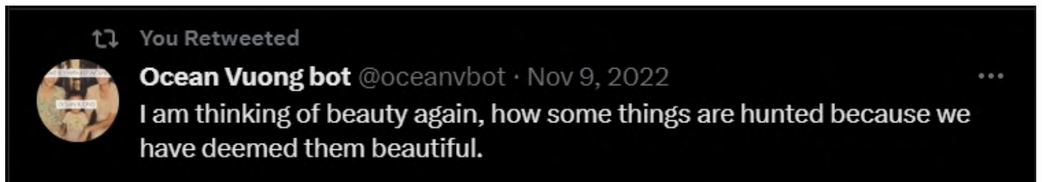
Spread of Hate Speech on Twitter



What is Hate?

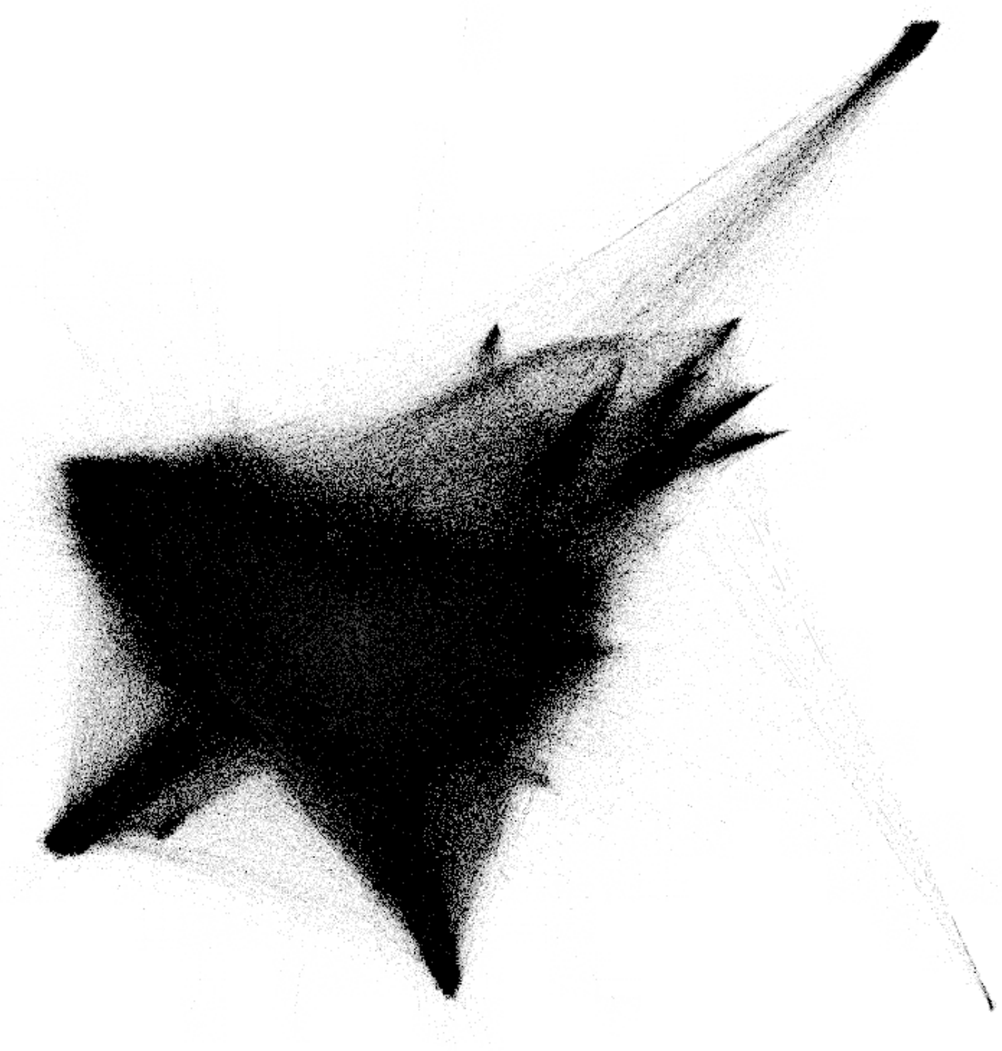
And why is this a SNA problem?

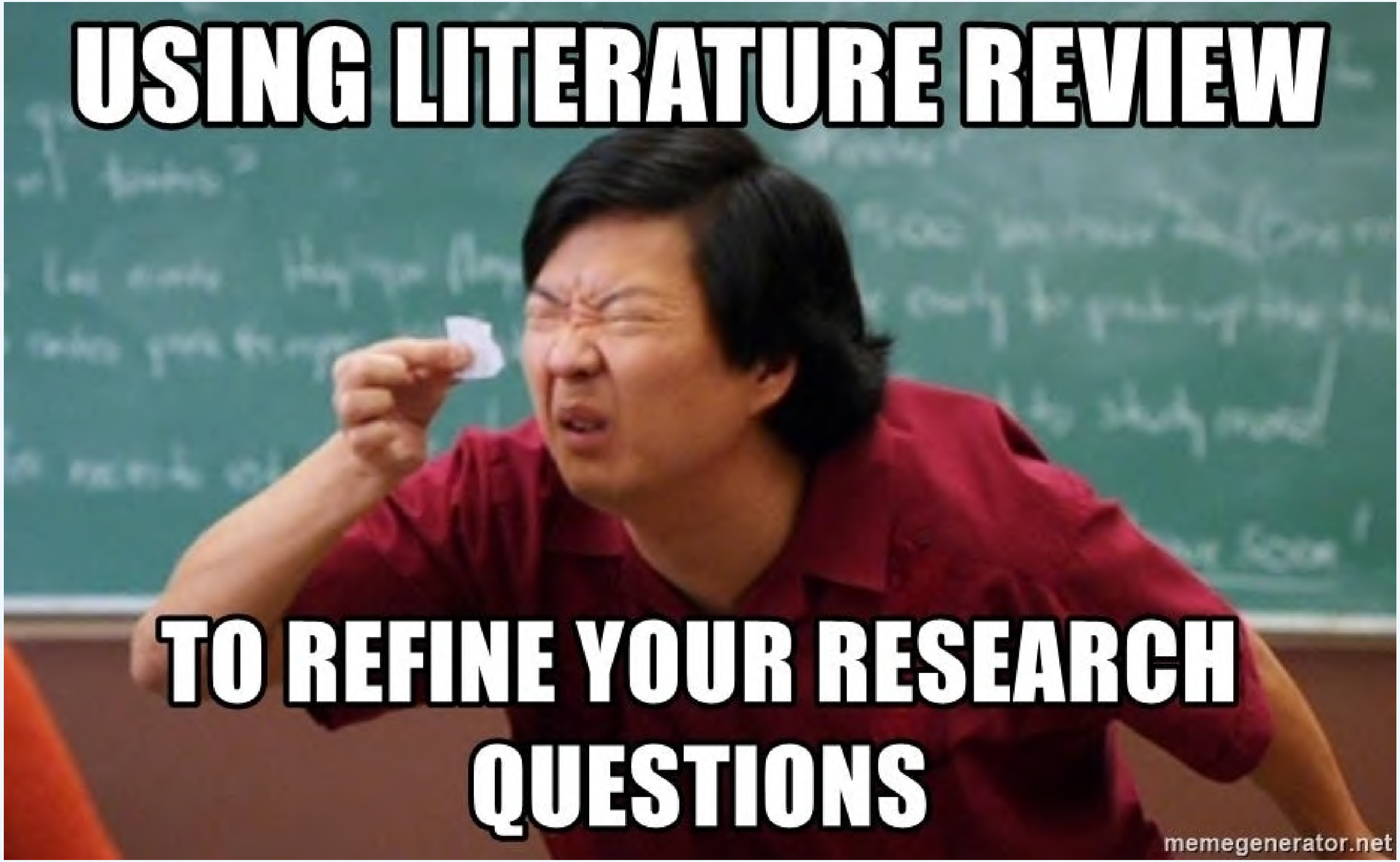
Hate Speech is a very broad problem and is context dependent. Each country has different laws of hate speech and how they deal with it from a policy perspective.

We define "***Toxicity***" as a *rude disrespectful unreasonable comment, that is likely to make someone leave a conversation.*

Hate with Sarcasm - not from a lexicon:

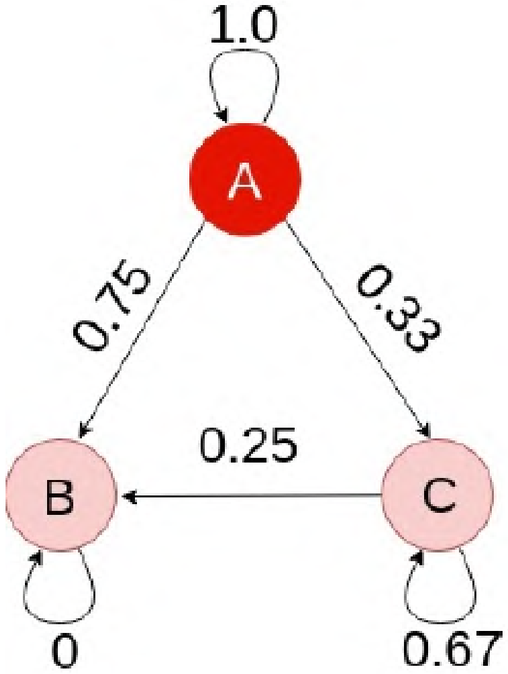
> *"Who convinced Muslim girls they were pretty?"*




USING LITERATURE REVIEW
TO REFINE YOUR RESEARCH QUESTIONS
memegenerator.net

Review of Literature - Previous Work

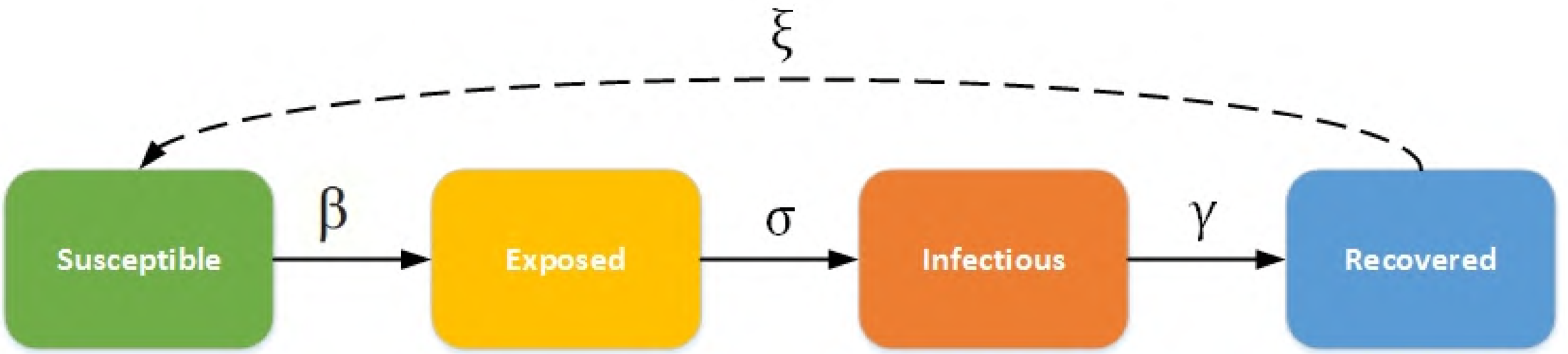
- **Spread Activation Modelling**
  - The theory of spreading activation proposes that the activation of a semantic memory node may spread along bidirectional associative links to other related nodes.
  - Topic modelling - detecting topics using LDA (latent topic detection)
- **Belief Propagation**
  - Belief propagation is used to model how users' beliefs are influenced by their neighbors' posts and reposts
  - Based on the Gab Dataset.
  - The authors found that posts made by hateful users tend to spread farther, faster, and wider than those made by non-hateful users.



