



# INFLUENCE NETWORK IN TWITTER

Presented by - Deekshant Mamodia (MT19119)  
Jain Piyush Pradeep (MT19122)

# Problem Statement



How to use tweet, retweet and user data to identify and characterize the user is influencer or not?

# Motivation



Influencer is someone who has a capability to spread his/her ideology among large group of people generally for own personal benefits.

Opinion leaders, actors, innovators, etc can be considered as a influencers.

## **Why finding influencer is important -**

Useful for tasks like viral marketing, product promotion, behavior adoption and analyzing epidemic spreading.

# Influence network



Generally in social media network some user tries to influence a large group of user. Influence network is used to measure which set of user are more influential in network.

Influence network can be generalized or can be related to certain topics such as CAA, Article 370 in twitter network.



Mir Suhail @mirsuhail · Feb 26

#india #caa #muslimGenocide



247 6.9K 13.8K

Positive influencer with large number of followers.

Negative influencer despite having less number of followers.



BJP Karnataka @BJP4Karnataka · Feb 18

Dear Opponents of #CAA,

Protest as much as you want, peddle as many lies you want and create as much disturbance you want.

CAA WILL NOT BE WITHDRAWN !!!

No Force can prevent PM @narendramodi Govt in fulfilling the wish of Mahatma Gandhi & Dr B R Ambedkar.

