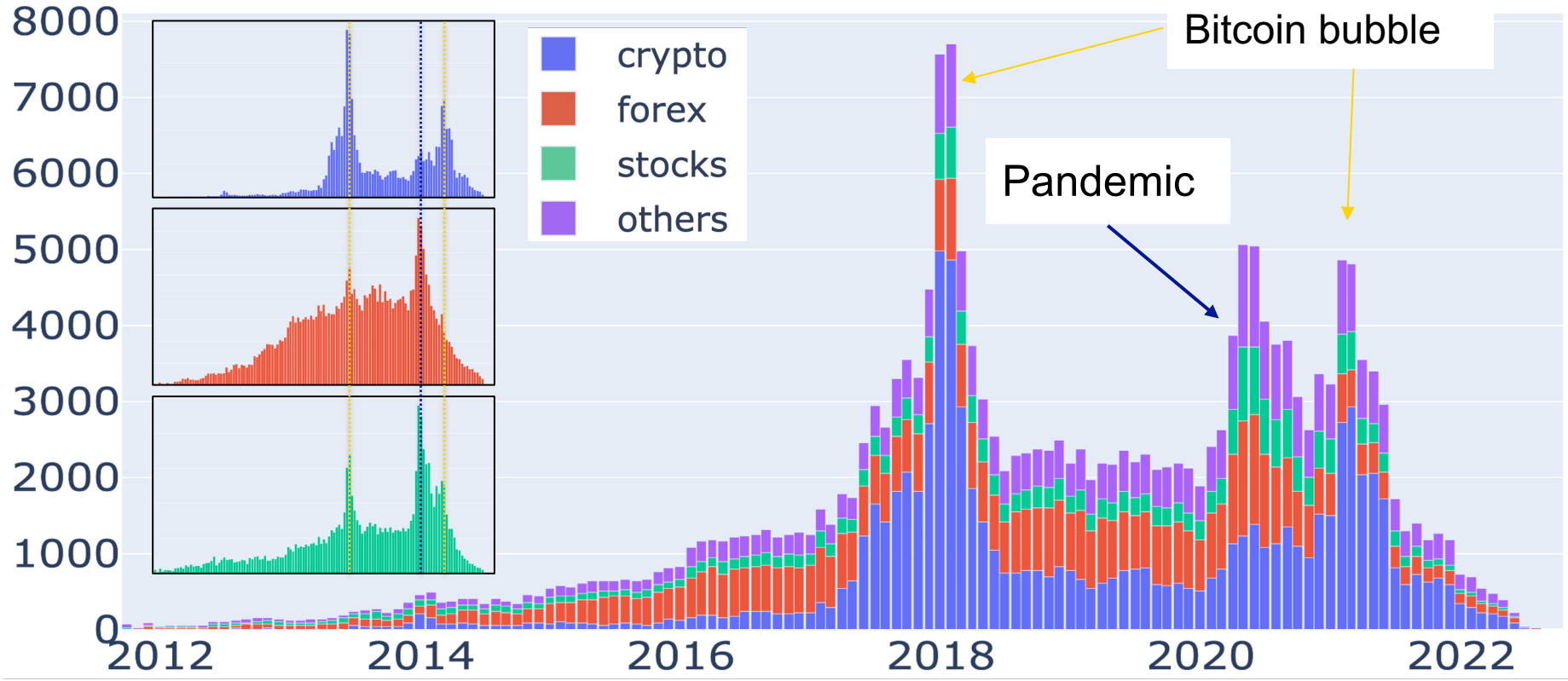
[13:44] Fuun: 😂

[13:45] Buckztr: Savages btc https://www.tradingview.com/x/uNIQAGO3 1hr bully btc

[13:50] Fuun: sad individuals, but relax.
still.