Nagar, S., Gupta, S., Barbhuiya, F. A., & Dey, K. (2022). Capturing the Spread of Hate on Twitter Using Spreading Activation Models. In Complex Networks & Their Applications X: Volume 2, Proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021 10 (pp. 15-27). Springer International Publishing.

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019, June). Spread of hate speech in online social media. In Proceedings of the 10th ACM conference on web science (pp. 173-182).

SIR* Models



- dS/dt = - bSE + fR
- dE/dt = bSE + cR - dE
- dI/dt = dE - eI
- dR/dt = eI - fR - cR

Yo, P. & Liu, S. (2006). SEIR epidemic model with delay. The ANZIAM Journal, 48(1), 119-134.

Dataset → Ribeiro Dataset

"Like Sheep Among Wolves": Characterizing Hateful Users on Twitter

What does the dataset look like?

Stats	
Rows	19.58 M
Columns	24

After **Perspective API** →

Jan 2017 - Oct 2017

Stats	
Rows	17.22 M
Columns	26

user_id	qt_flag	week
tweet_id	rt_flag	month
tweet_creation	rt_text	Toxicity
tweet_text	NewDate Format	Severe Toxicity


I DON'T KNOW
HOW TO "CLEAN" DATA
AND AT THIS POINT
I'M TOO AFRAID TO ASK
imgflip.com

Twitter Re-Tweet Graph

Nodes - 100,386 Users → 99,986 Users

Edges - 2.28 M → Directed

Clustering Co-efficient - 0.056

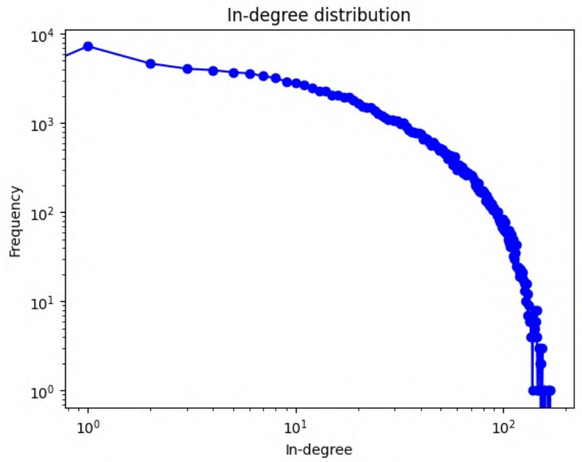
Assortativity - 0.104

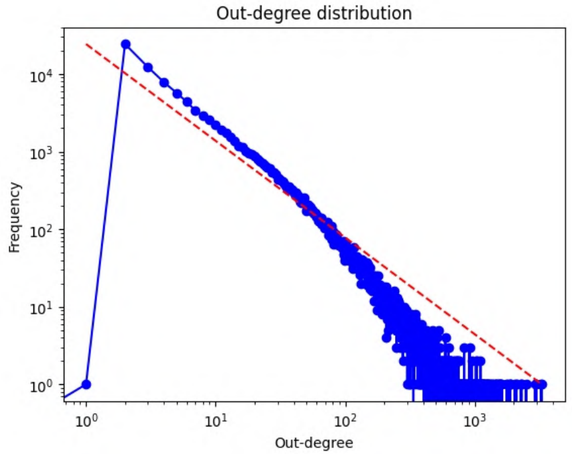
Bow-Tie Structure

- Strong - 91,914
- In - 8,471
- Out - 1

Modularity - 0.62

- Number of Communities: 18

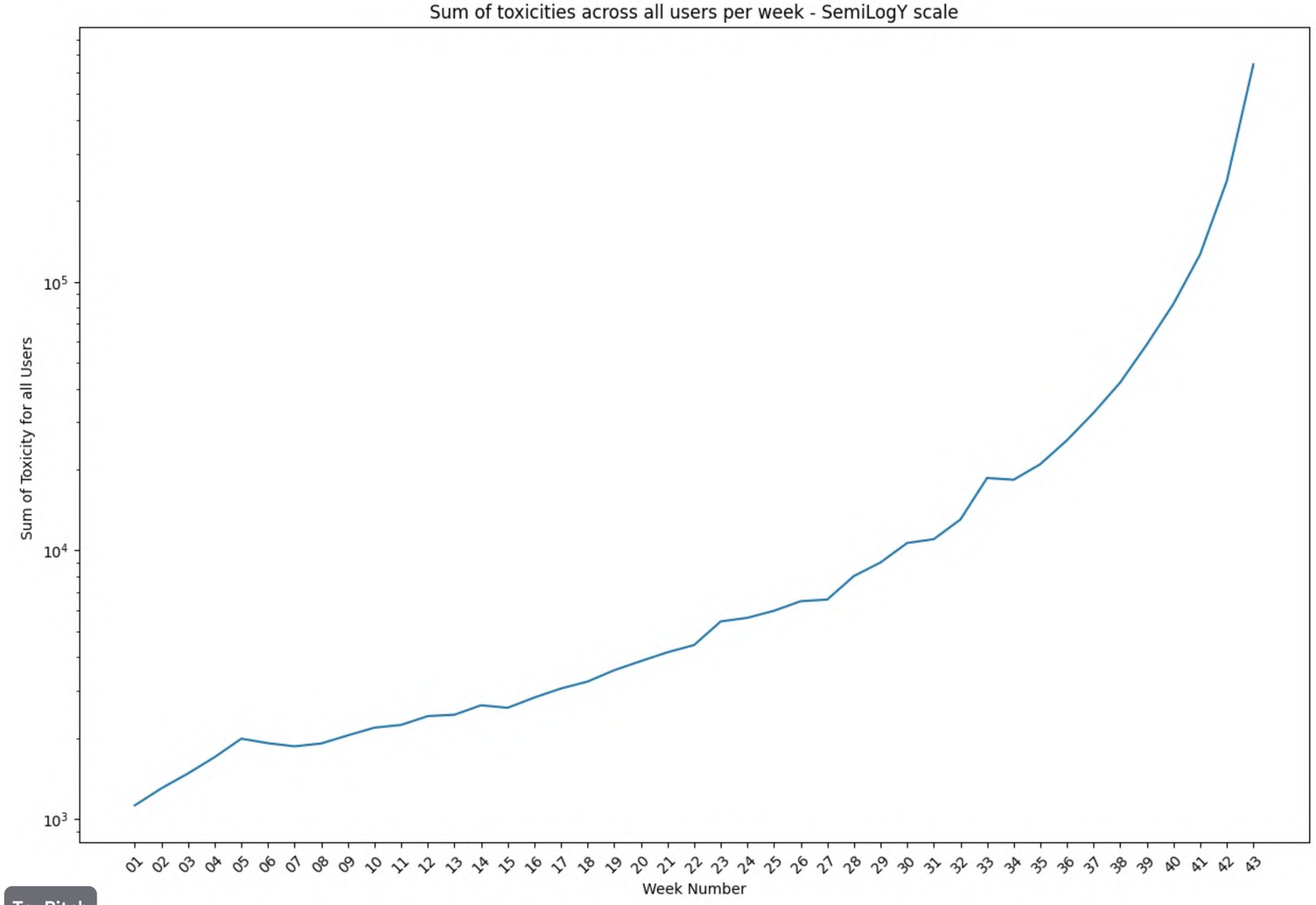


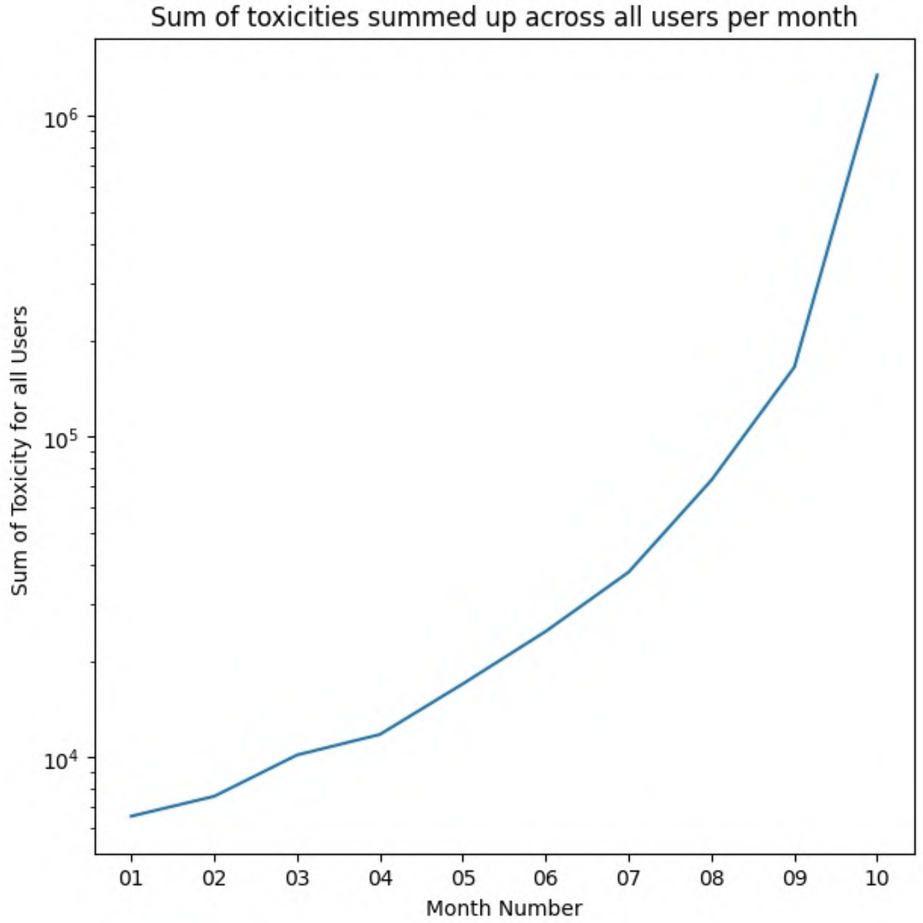


Try Pitch

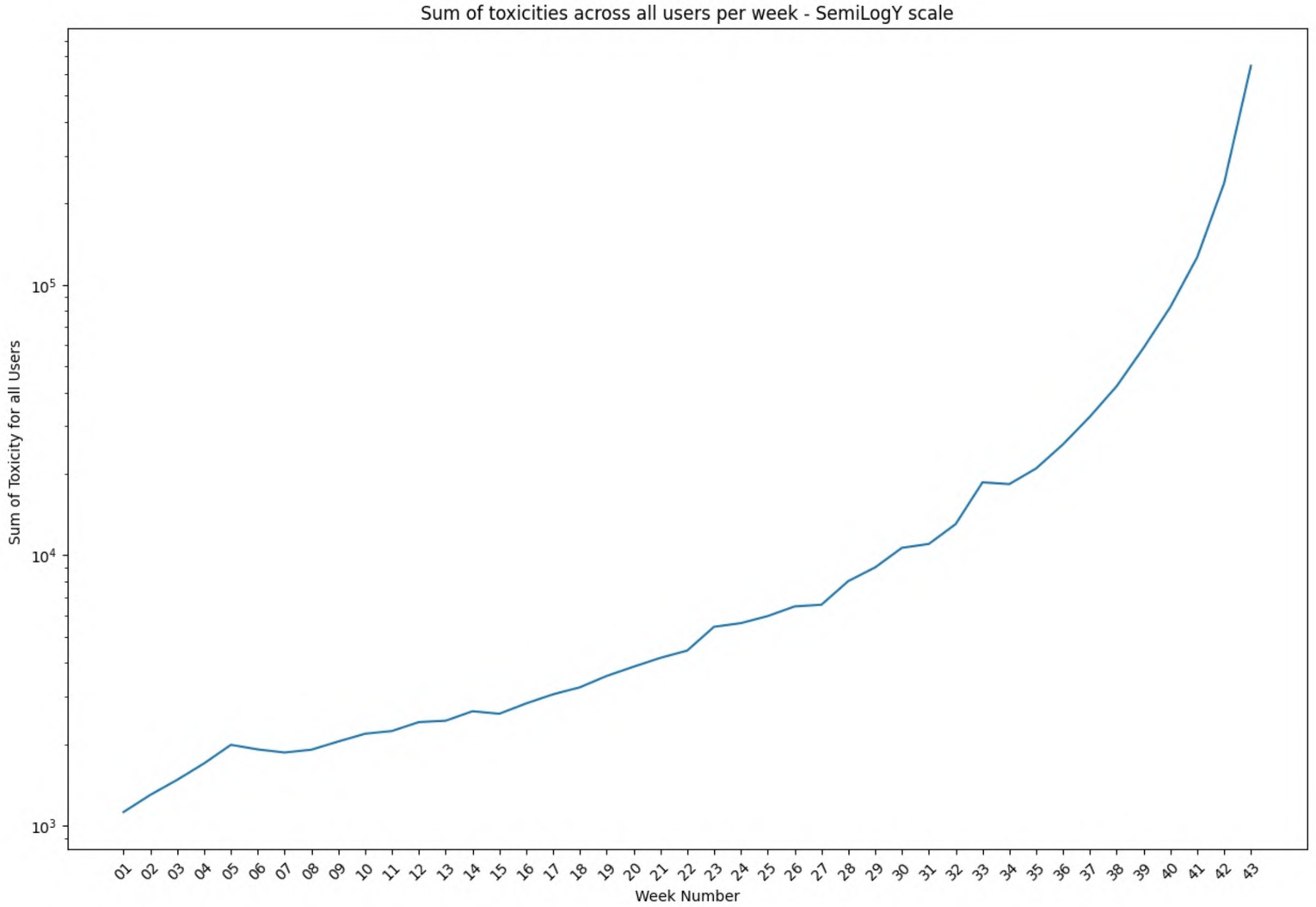
Analysis of the Dataset

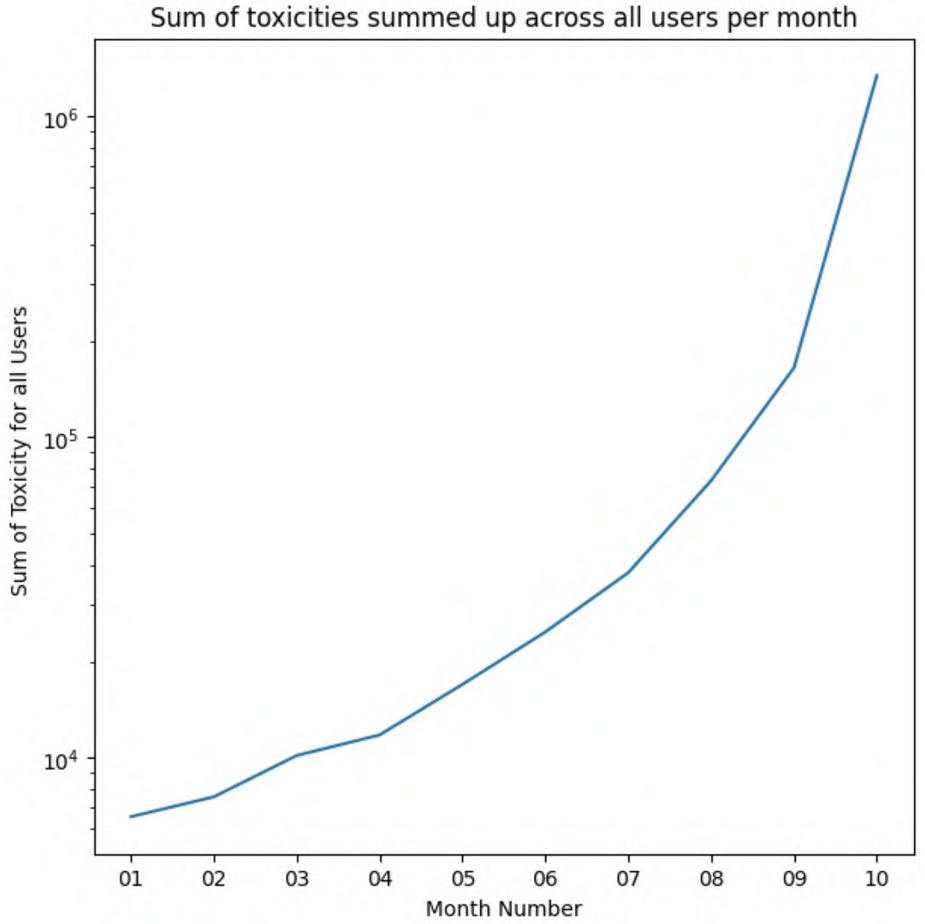
What is happening to the "*Total Toxicity*" in the network?





What is happening to the "*Total Toxicity*" in the network?

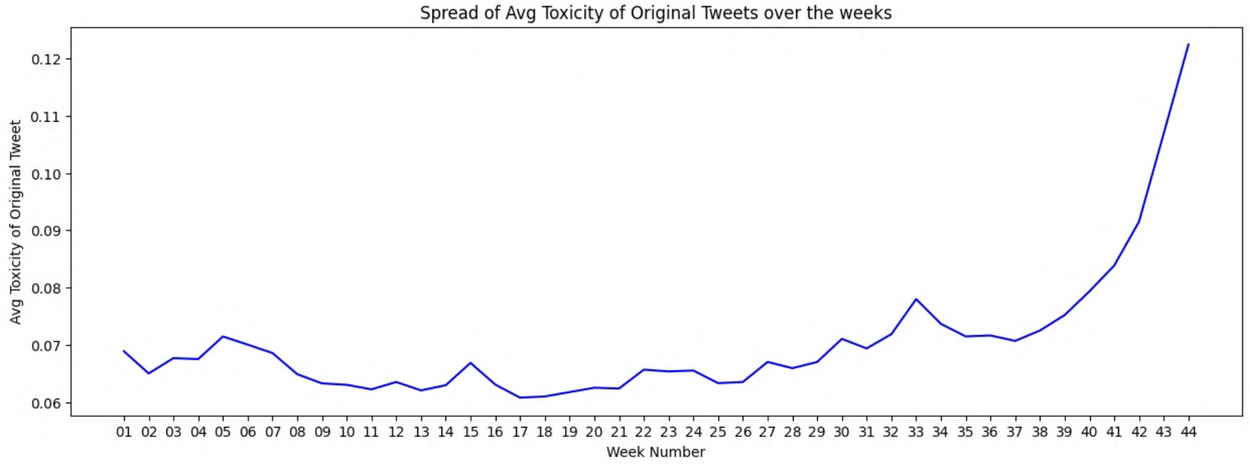


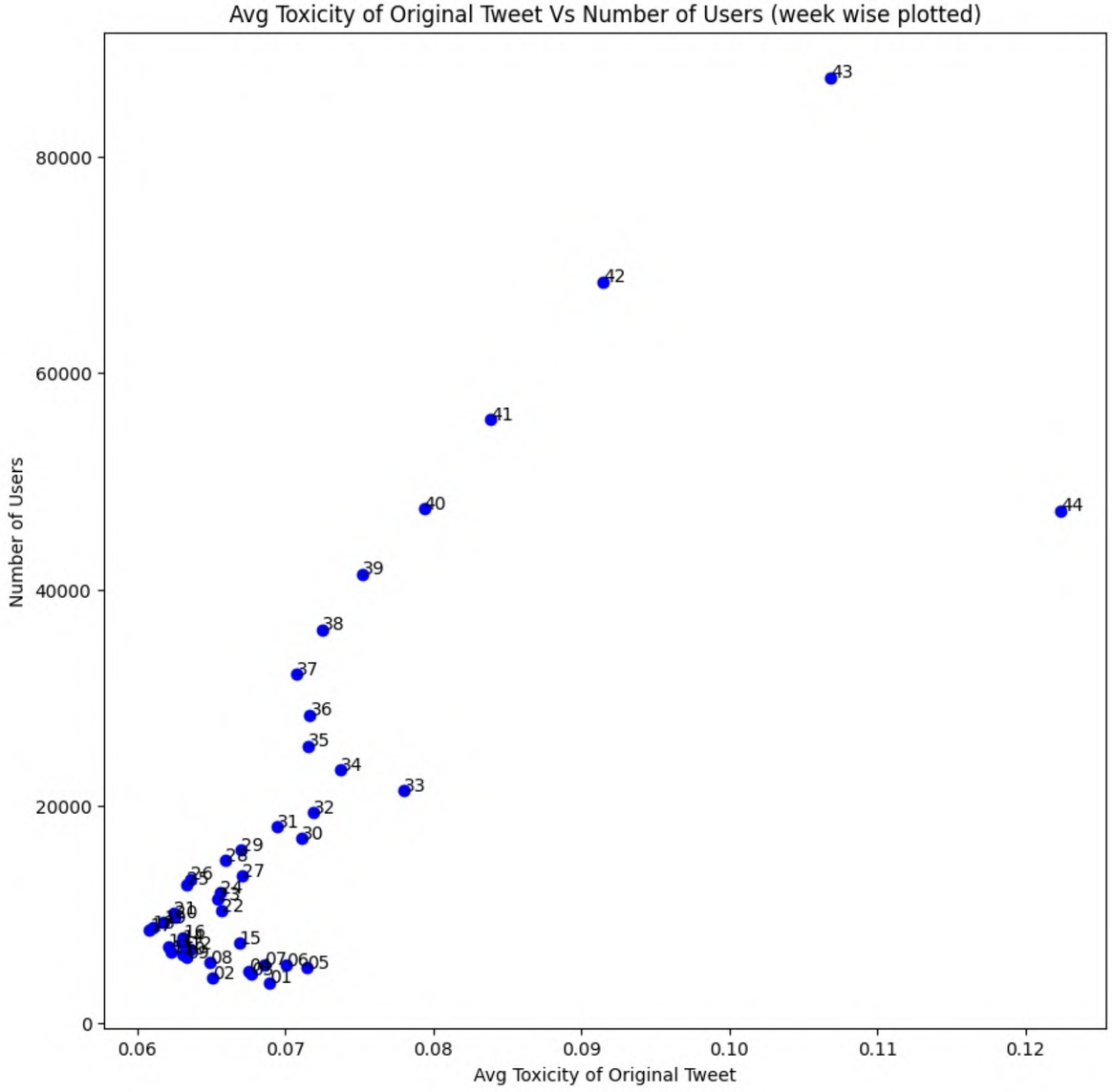


Plot 1,2 → Total Sum of Toxicity Increases weekly

What is happening to the **Average Toxicity** for Original Tweets?