#IndiaSupportsCAA

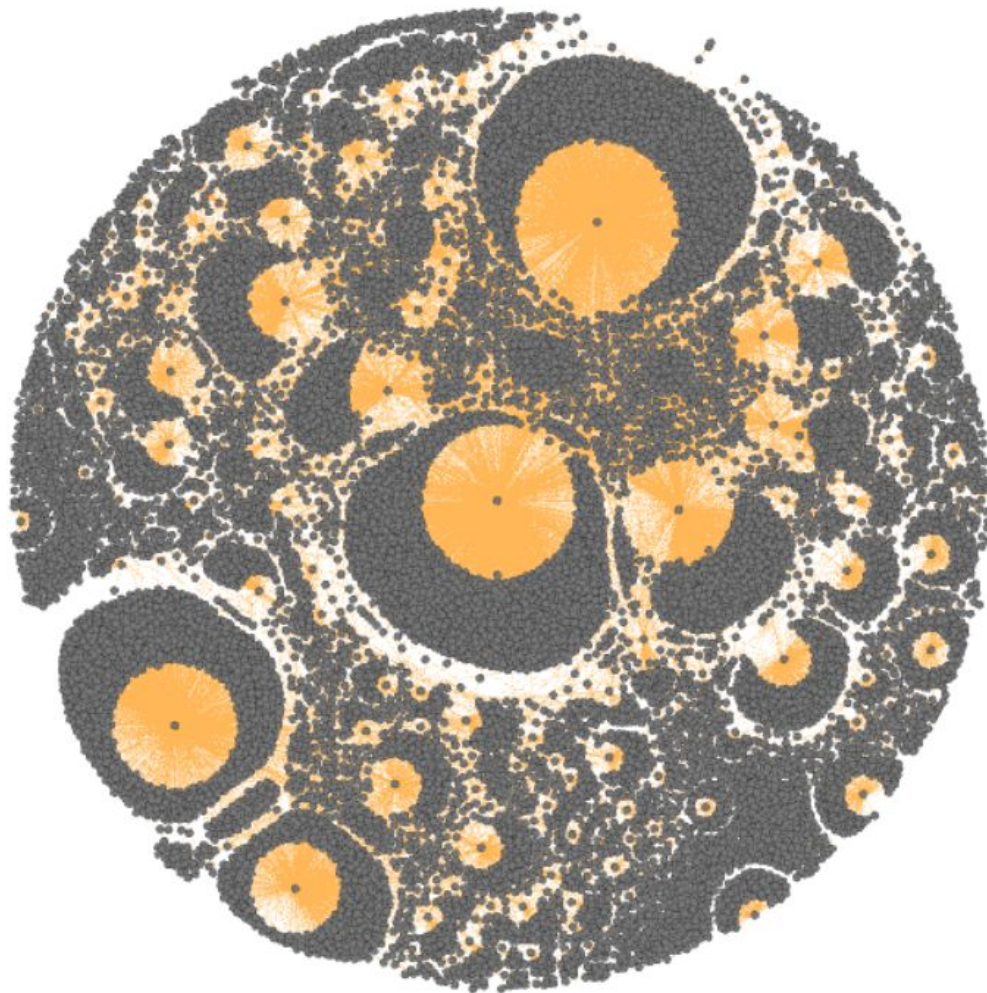
166 1.7K 6.5K

# Retweet Network Properties



Features	Values
Nodes	86194
Edges	205814
Average Degree	4.7590
Average in/out Degree	2.3795
Average Clustering Coefficient	0.01045
Average Neighbour Degree	1.01747