[13:51] Buckztr: https://www.tradingview.com/x/jd63pHsk 1hr btc bull
Snotty Nose Rez Kids - The Warriors https://www.youtube.com/watch?v=1lyZlj1
Ts btc
https://www.tradingview.com/x/d7UR3JbW 1hr bull btc check vol trend..trendz change

# The platform got popular with crypto bubble and covid-19



# Spam and toxicity are most prevalent

Manually label **removed** comments (κ= 0.734)

## Spam (35.8%)

*"Nice work mate, <…>*
*I had the option to make 9.4bTc across the board month from executing/replicating exchanges signals with tips and data from Mrs Regiane on* ***Telegram @R.e.g.i.a.n.e.c.r.e.s.t,"***

*"<…> Verify your address by sending 2 - 5 ETH to the address below and you will receive 20-50 ETH! This offer will last 2h!* ***Address ETH: 0x6aF562F7343DA3122a71C9350c6Cd6A0eA8107F5"***

## Toxicity (30.8%)

Sub-category "Insults": 81.8%
> Reddit (Kumar et al. 2023)

*"Stop the sharing!!! bullshit"*

*"this chart gave me cancer"*

## Undefined (33.4%)

*"Good work"*

Reputation manipulation??

# What are the reasons of comment removal?

| Category | Ratio | Break down | Ratio | Reddit* |
|---|---|---|---|---|
| Toxicity | 30.8% | Insult | 81.8% | 63.4% |
| | | Identity Attack | 5.8% | 14.2% |
| | | Call to leave | 9.1% | 12.0% |
| | | Threat | 2.0% | 5.5% |
| | | Sexual Aggression | 1.3% | 2.8% |
| | | Identity Misrepresentation | 0.0% | 1.6% |
| Spam | 35.8% | | | (Kumar et al, 2022) |
| Undefined | 33.4% | | | |