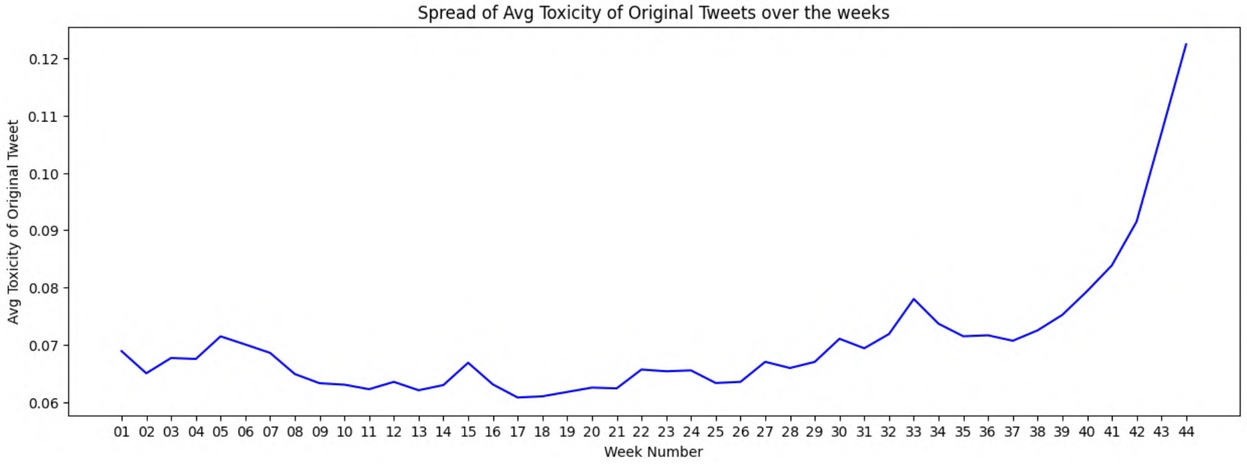
Original Tweets

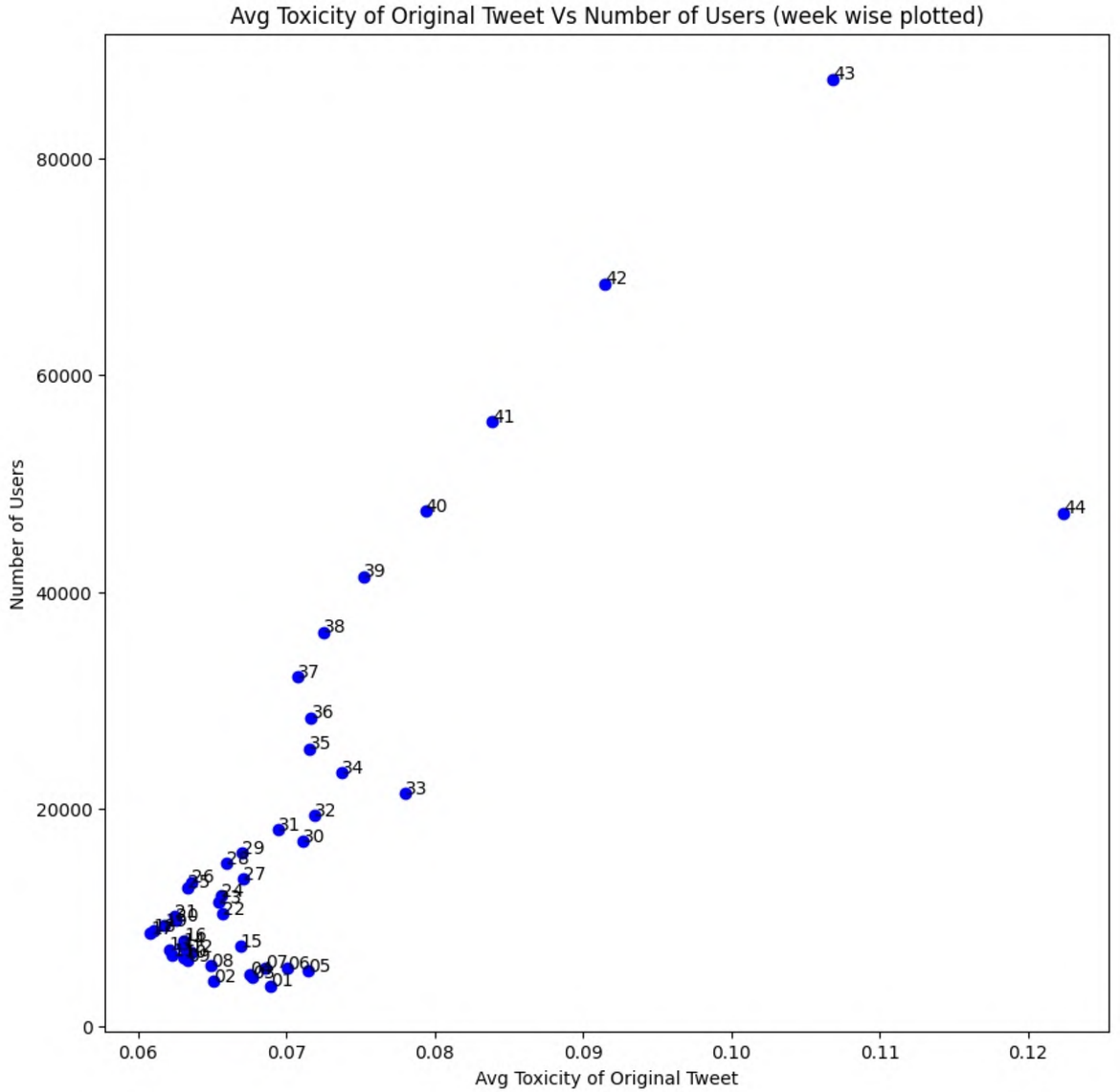




Try Pitch

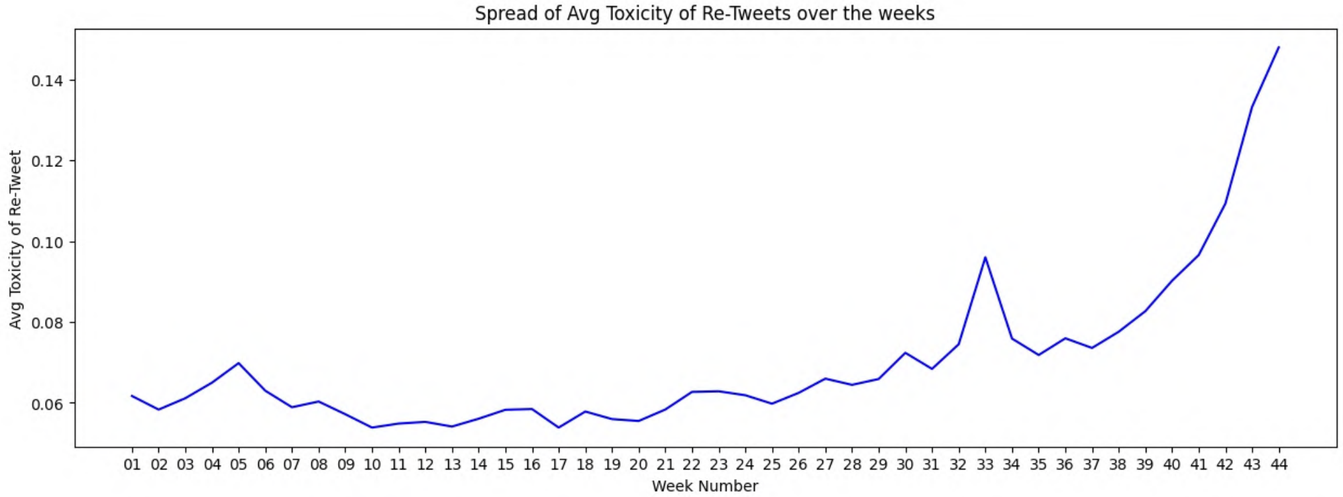
Original Tweets

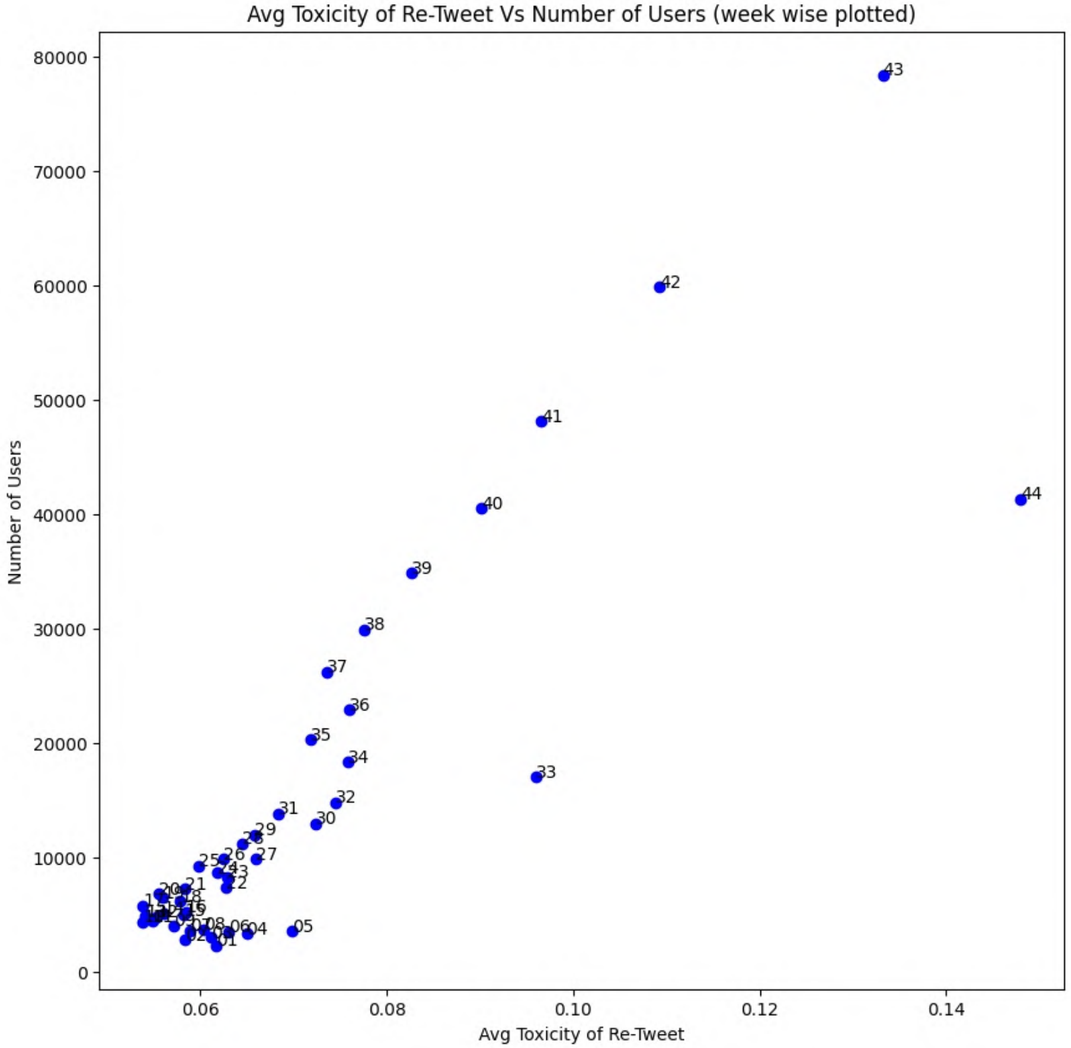




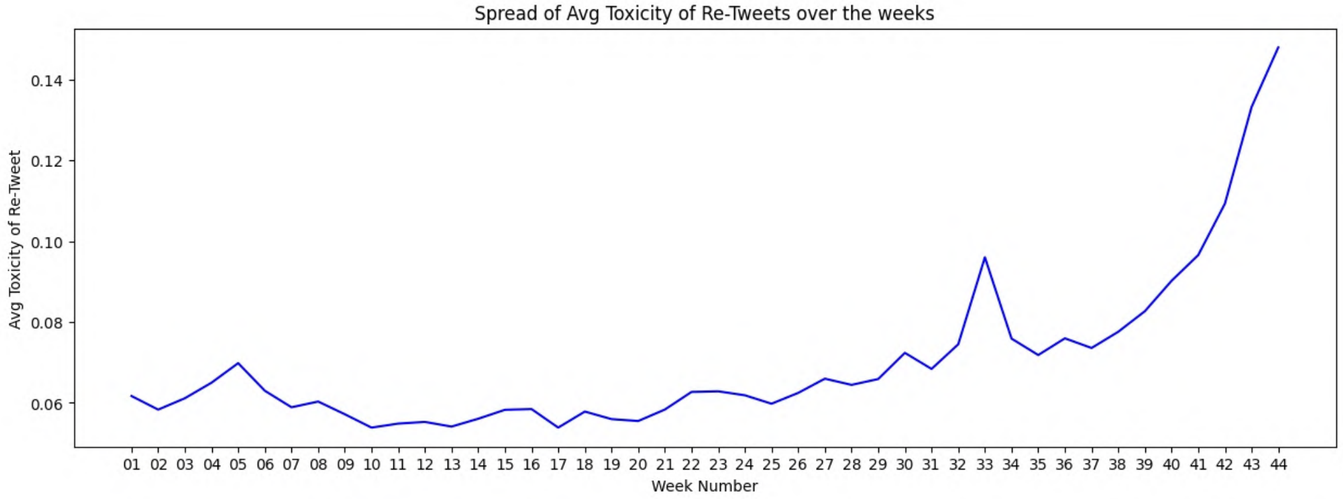
Plot 3 → Average Toxicity of Original Tweets show a steady increase

