Social Media and Network Sciences

# Twitter Retweet Network Analysis

DSM160-2024

# Abstract

X (formally Twitter), is a massively popular social network, formally categorized as a microblogging network. A network where followers follow influencers, and opinions (Tweets) by influencers are broadcasted with a specific Hash-tag/s (#ChatGPT for example), these tags build a social circle encompassing the communications in the form of Tweets, Retweets, Likes, etc. We will be analyzing a collection of Retweet data gathered from Twitter’s API access for 6 tags, #chatgpt, #bitcoin, #BoredApeYatchClub, #Eth, and #uranium. The report will explore each network structure using Network science techniques and a comparative analysis will highlight the differences between them. For this purpose, a Retweet network will be constructed with each Original tweeter and re-tweeter assigned as Nodes and the link between them will specify the flow of information.

# Keywords

Network Science, Twitter (now X), Retweet analysis, Crypto tweet network

# Data

The data being analyzed is a static collection of these networks, hence the analysis will also focus on the state of things at the specific time and duration. There is an incentive to have a larger time frame for analysis to extract a deeper and more accurate understanding of the network, we define the data in terms of its 4 Vs.

### Volume

Each retweet data contains in addition to the tweet text, a large amount of metadata, and in the case of it being a retweet, it also contains the metadata for the original user. From the provided extract of data, an average Tweet is 6KB in size, and one-day data amounts to 25MB.

### Velocity

Tweet rate for the networks is as below, #Etherium and #ChatGPT are similar and the fastest in terms of new data generation. With 6KB of new data generated every ~30 seconds.

|  |  |  |  |
| --- | --- | --- | --- |
| **Network** | **Rate of Tweet** **(H:MM:SS)** | **Measurement Period** | **Date** |
| #ChatGPT | 0:00:29.861600 | 1 day, 17:28:28 | 2023-02-06 |
| #Bitcoin | 0:00:16.338431 | 22:51:04 | 2023-02-07 |
| #boredApeYachtClub | 0:02:42.061663 | 9 days, 12:30:25 | 2023-02-27 |
| #energy | 0:00:31.479800 | 1 day, 19:43:19 | 2023-02-06 |
| #eth | 0:00:28.380324 | 1 day, 15:25:30 | 2023-02-07 |
| #uranium | 0:01:43.696600 | 6 days, 0:01:23 | 2023-02-06 |

### Variety

The retweet data is contained in a JSON file, for this analysis, we are limiting ourselves to the text-based metadata of each retweet. However, the file also contains links to videos, images, and web links.

### Veracity

In a social media setup, there is a good chance of data misrepresentation or irrelevancy to the topic, other factors such as bots and spam also contribute to the challenges faced during analysis. Which is why it is of benefit to identify these players.

### Legislation and Bot/Spam consideration

## General Data Protection Regulation (GDPR) imposes strict rules on personal data handling (https://gdpr-info.eu­­), it operates in the European Union. It defines a legislative ground by defining the principles, rights, and restrictions such as Lawful processing, Consent, code of conduct, transfer of data, and penalties.

## Other sources to keep in mind are X’s term of service and developer agreement, where the sharing restriction of API accessed data and its utilization for surveillance.

# Processing the Data

The JSON files are loaded into Python using the JSON library, for each re-tweet in the network, the information on the original tweeter and the re-tweeter are captured along with their follower counts. These *users are designated as nodes*, each re-tweeter is linked to its original tweeter via the edges. The larger the interaction between two such nodes, the greater will be the weight of the edge. The nodes themselves are weighted by the number of their followers and also attributed to the number of tweets/re-tweets sent by them. A network library is used for the creation of the networks. Pyvis library is used for creating the visualization for each of the networks.

The resulting networks are directed and weighted as per the above-mentioned attributes. Within each network, there are many weakly connected components, visualized as a star-like structure, where each compartment acts as its re-tweet network centered around an original tweeter (influencer). *The directivity of links represents the flow of information from the tweeter to the re-tweeter*.