Scaled down version  
of retweet network  
visualization in gephi



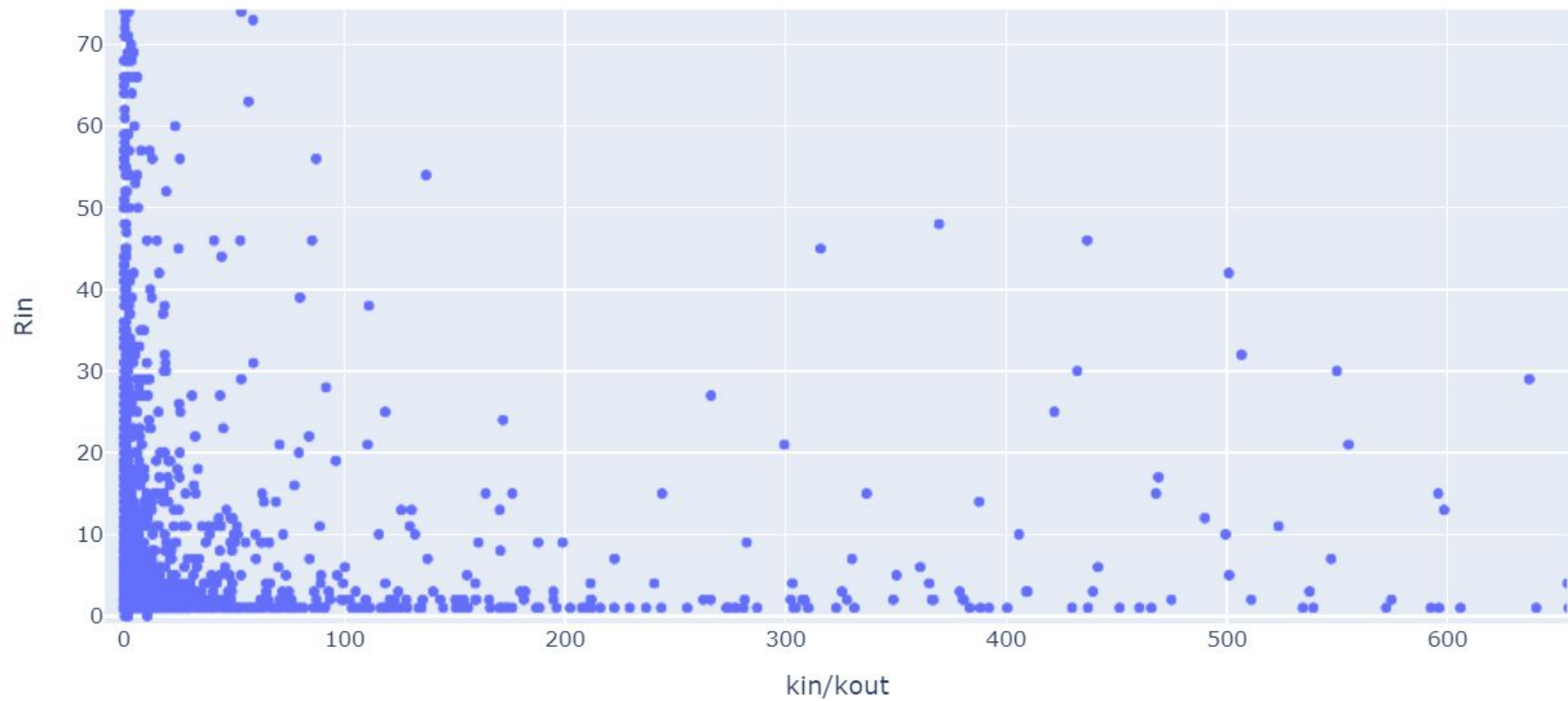
# Baseline Model 1 <sup>[1]</sup>



- Dataset of baseline based on twitter -
  - Follower network - 9782 nodes, 221498 edges
  - Retweets network - 7747 nodes, 53454 edges
- Baseline paper uses two approaches
  - Information spreading mechanism
    - popularity( $K_{in}/K_{out}$ ) vs retweet count.
  - User's Participation
    - User enthusiasm
    - User engine
- Above measures also calculated 6 large communities and 26 smaller communities.