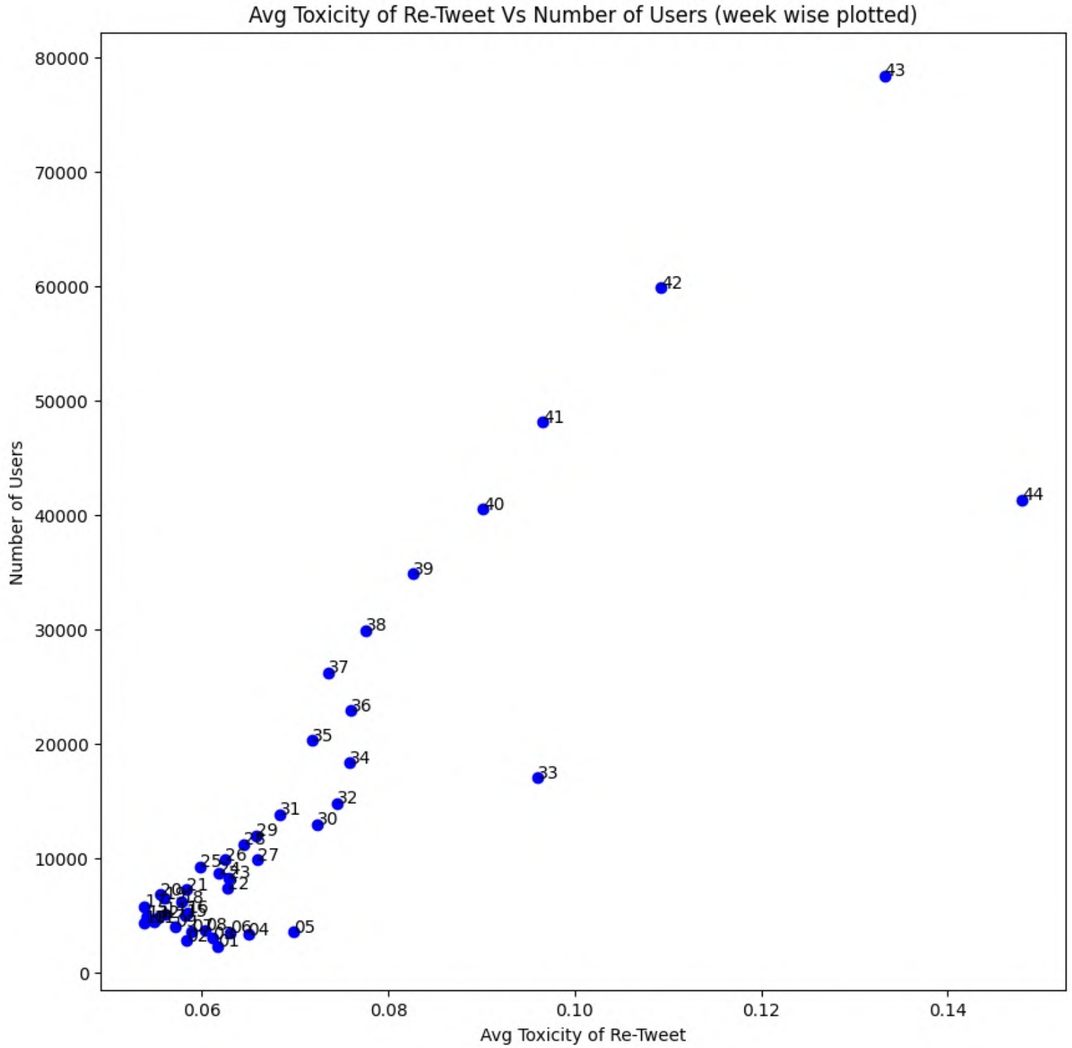
Re-Tweets





Re-Tweets

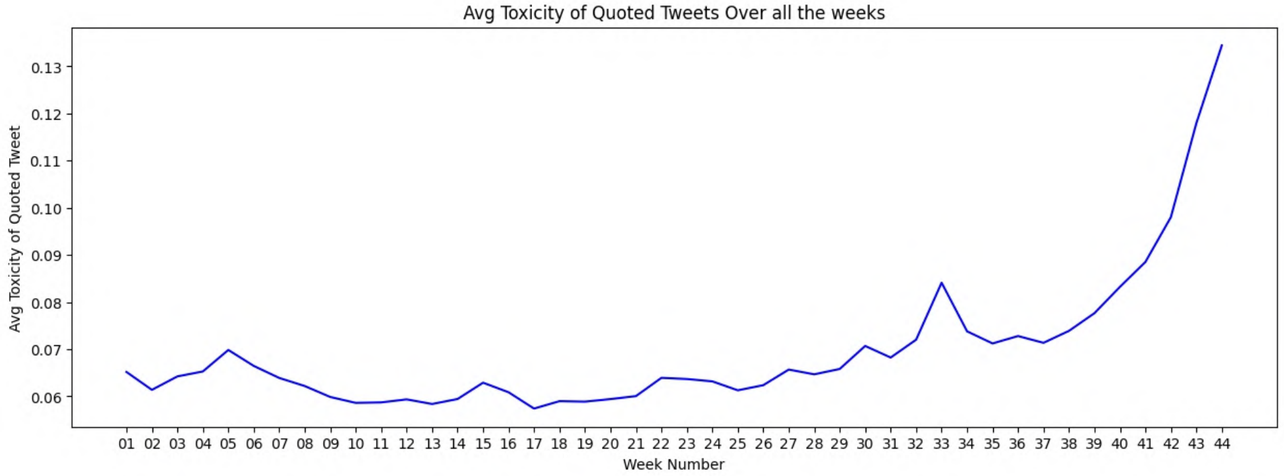


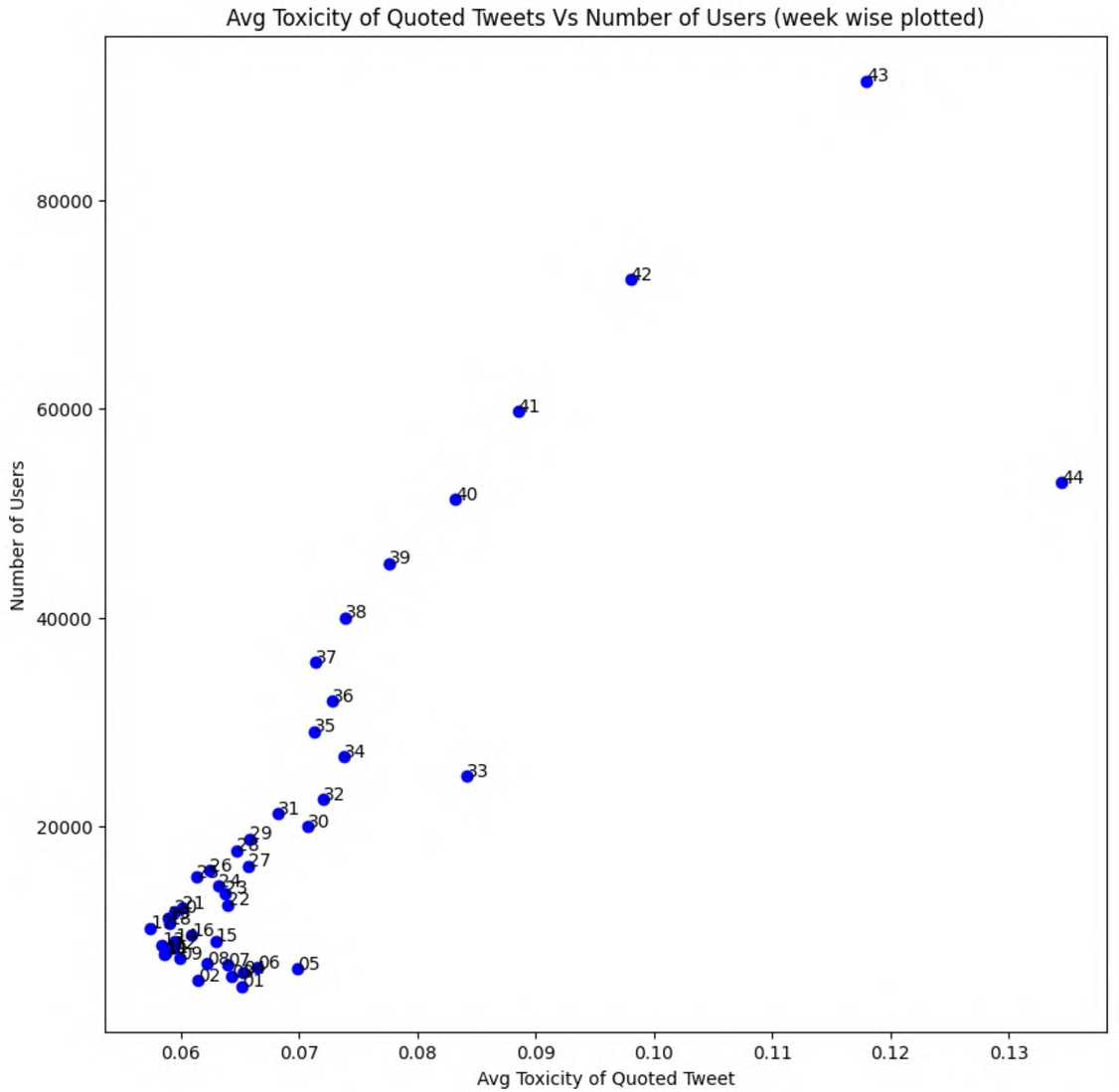


Plot 4 → Average Toxicity of Re-Tweets show a steady increase.

The average toxicity is a bit higher than the Original Tweets.