# Networks

After creating the networks, some key observations distinguish the networks. Here we will describe the network from the visual **Graph perspective** and from its **Node characteristics** (such as Top 10 degrees, Tweets, Retweets, etc.) and finally from the perspective of the **Network’s KPIs** (such as assortativity, degree, density, clustering, etc.), we will also briefly analyze the tweets to get a sense of the hot topics being discussed.

## Graph Representation

The full network views in Figure 1 offer some insights into each of them, there are similarities in terms of the retweet networks among similarly themed hashtags.

### #Bitcoin: A cryptocurrency hashtag

Around 10 core nodes can be identified from where the most original tweet activity is taking place and retweeters are mostly connected to either one of them. Node size represents the tweet activity by both the tweeter and the retweeter, no standout node in terms of high tweet count is observed in the core network. #Bitcoin is the most similar to #ChatGPT in its network architecture, there are many core nodes and the retweet activity is also distributed among a good amount of retweeters. Edge weights that represent the retweet activity among two nodes (tweeter->retweeter) are generally consistent, which suggests no specific relationship is strong (at least within the data collection period). The k-core(=2) (Fig 2) represents a difference in its architecture vs #ChatGPT, here most retweet activity is retained as the retweeters are connected to multiple nodes.

### #ChatGPT: One of the main players in Artificial intelligence

Similar to #Bitcoin, Core is dominated by many original tweeter nodes (main ones; Ishansharma7390, johnvianny) and similarly many retweeters are connected to either one of these nodes. Two key retweeter nodes are identifiable which connects to multiple nodes (uCloudifyAI, SlothGPT). Like #Bitcoing many high-degree weakly connected commponents are observed.

### #BoredApeYatch :

A very unique network, which is dominated by two Tweet nodes (AshtonNFTs, oesnetwork) many retweeters are connected to either one of these nodes, and there is a group of connecting retweeters that connect to both nodes.

### #energy: Represents the topic of energy

Lead by only one tweet node (MikeHudema) and many small nodes with their accompanying retweeters. A high amount of low-degree weakly connected components are observed. This is represented in the k-core(=2) graph (Fig 2).

### #eth

### #uranium: Represents the mineral Uranium

A very densely connected network, where the nodes are interconnected via retweeters which are not selective in retweeting a particular node. A similar observation can be made from the k-core(=2) graph which retains most of the core network suggesting a high amount of interconnectivity.

### Full Network

|  |  |  |
| --- | --- | --- |
| **#ChatGPT** A group of fireworks in a circle  Description automatically generated | **#Bitcoin** A group of red and blue dots  Description automatically generated | **#BoredApeYatch** A red and blue fireworks  Description automatically generated |
| **#energy** **A group of red and blue dots  Description automatically generated** | **#ethA network of red and blue dots  Description automatically generated** | **#uranium** **A red and blue network  Description automatically generated** |

Figure 1 Re-tweet Network Graphs of all Hashtags

### K-Core = 2 Network

|  |  |  |
| --- | --- | --- |
| **#ChatGPT** | #**Bitcoin** | **#BoredApeYatch** |
| #energy  A diagram of a network  Description automatically generated | #**eth** | **#uranium** |

Figure 2 K-Core(2) Re-tweet Network Graph of all Hashtags

## Node Statistics and Insights

This section will focus on the Node level statistics, looking at the Top nodes in terms of different metrics.

### Top Nodes: Degree

The ranking below shows high-degree nodes of each network, along with the distribution of Out and In degrees, the majority of high-degree nodes have higher out degrees which represents a higher number of retweeters for that particular node. This does not always mean a higher tweet count. The biggest node among the Networks is from #BoredApeYatch: AshtonNFTs with 1,767 out-degrees.