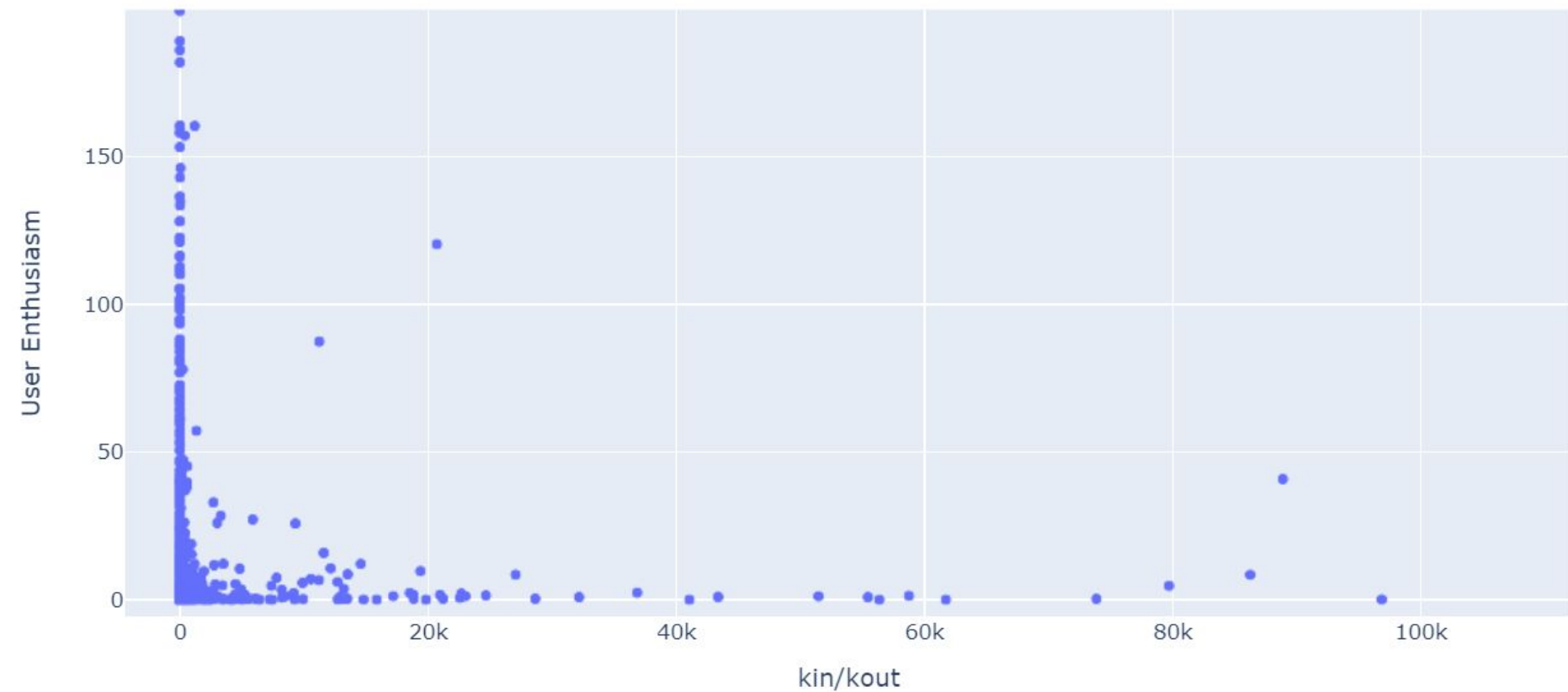
Rin transmission vs  $K_{in}/K_{out}$





Top 5 Kin/Kout	Top 5 Retweet Count
Kamal R khan	Voice_For_India
Paranita subhash	Sadhguru JV
NDTV	Piyush Goyal
Sadhguru JV	SinghShewta04
ANI	Hasanminhaj