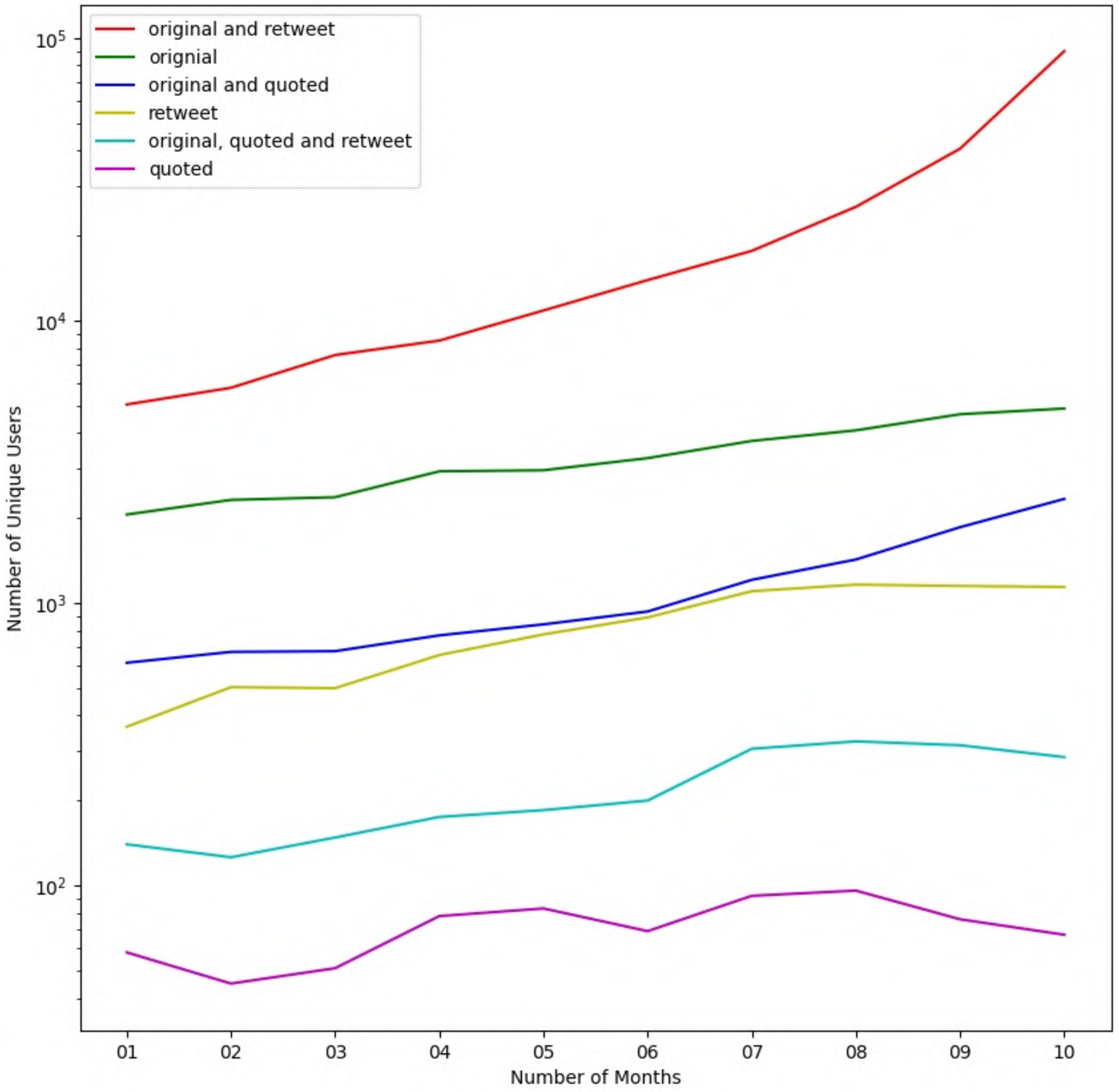
Quoted Tweets



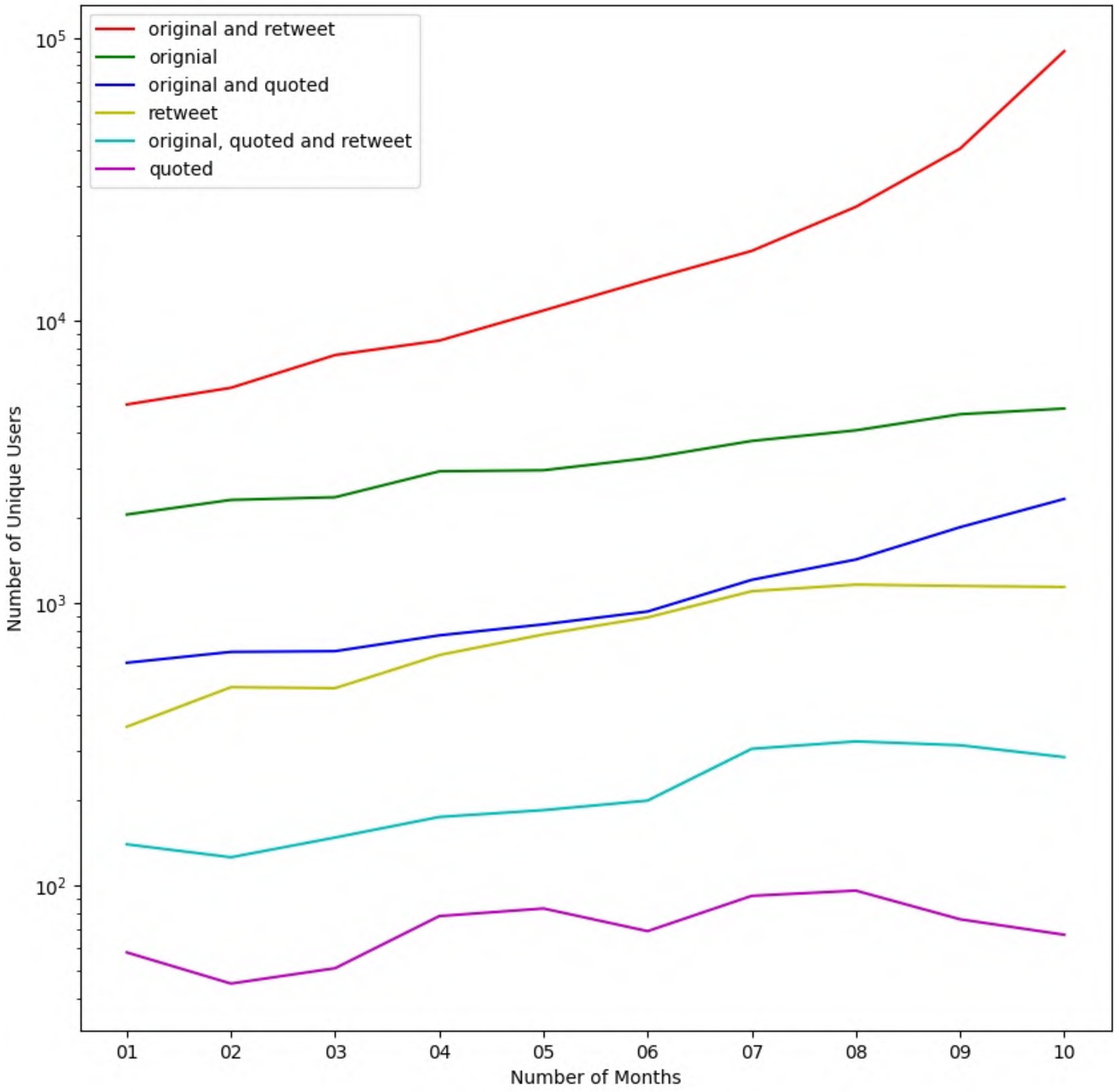


Plot 5 → Average Toxicity of Quoted Tweets show a steady increase

How are users with their tweet categories spread across the months?



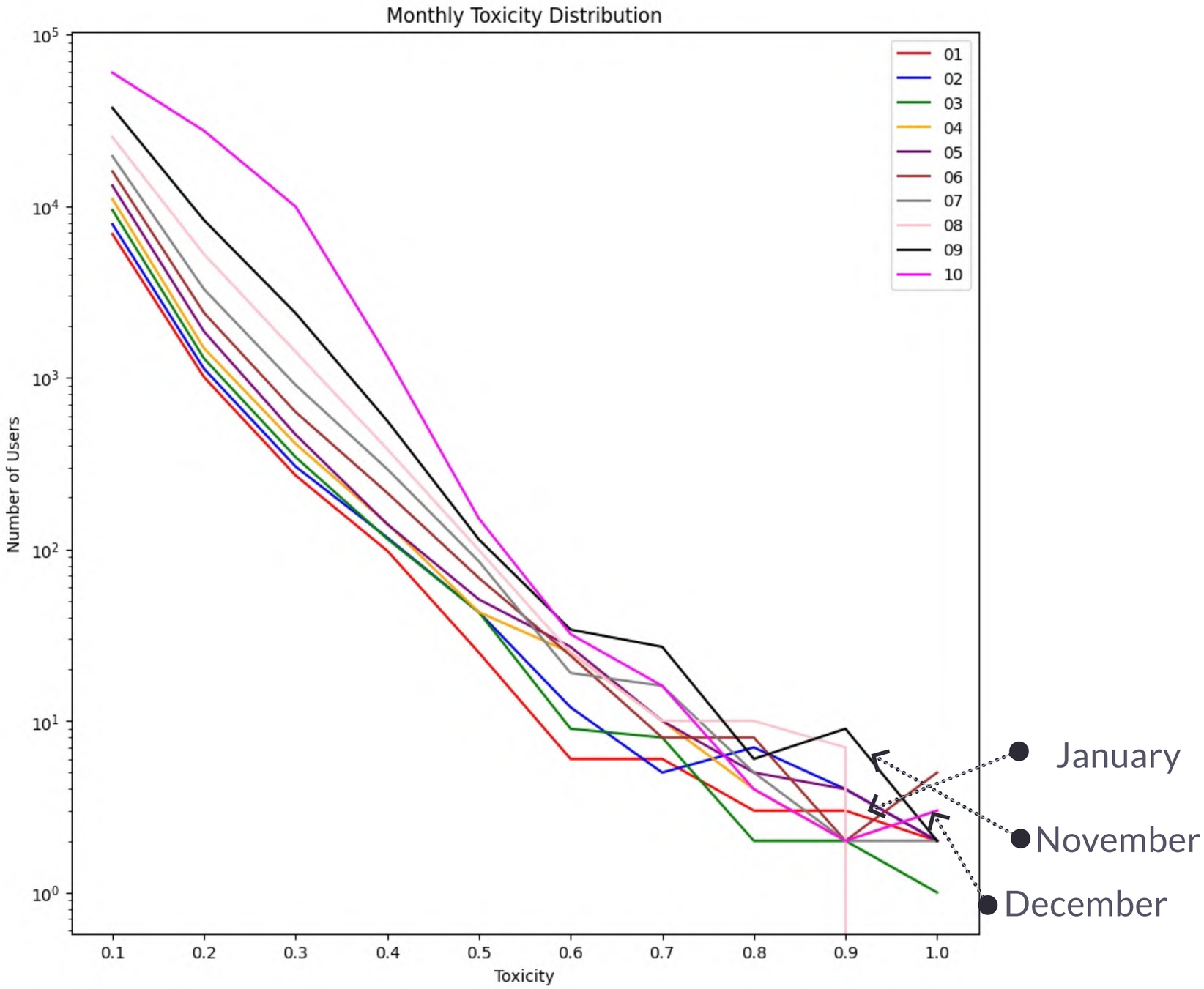
How are users with their tweet categories spread across the months?



Plot 6 →
Majority of the Users tweet original tweets and retweets.

Re-tweets spread more toxicity.

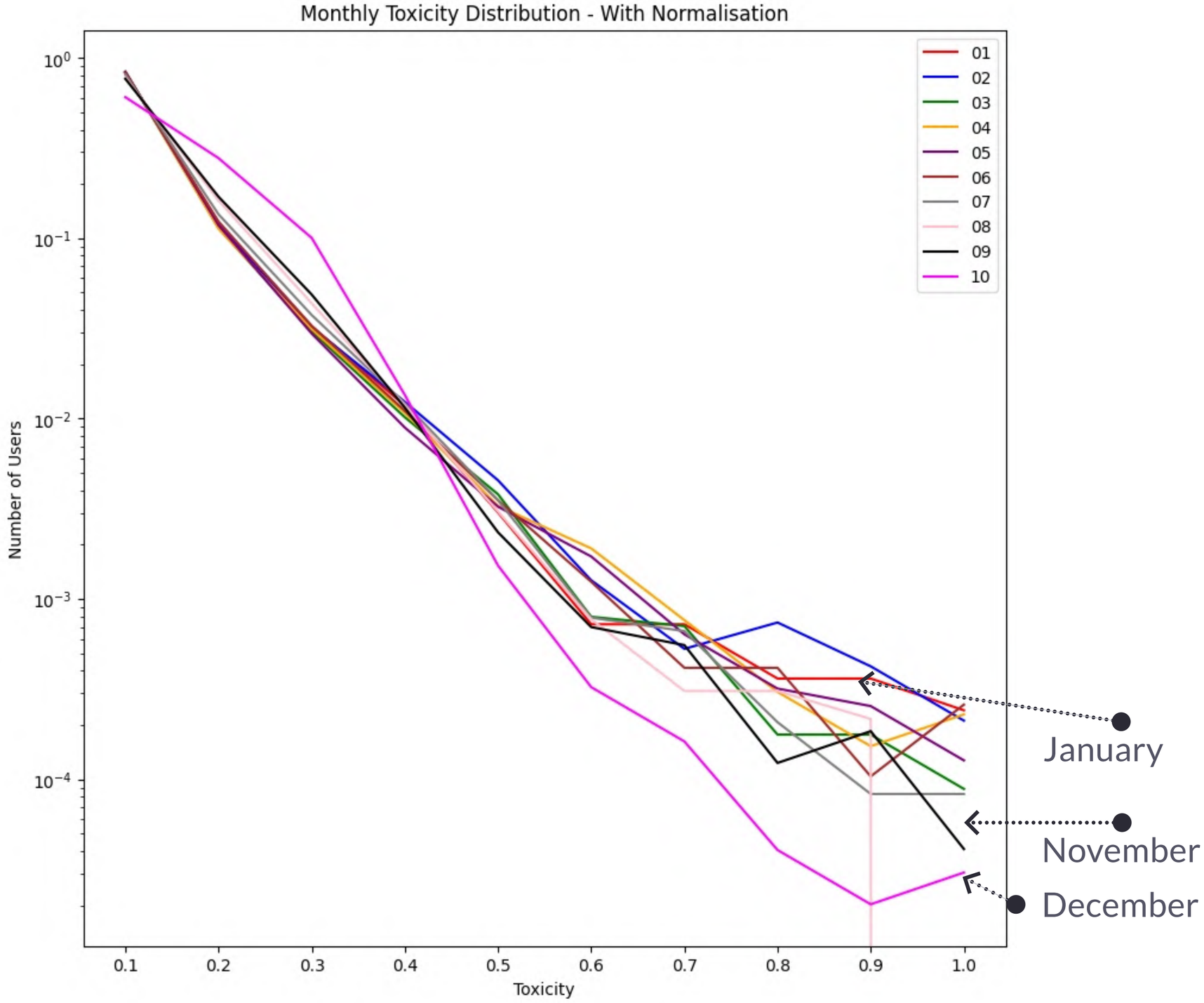
What is happening to the User Toxicity across the months?