Figure 5 shows the degree distribution per degree range, some clear differences are observed between them, similar analysis is done when calculating the CCDF of degree; refer to the discussion there  
  

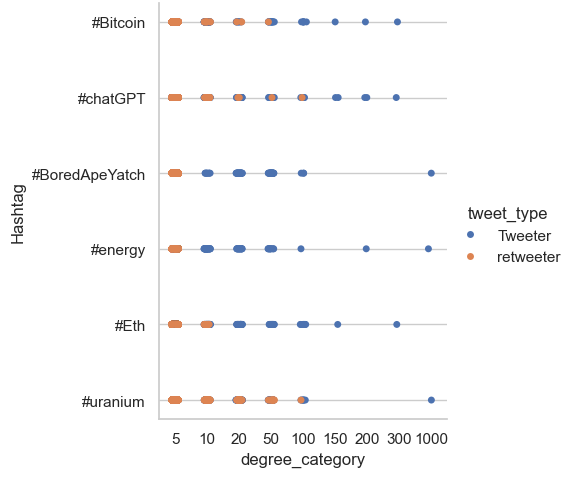


Figure 3 Degree Distribution

### Top Nodes: Total Tweets

The nodes with the highest amount of unique tweets/retweets are shown in the figure below. The stack bar chart is used to represent the retweet vs tweet ratio, as certain nodes are higher in original tweets such as ibis\_sacred (#Bitcoin) has the highest original tweet count in all networks (~50), whereas AndUranium(#uranium) has the highest retweet count (~175).

A graph with multiple colored bars

Description automatically generated with medium confidence

Figure 4 Nodes - Tweets vs Retweets

### Tweets vs Retweets

Looking at the network distribution of Tweets to retweet quantity distribution, some interesting observations can be made, all networks except #BoredApeYatch have an average of 20-25% original tweets, remaining are retweets. #BoredApeYatch has only 3% original tweets, the remaining are retweets.

A group of pie charts

Description automatically generated

Figure 5 Tweets vs Retweets

### Top Nodes : In Degree

A look into the top retweeters in terms of connection to multiple original tweeter, could potentially identify a SPAM or bot actor, for example, uCloudifyAI(#chatGPT) and SlothGPT(#chatGPT) have the highest retweets from multiple sources than others (80).

**Top Nodes – In Degree**A graph of a number of objects

Description automatically generated with medium confidence

Figure 6 In Degree Nodes Comparision

### Top Nodes: Virality

Virality is calculated as the ratio of Out-degree to No. of followers, this would suggest that an original tweeter with low followers had a high number of retweets, possibly due to the value of his tweet (not because of his follower count), Baronbeen002(#chatGPT) scores the highest.

**Top Nodes – Virality**A screenshot of a graph

Description automatically generated

Figure 7 Virality Nodes Comparition

## Network Statistics

All the important statistics are calculated and compared among the networks for a relative comparison.

### Nodes & Edges

#BoredApeYatch has the highest number of total nodes and edges (4000) with #uranium having the lowest count. Relatively similar count of nodes in the rest.

### Density

#uranium has by far the highest density, suggesting a higher connected network, it can also be compared by the higher edges-to-node ratio of #uranium.

### Average Degree

#uranium shows the highest average degree, it is also another indicator of a more connected network

### Assortivity

Generally, lower assortivity since retweeters connect to high-degree tweeters and rarely connect to other retweeters. Only #boredApeYatch shows a positive assortivity, suggesting a connection between retweeters.

### Weakly Connected components

#boredApeYatch has the lowest components, that is another sign of a more connected network, especially when the network has the highest count of nodes, we can see from the K-core graph in figure-2 that most of the nodes are connected within the core component.

A group of colorful bars

Description automatically generated with medium confidence

Figure 8 Network Statistics

## Complementary Cumulative Degree distribution

The degree distribution of 6 networks is plotted here using a log scale so that all degree numbers are represented without outliers affecting the visibility of the graph.

A clear difference is #uranium with its high CCDF consistent at medium degree numbers, suggesting a well-connected network and relatively more interaction between the nodes than the other networks.

#BoredApeYatch has a long tail at the higher degree values with low CCDF, suggesting few nodes with high degrees can be the main influence in the network. Lower relative (compared to others) probability at a low degree suggests less amount of weakly connected components.

#energy has a relatively higher CCDF at a high degree with and higher CCDF at a low degree as well.

