Twitter network analysis shows few focused followers elevate influential leaders

William B. Andreopoulos[*]
Department of Computer Science
San José State University
San Jose, CA, USA
william.andreopoulos@sjsu.edu

Fei Pan[*]
Department of Computer Science
San José State University
San Jose, CA, USA
fei.pan@sjsu.edu

Mohit Sonwane[*]
Department of Computer Science
San José State University
San Jose, CA, USA
mohit.sonwane@sjsu.edu

Abstract - **Twitter influencers have played a key role in disseminating ideas. However, understanding the dynamics of how subnetworks of influence emerge is a matter of contention. We analyzed 286,000 tweets over a period of 3 months to understand the distributions of replies and retweets, and understand how they relate to leaders and followers. We found there exist many followers who reply or retweet few times and few who reply many times. While low influence leaders get retweeted much more in aggregate and by more followers than high influence leaders, any users retweeting them do not focus on any one individual, which may explain their remaining low influence. On the other hand, the high influence leaders are retweeted or replied to by a smaller number of followers, but they also attract more focused attention per leader than low influence leaders. In our data sample a relatively modest or small number of followers contribute to elevating leaders in the Twitter network, rather than a general wave propagating throughout everyone in the network.**

Keywords—Twitter distribution network analysis follower leader

I. INTRODUCTION

Followers are particularly important for Twitter influencers because they are the primary metric that brands and advertisers use to evaluate the reach and influence of an influencer. Influencers are individuals who have a significant following on social media, and their opinions and recommendations can impact the behavior of their followers. For Twitter influencers, having a large and engaged following is crucial for building their brand, increasing their influence, and monetizing their content through partnerships and sponsorships. The more followers an influencer has, the greater their potential reach and the more valuable they become to brands and advertisers looking to promote their products or services to a wider audience. Additionally, followers are important for influencers because they can help to boost their credibility and authority. When an influencer shares content that resonates with their followers, it can generate likes, retweets, and comments, which can help to increase the visibility and virality of the content [1].

Twitter replies and retweets can help to promote influencers by increasing their visibility, providing social proof, amplifying their message, and encouraging engagement with their content. Replies and retweets can help promote influencers as follows:

Increased Visibility: When an influencer's tweet gets retweeted or replied to, it is exposed to a wider audience, increasing the potential reach of the influencer's message.

Social Proof: When people see others replying to or retweeting an influencer's tweet, it can provide social proof that the influencer's content is valuable or worth sharing.

Amplification: Replies and retweets can amplify the message of an influencer, increasing the likelihood that the message will be seen and shared by others.

Engagement: When users reply to an influencer's tweet, it can lead to increased engagement with their content. This can help to build a community around the influencer, which can further enhance their visibility and influence [7].

II. RELATED WORKS

Guilbeault et al. (2021) studied measures of complex path length and complex centrality that could improve the capacity to identify the central individuals who spread complex contagions. They found the classical measure of path length fails to define network connectedness and node centrality, which would help to detect node influence in the spread of complex contagions [2].

Lu et al. (2014) studied the degree distribution of retweets in the Twitter network [3]. They present a hypothesis that the frequency distribution of retweets follows a power law distribution by analyzing the retweets data. A simulation done produced a power law distribution for retweets. Tong et al. (2016) also hypothesized that the Twitter network comprises of two subnetworks that follow a power law distribution and present a formal model. They extracted two subnetworks, the Social Network containing all mutual relationships, and the Information Network containing all the one-way relationships to look for the power law distribution [6].

Zhao et al. (2013) utilized the lessons learned to build recommendation systems that are not sensitive to making popular users more popular [4]. Lau et al. (2012) present novel dynamic topic modelling-based methodologies to track emerging events in microblogs such as Twitter [5]. Ramponi et

[*] the authors contributed equally

al. (2019) validated the community structure of Twitter on the basis of content, which is similar within groups [8]. Paudel et al. (2019) extracted topic embeddings to describe user interactions and evaluated them based on friendship recommendations and retweet behavior prediction [9]. Grandjean et al. (2016) studied the small-worldness of Twitter networks by explaining clustering of users on the basis of linguistic groups [10].