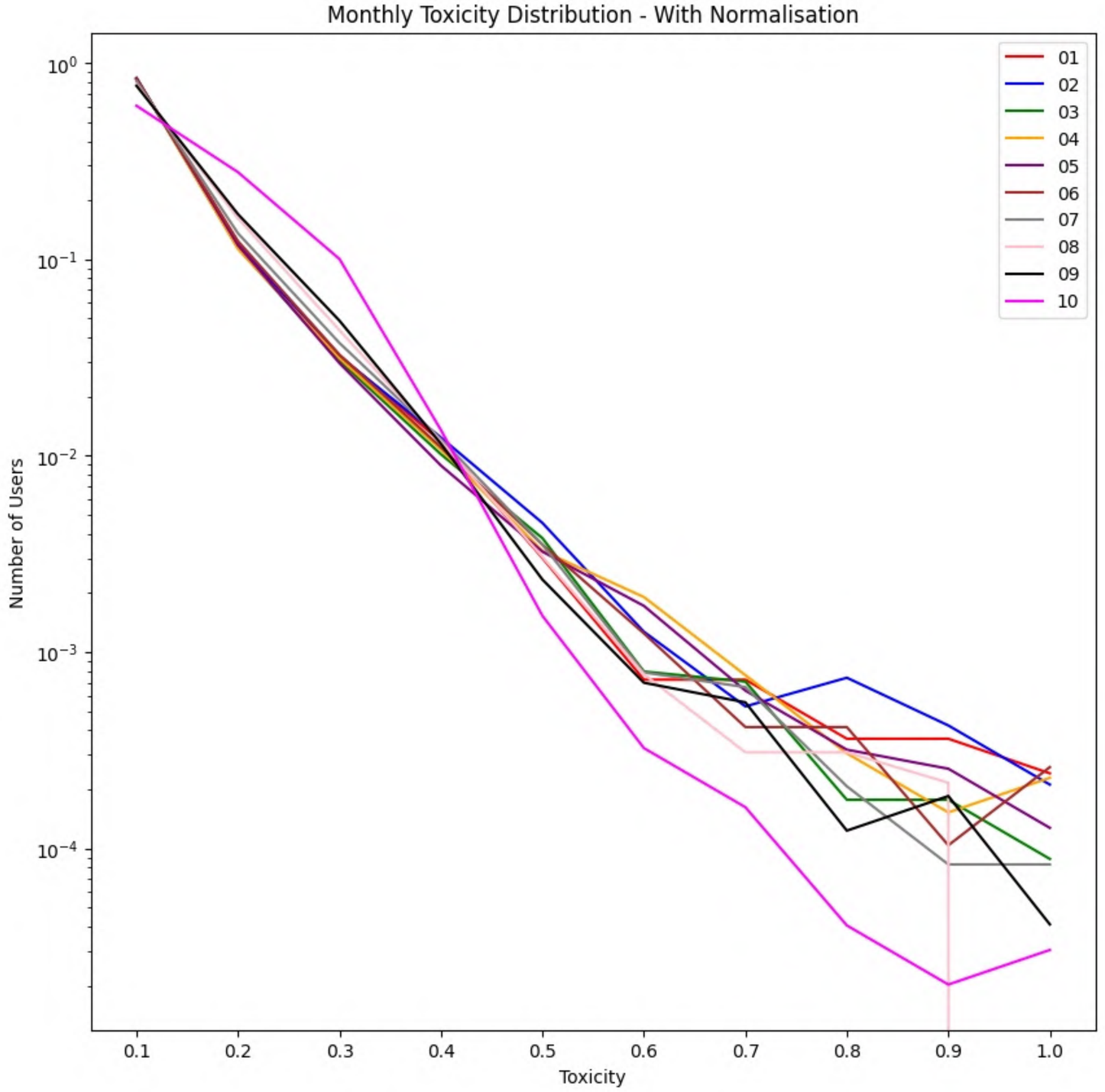


Fraction =

Number of users in that bucket

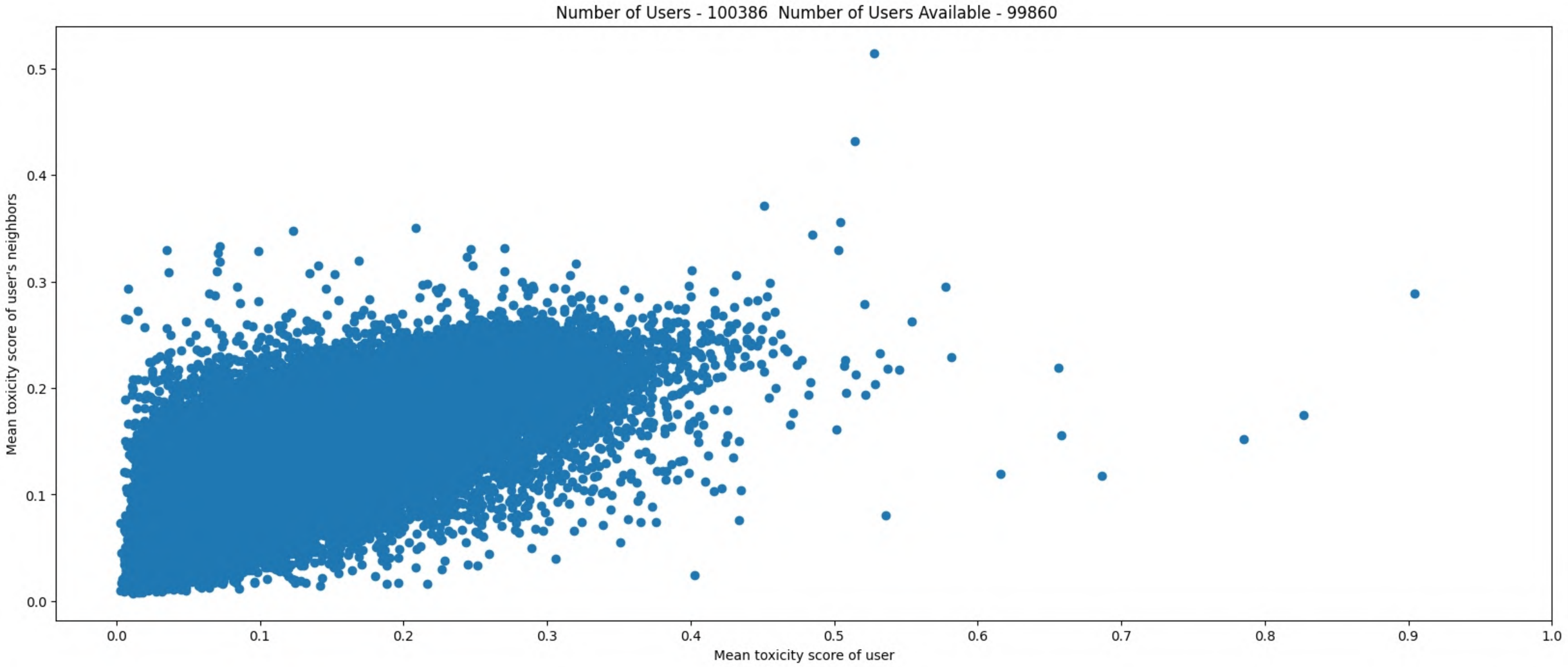
Number of Unique Users in that month



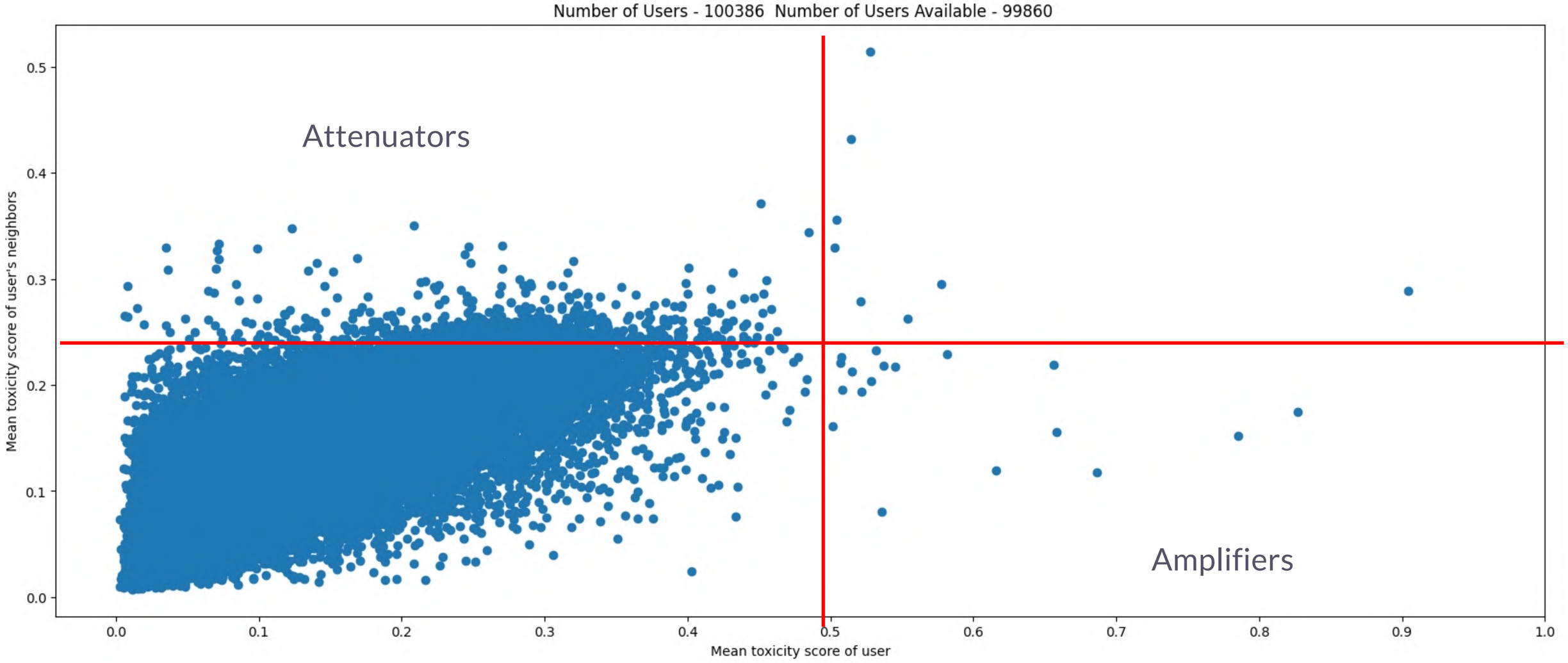


Plot 7 → All the months have users with high toxicity.