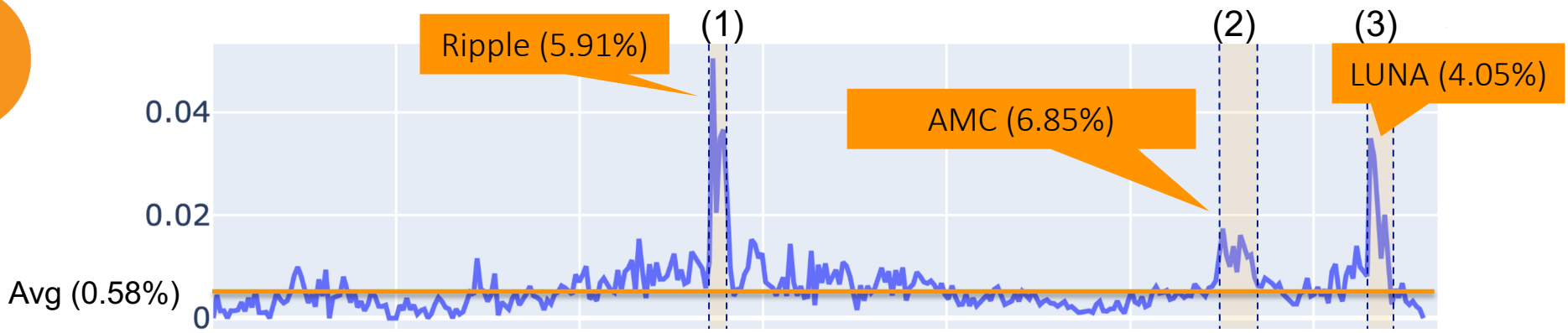
Cohen's kappa = 0.734

# URL shortener is still more likely to be in removed comments

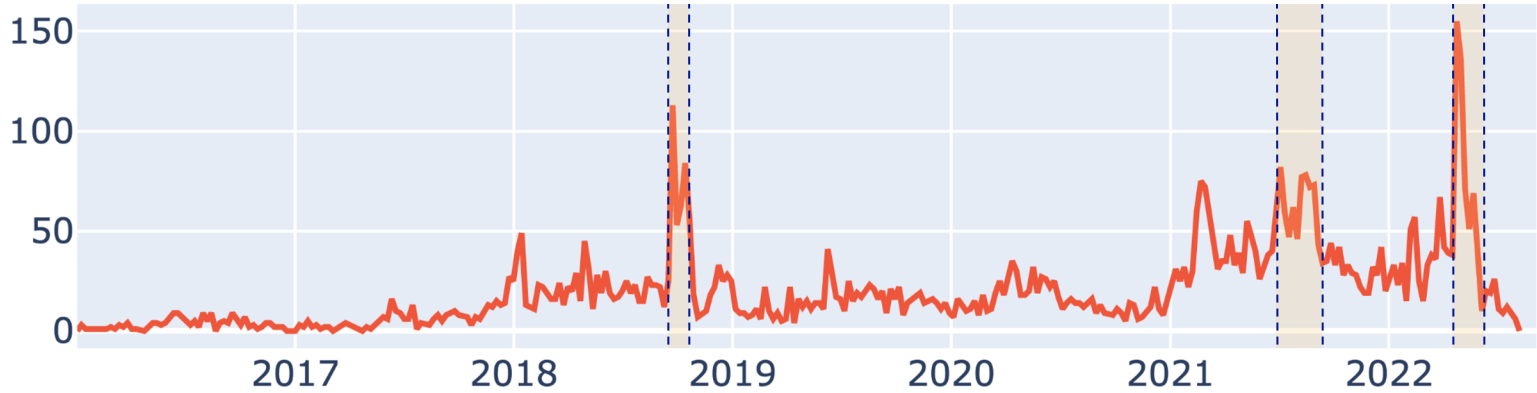
| Domain | Count | | Probability | | Likelihood ratio* |
|---|---|---|---|---|---|
| | Removed (0.3%) | Normal (99.7%) | Removed | Normal | R/N |
| bit.ly | 88 | 218 | 0.525% | 0.0041% | 126.80 |
| tinyurl | 18 | 45 | 0.108% | 0.001% | 125.64 |
| is.gd | 71 | 2 | 0.424% | 0.000% | 11150.87 |
| goo.gl | 16 | 277 | 0.096% | 0.005% | 18.14 |
| invst.ly | 3 | 207 | 0.018% | 0.004% | 4.55 |
| all | 196 | 734 | 1.171% | 0.014% | **83.88** |

*(Thomas et al, 2011)

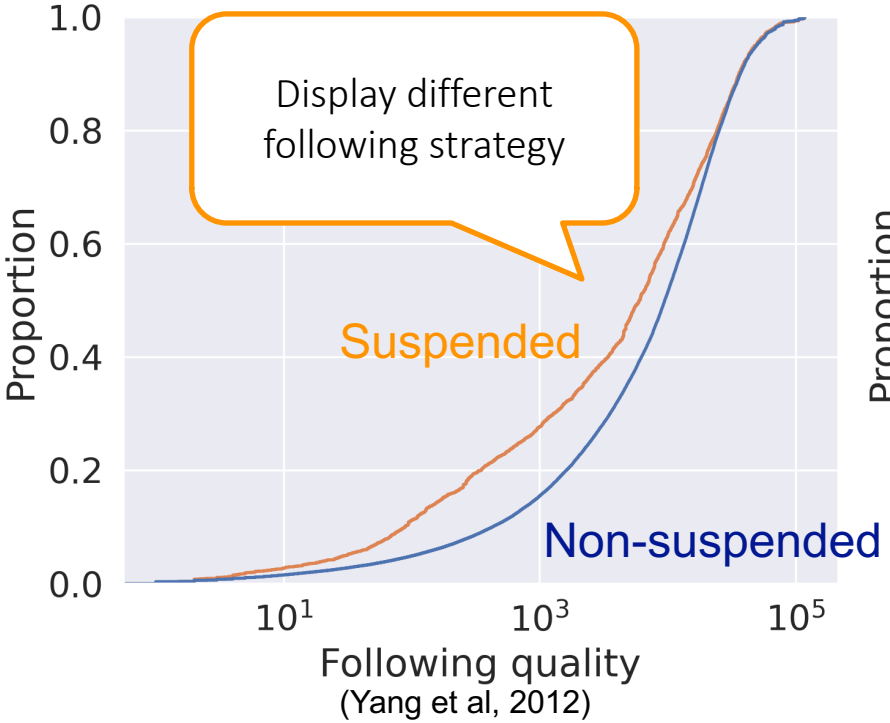
The ratio of articles with removed comments