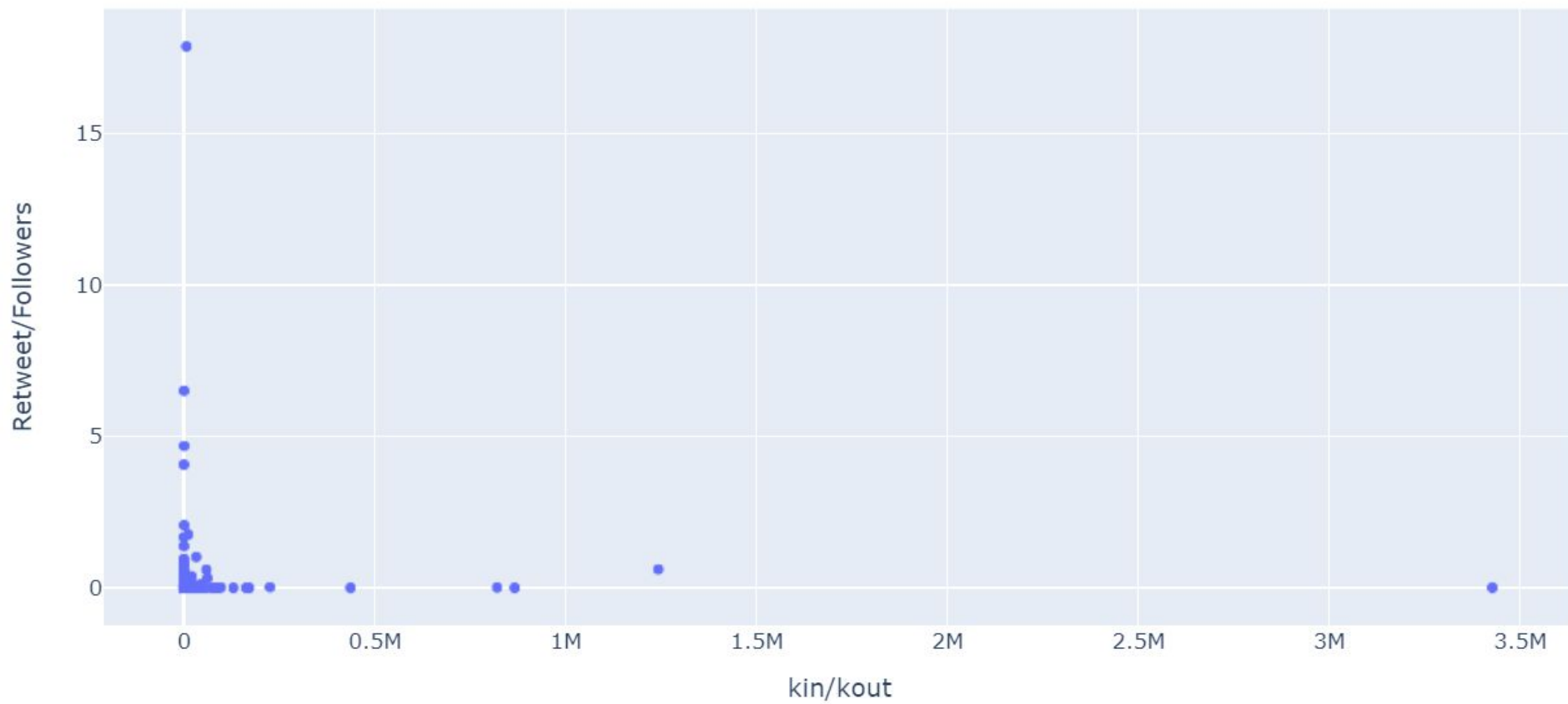
## User Enthusiasm vs Kin/Kout





Top 5 Kin/Kout	Top 5 User Enthusiasm
Kamal R khan	Pradeep76482189
Paranita subhash	aumchem
NDTV	Shivraj513
Sadhguru JV	ImamNazre3
ANI	Ayush665

## User Engine vs kin/kout





Top 5 Kin/Kout	Top 5 User Engine
Kamal R khan	only_hindutva
Paranita subhash	mirsu hail
NDTV	hasanminhaj
Sadhguru JV	Voice_For_India
ANI	officeofbagga

# Baseline Model 2 <sup>[2]</sup>



- Dataset of baseline based on twitter -
  - Austria top 10 users having highest austrian followers.
- This baseline proposed new metric named SNP (**S**ocial **N**etworking **P**otential) for ranking nodes in the network and compared the same with existing ranking methods authoritative score, page rank and Klout score.
- Author proposed 3 indicators -
  - Follower Followee Ratio
  - Retweet Ratio
  - Interactor Ratio



Top 5 Kin/Kout	Top 5 SNP User
Kamal R khan	officeofbagga
Paranita subhash	Bawanaaved
NDTV	Voice_for_India
Sadhguru JV	mirsuhail
ANI	Hasanminhaj





# Proposed Solutions

# Feature Generation



Before applying any model sentiments of tweets were analyzed from tweet text and after that following features were extracted or derived from the dataset.

- Tweet time
- Retweet time
- Tweet likes
- Tweet retweet
- Follower count
- Statuses count
- Tweet sentiment
- Activity measure
- Topical Strength
- Follower rank
- Popularity
- Degree centrality
- Follower-Followee ratio

# Machine Learning Models



- Develop a Machine Learning models for the identification of influencer.
- Make features using user properties and graph properties.
- Make correlation matrix between features and remove highly correlated features.
- Develop the ground truth by applying the threshold value on each feature.
- Use three models:
  - Random Forest
  - Artificial Neural Network
  - Multi layer perceptron
- Classify the user as Influencer or non-influencer.

# Graph Neural Network



- Used semi-supervised graph convolution network as normal neural networks are not able to capture the structure information from the graph.
- Total of 6 features were extracted from the dataset and feature vector is generated for each nodes.
- Initially ground truth annotation for 200 nodes were defined by analyzing information from user real twitter accounts.
- Developed two layer graph neural network from the network.
- Classify the user as Influencer or non-influencer.

# Detection through Retweets



- Retweet Detection in the tweeter network is an important task as it helps us to detect which tweet the user is going to retweet next related to the particular topic.
- Two graphs were used for this purpose-
  - Follower-Followee network from retweet graph.
  - User-retweet graph.
- For the task of the next possible retweet detection, some edges were removed from the retweet-user network.
- Use previous tweets to analyze previous tweets features like language, sentiment, etc.
- Use follower-follower graph to find follower related data.
- Use information from both graph to find next possible retweet.