Average Toxicity - User Vs Neighbour



Average Toxicity - User Vs Neighbour



Plot 8 → Users show 3 types of behaviors:

- Amplification - Attenuation - Similarity

We start building up!!

We look at the distribution of the difference

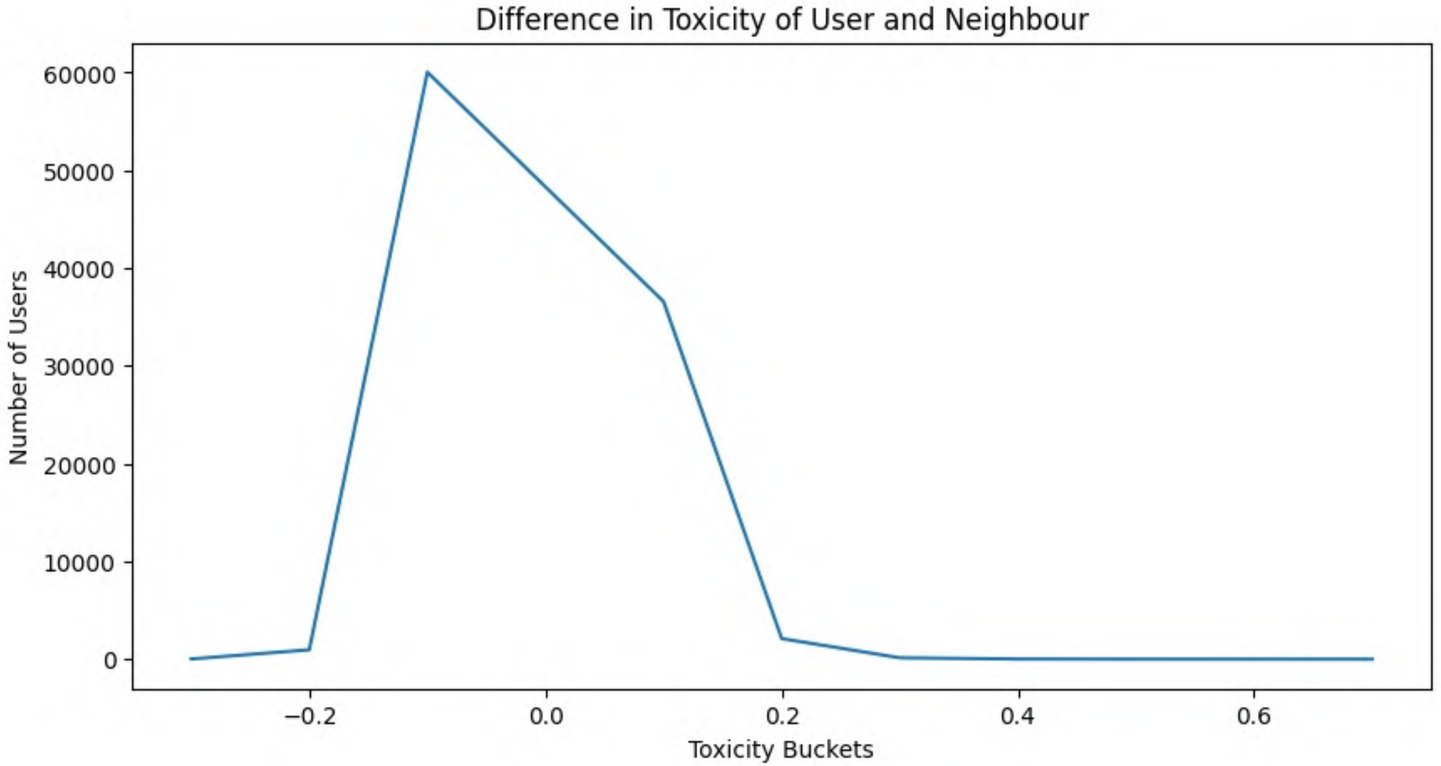
- diff_dict = user_tox - neighbor _tox

This distribution isn't normal

- Shapiro–Wilk → Negative
- Kolmogorov–Smirnov → Negative

This is where IQR comes in!!

IQR is robust measure of variability



⟶ Find Thresholds to Categorise Users

⟶ Find Shifts for each User Category

Fraction of Users

Attenuators - 2%

Amplifiers - 5.5%

CopyCats - 92.5%

Shifts

Attenuators - (-0.1038)

Amplifiers - (+0.1605)

CopyCats - (-0.0053)

What do we conclude?

1. **Counter to SPA** → Energy is not conserved

   ⟶ We saw Total Toxicity and Average toxicity per tweet increase over time

   We saw there were more users with higher toxicity values as time increased.

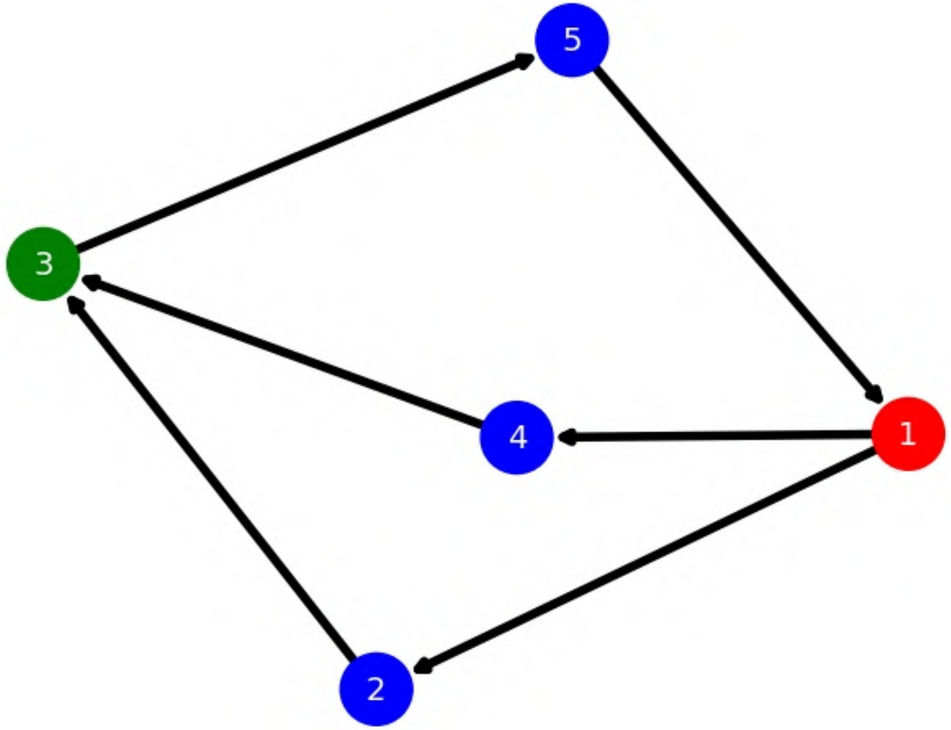
2. **Counter to SIR*** → The SIR model gives us a discrete state of the disease, but not the extent.

   ⟶ If there is a discrete state, that means there is a threshold.

   The threshold doesn't factor the amount of hate below it.

So what ahead now?

Lets start working with Toy Graphs now



We need to think of a cool name!!

There are two ways of going about it

- Sum and Average

Time	1	2	3	4	5
0	0.9: 1, 0.7: 2				
1		0.8: 1, 0.6: 2		0.8: 1, 0.6: 2	
2			0.6: 2, 0.4: 4		
3					0.5: 2, 0.3: 4
4	0.6: 2, 0.4: 4				

Testing - Random Graphs

ER Graphs → Graphs could be disconnected

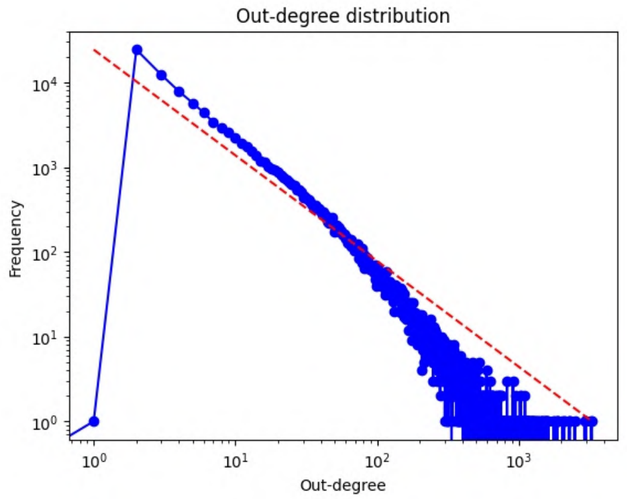
WS Graphs → Probability of edge creation. This wouldn't help us replicate our twitter graph

BA Graphs → Scale Free and Real Life resembling.

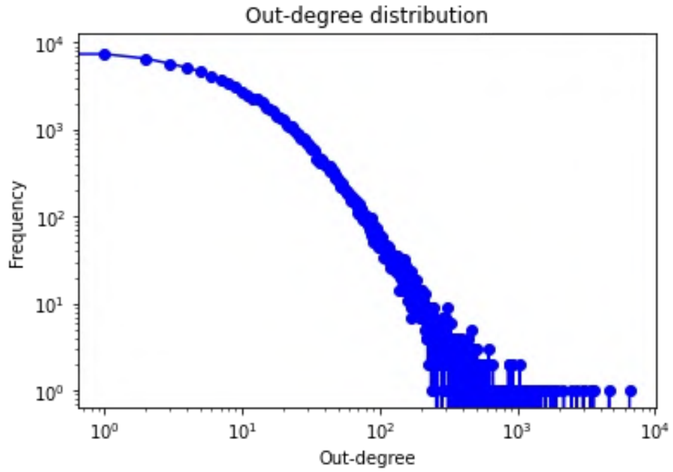
To resemble our Re-Tweet Graph

- Tweaked the BA graph creation
- Reversed the Edges
- The Outdegree of the graph - Outdegree of our retweet graph.

Re-Tweet Graph



Modified BA Graph



Results - BA Graphs

No: of Nodes	m	No: of Edges	Time (steps)	Sum of Total Toxicity	Avg Toxicity Per User
100,000	5	499,771	55	1.19E+11	1.19E+06
			59	1.00E+11	1.00E+06
			58	1.21E+11	1.21E+06
			58	1.31E+11	1.31E+06
			56	8.67E+10	8.67E+05
			57	1.04E+11	1.04E+06
			54	1.73E+11	1.73E+06
			53	1.42E+11	1.42E+06
			54	1.23E+11	1.23E+06
			55	2.01E+11	2.01E+06

No: of Nodes	m	No: of Edges	Time (steps)	Sum of Total Toxicity	Avg Toxicity Per User
50,000	5	249,796	52	1.69E+10	3.37E+05
			52	1.65E+11	3.31E+06
			55	2.53E+10	5.05E+05
			45	8.97E+10	1.79E+06
			52	1.07E+10	2.14E+05
			51	4.87E+10	9.75E+05
			53	2.08E+10	4.16E+05
			46	2.27E+10	4.54E+05
			54	2.15E+10	4.30E+05
			51	3.02E+10	6.04E+05

Total Sum of Toxicity High Indegree High Outdegree

Nodes	m	edges	time	toxicity	toxicity - attenuators	toxicity - amplifiers	toxicity - attenuators	toxicity - amplifiers
10,000	3	29,937	31	3.63*10^5	1.05*10^5	3.09*10^6	2.31*10^5	2.5*10^5

Future Work

- Create a Model that would generate random twitter re-tweet graphs.
  - This would facilitate better testing for twitter retweet graphs.

- Testing the Model on the BA graphs where each tweet has an **age (**information value) and it decays over time.

- Think of Strategies to Mitigate Hate Speech.

Thank You!!