III. Methodology

A. *Dataset*

All the data Twitter generates is essentially in the public domain. The first instinct is to web-scrape tweets but Twitter has strict no-tolerance policies for such an activity. Twitter understood the necessity of making their data accessible to the researchers and decided to improve their external API for tweet extraction, including a timeline, user tweets, and tweet search.

To make use of these tweets for the purpose of data analysis, they need to be in a network format. We represent the network as a directed adjacency list with pairs of follower-leader, where follower retweeted or replied to leader [4].

B. *Technicals*

Each tweet is required to contain an user id and at least one hashtag. The user id is always a required field but a hashtag is not necessary for retrieving a tweet. To collect tweets data, Tweepy was used to leverage Twitter API v2. With keys and tokens from twitter developer account for authentication, the endpoint of `search_all_tweets(query,...)` was called and an operator "has:hashtags" was added in the query to get the tweets that meet our demands. However, this operator is a conjunction required, and needs to work with other standalone operators. We added an user id to the query (query = from: xxxx(user id) + has:hashtags), where user ids can be collected from a user P's followers and are ensured as active users with a threshold for an average daily post per user. Here, we assumed P's followers are neutral enough with less bias. With a user id list, we requested tweets iteratively for each user. We assigned a query string with user id, limit the tweets must contain one or more hashtags and for English language only. Limited dates range from 01/01/2022 to 04/30/2022. Finally, 286,044 tweets (with hashtags) among 388 users were collected and the corresponding tweets information was parsed from dictionary to dataframe and then saved to .csv files.

IV. Experiments

We retrieved a Twitter reply network of 8,321 follower-leader pairs, where follower replied to leader. We also retrieved a Twitter retweet network of 24,927 follower-leader pairs, where follower retweeted leader. For replies 2,766 of those replies were to “high influence” users, defined as those who received 3 or more replies overall; 5,555 of those replies were to “low influence” users, defined as those who received just 1 or 2 replies. For retweets 8,334 of the retweets were to “high influence” users who got retweeted 3 or more times overall, and 16,593 retweets were to “low influence” users who got just 1 or 2 retweets.