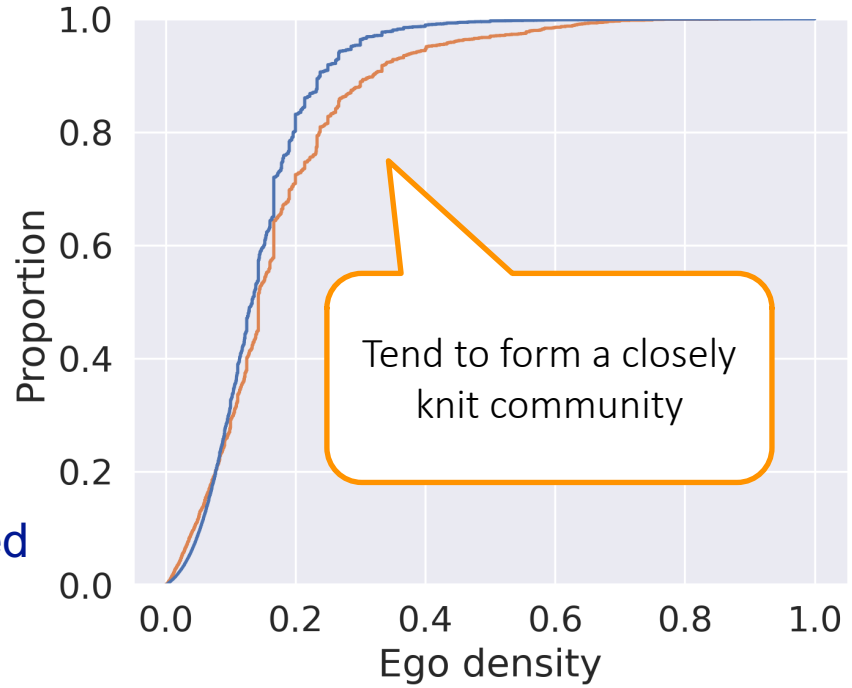



# of articles with removed comments



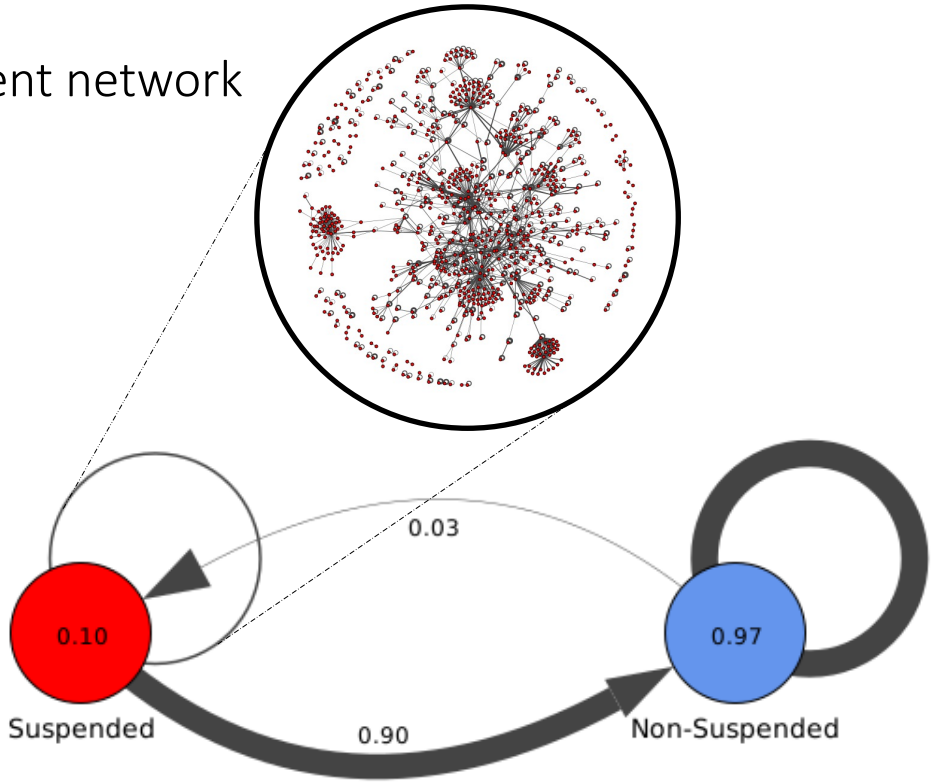
# Suspended user have different characteristics