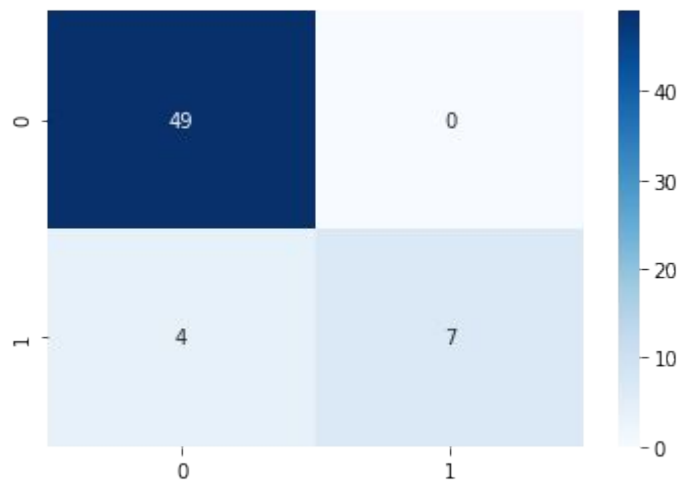
# Results



- The table shows the evaluation metrics of the models

Model Name	Accuracy	Precision	Recall	F1-score
Random Forest	90.52	90.05	90.18	90.13
ANN	94.44	94.04	93.90	93.98
MLP	96.34	96.12	96.20	96.25
GCN	93.33	96.22	81.81	86.92

# Results



- Confusion matrix of GCN shows the non-influencer nodes are correctly predicted by the model.
- None of the non-influencer is influencer.

# Sample output for MLP



Influencer supporting CAA. He was famous for series of posts in the favour of CAA which had high number of retweet



# Sample output for GCN

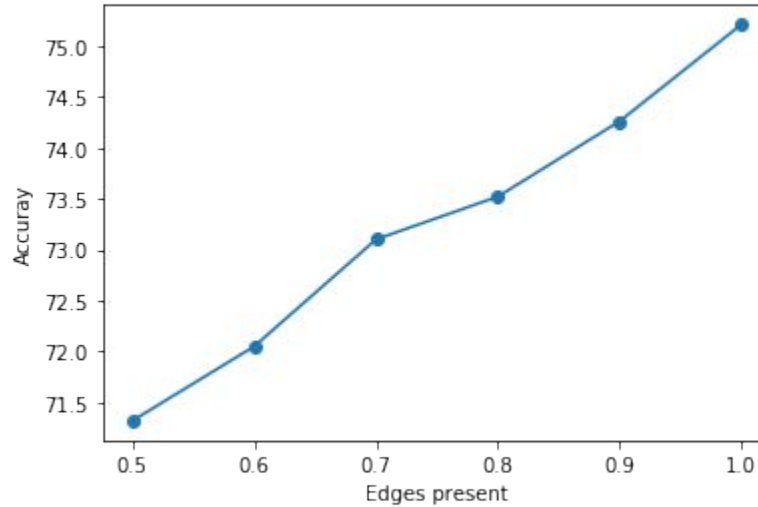


Influencer supporting CAA. He was famous for series of posts in the favour of CAA and one of the tweets he posted became too popular.



# Results for detection through retweets

- Accuracy for this model was around 72.35% and the accuracy of the model does not change much even though we remove up to 50% of previous retweet data.



# Challenges



- Data retrieval takes time.
- Difficult to incorporating multiple metadata at the same time.
- As the number of non influencer are far more than number of influencer, due to this reason it was hard to train the models.

# Conclusion



- From the above solution we can identify the influencers upto certain extent and GCN results are very prominent as it can capture the topological properties of the graph also.
- Class imbalance problem.
- Verified user and follower information about nodes were needed.
- Graph Sage and Graph attention networks can also be tried in the future to find the influencer in the graph.



**Thank you**



## References

- [1] Zhang, Xin & Han, Ding-Ding & Yang, Ruiqi & Zhang, Ziqiao. (2017). Users' participation and social influence during information spreading on Twitter. PLOS ONE. 12. e0183290.  
10.1371/journal.pone.0183290.
- [2] Anger, Isabel & Kittl, Christian. (2011). Measuring influence on Twitter. I-KNOW. 31.  
10.1145/2024288.2024326.