#eth has a higher density of medium degree nodes, and very low high degree frequency, which could suggest a low number of influencer count, but still a more robust network where nodes are interconnected more often than for example.

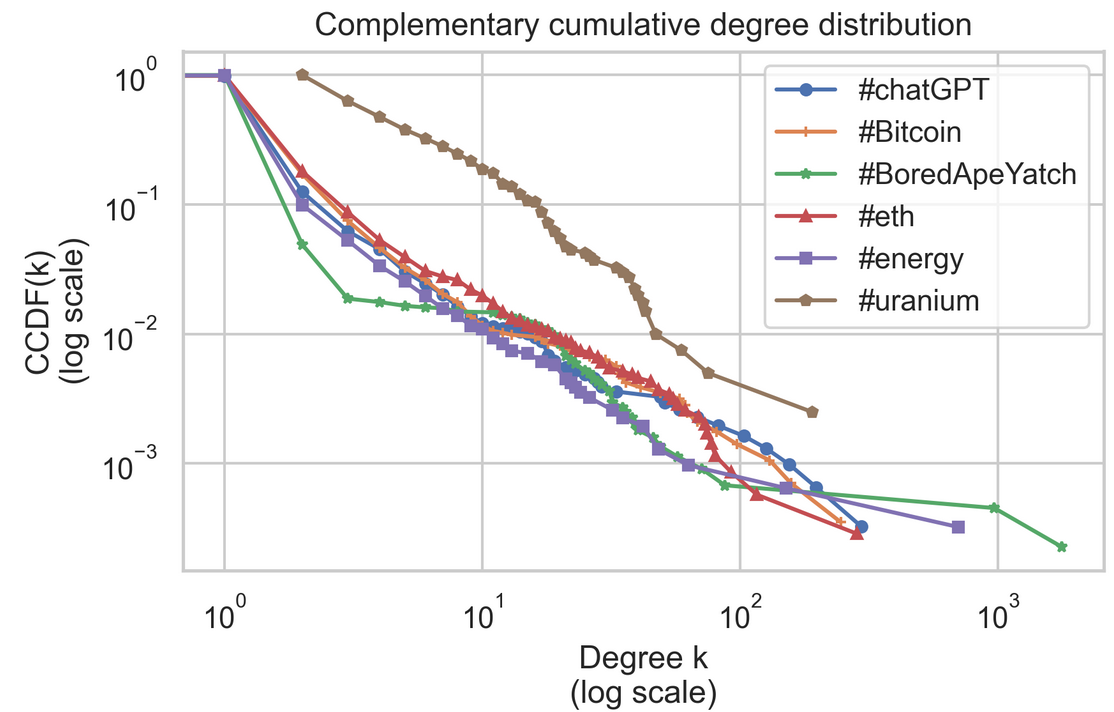


Figure 9 CCDF(k)

## Tweet Analysis

Below is an attempt to provide a sense of the content and hot topics being discussed for each Network, by creating a wordcloud for all the tweets in each network.

### #chatGPT

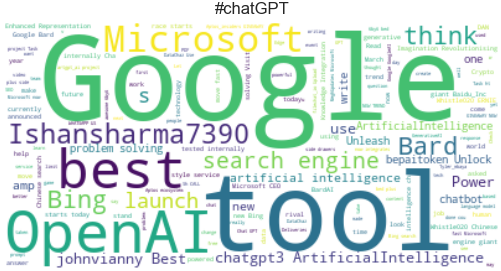


Figure 10 chatGPT Tweets wordcloud

### #Bitcoin

A close up of words

Description automatically generated

Figure 11 Bitcoin Tweets wordcloud

### #boredApeYatch

A close up of words

Description automatically generated

Figure 12 boredApeYatch Tweets wordcloud

### #energy

A close-up of words

Description automatically generated

Figure 13 energy Tweets wordcloud

### #eth

A close up of words

Description automatically generated

Figure 14 eth Tweets wordcloud

### #uranium

A close up of words

Description automatically generated

Figure 15 uranium Tweets wordcloud