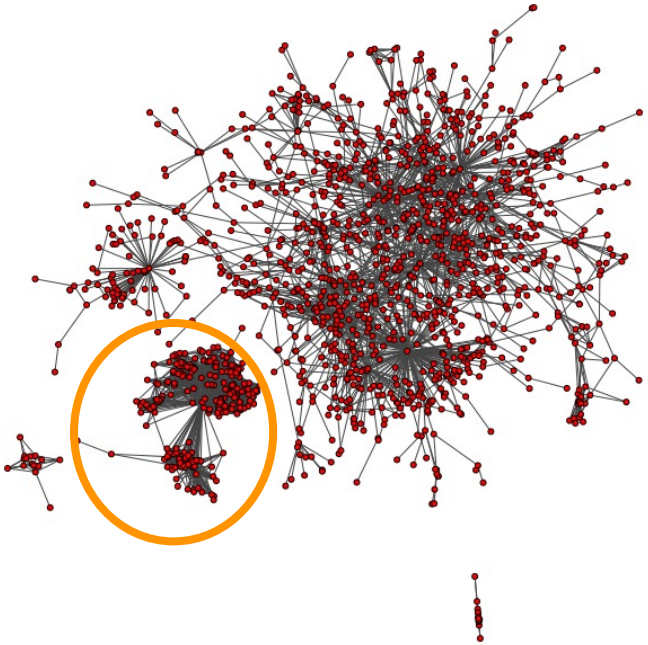


# Suspended user tends to form more closely-knit communities

Comment network



Following network

# Predict suspension by logit model

**Independent variables**

Violations

- # of removed messages
- # of moderated articles
- # of fake accounts connected (proxy from the **registration date**)

Pro users (pro or free)

Reputation tier (20% quantile)

Registration year

Interaction variables (types of violations * pro users)

**Logit model**
**Dependent variable**
Suspended (1)
Not suspended (0)

Does the level / type of violations lead to suspension differently depending on pro/free user?

# Fairness of content moderation?

All types of violations (+)

Pro user (-)

Reputation tier

Registration year

Interaction variables (-)

- # of removed messages
- # of moderated articles

Significant at 1% level

| variable | | coef. | std err | p-value |
|---|---|---|---|---|
| $\beta_1$ | const | -4.1073 | 0.055 | 0.000 |
| $\beta_2$ | # of removed messages | 0.0353 | 0.009 | **0.000** |
| $\beta_3$ | # of moderated articles | 0.0183 | 0.001 | **0.000** |
| $\beta_4$ | # of fake accounts (proxy) | 0.5051 | 0.018 | **0.000** |
| $\beta_5$ | All pro users | -2.3388 | 0.103 | **0.000** |
| $\beta_6$ | Tier 1 (Top -20%) | 0.5484 | 0.055 | **0.000** |
| $\beta_7$ | Tier 2 (Top 20-40%) | 0.3394 | 0.057 | **0.000** |
| $\beta_8$ | Tier 3 (Top 40-60%) | 0.0585 | 0.058 | 0.316 |
| $\beta_9$ | Tier 4 (Top 60-80%) | -0.373 | 0.064 | **0.000** |
| $\beta_{10}$ | Registered 2017-2018 | 0.1943 | 0.047 | **0.000** |
| $\beta_{11}$ | Registered 2018-2019 | 0.3217 | 0.051 | **0.000** |
| $\beta_{12}$ | Registered 2019-2020 | 0.2483 | 0.049 | **0.000** |
| $\beta_{13}$ | Registered 2021- | -0.0566 | 0.064 | 0.374 |
| $\beta_{14}$ | # of removed messages * pro | -0.0105 | 0.013 | 0.421 |
| $\beta_{15}$ | # of moderated articles * pro | -0.0145 | 0.002 | **0.000** |
| $\beta_{16}$ | # of fake accounts * pro | -0.405 | 0.05 | **0.000** |

# Possible mitigations for each attack

Toxicity

ML toxic detection with domain adaptation

Spam/Fraud

Use of URL shortener will help

Reputation manipulation

Look at their attributes (following quality, ego density, registration date ) or network (comment/following)