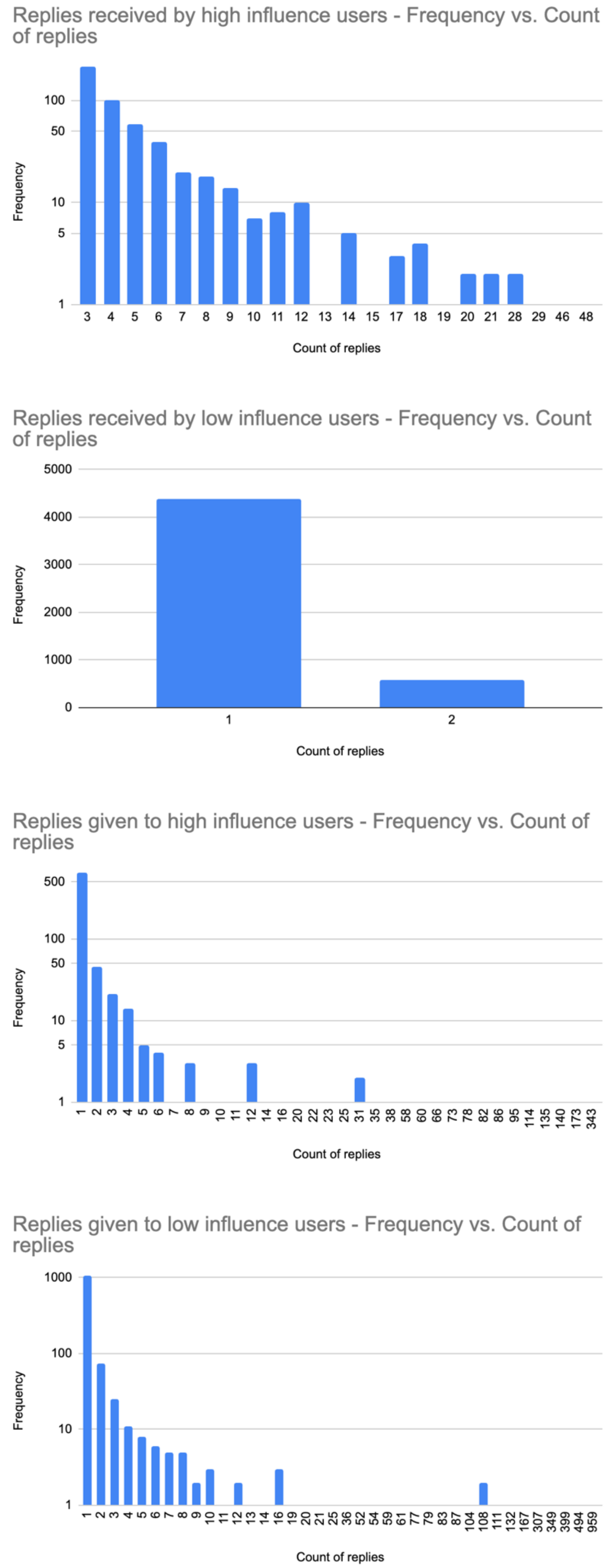

Figure 1 – Replies histograms

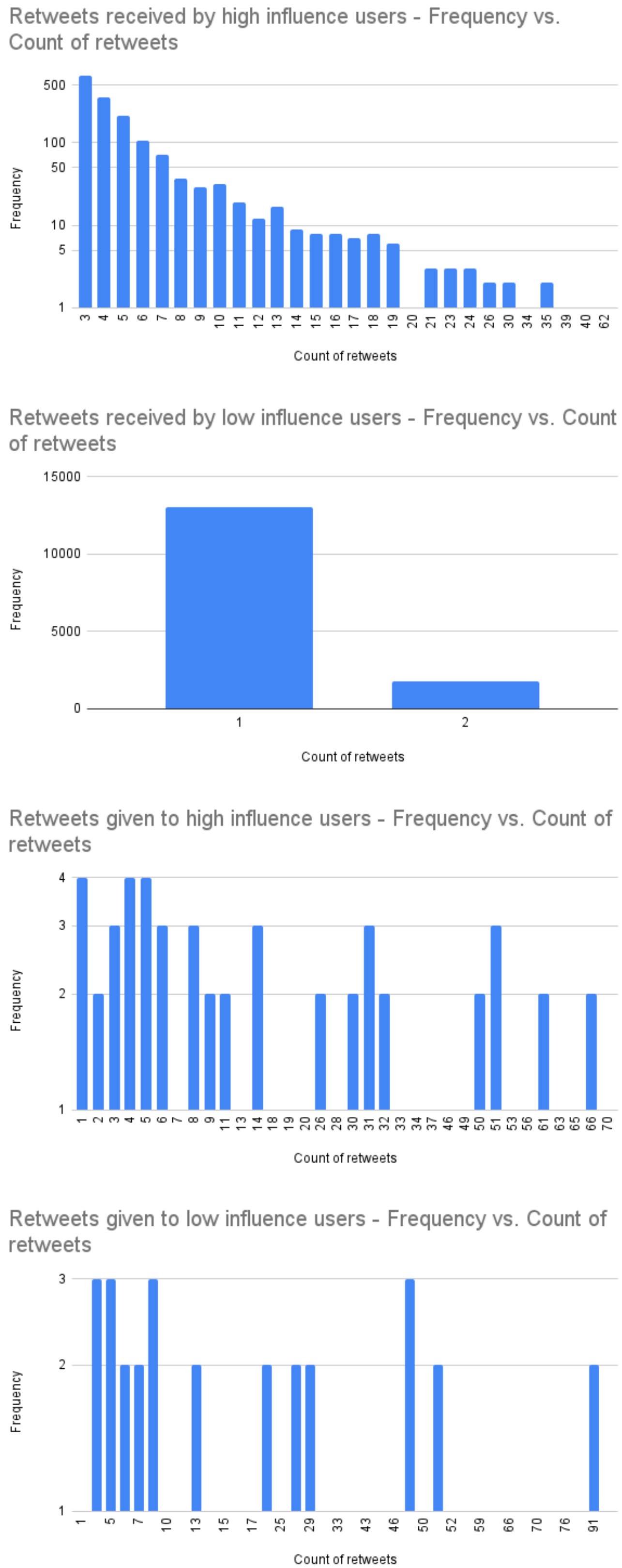

Figure 2 – Retweets histograms

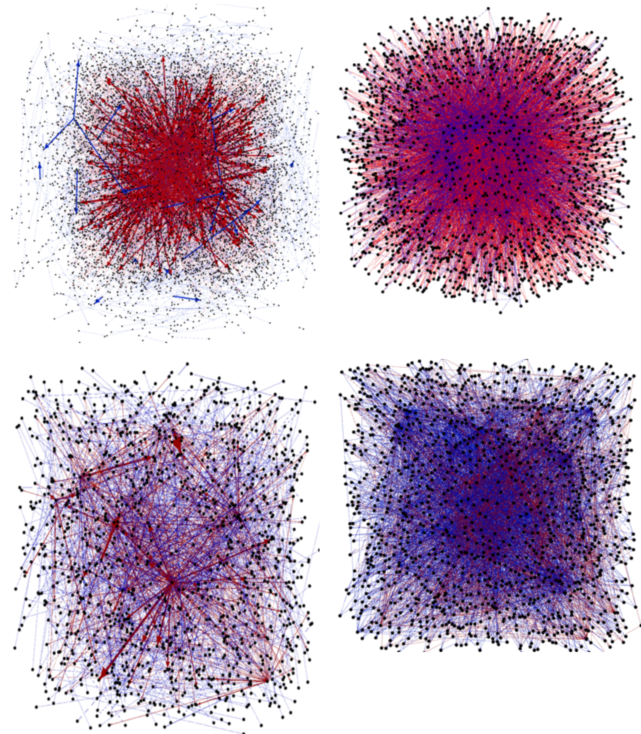

Figure 3 - *Top left*: High influence replies graph. *Bottom left*: Low influence replies graph. *Top right*: High influence retweets graph. *Bottom right*: Low influence retweets graph. These were produced with the Gephi network plotting tool. Red edges are replies of retweets emanating from top 40 repliers or retweeters and blue edges emanate from all others. The density of red edges shows that the top 40 repliers and retweeters are much more active in the high influencer than in the low influencer networks.

Next we analyzed the followers' distribution in the replies network. While there are 1239 followers who replied to 4966 low influence leaders, they replied 5555 times in total (4-5 times each), which suggests they each distribute their replies across many low influence leaders and are not focused on any leader in particular. The 515 highly influential leaders who got 3 or more replies each, on the other hand, got half as many replies in total (2766) from a modest number of 774 followers. This suggests that an influential leader may be elevated by a relatively small number of focused followers. The reasoning is that most potential follower users are unfocused and distribute their attention across many other leaders, thus any spark of attention from a modest-sized group can distinguish a potential leader.

To understand this better we compared the top 40 vs. bottom 40 repliers next. Almost all top 40 repliers reply to both high and low influence leaders (39 and 40), but they reply twice more to low than high influence leaders (4089 vs. 1956), which is explained since there are twice more low influence than high influence users. The top 40 repliers replying to both leaders who are high or low influencers (39 vs. 40), combined with many more replies to low influence (4089 vs. 1956), suggest the top repliers as a group focus neither on highly influential users nor any users in particular.

The bottom 1554 repliers (who each replied just once) also replied almost twice more often to low influence (992) than to high influence leaders (562). As expected, the bottom repliers reply much less than the top 40 repliers above overall: 562 vs. 1956 replies to high influence leaders, and 992 vs. 4089 replies to low influence leaders. The bottom repliers produce fewer replies to many more leaders than top repliers, thus distributing their attention across a large number of mostly low-influence leaders. On the other hand, the top repliers and retweeters respond to fewer leaders many more times than bottom repliers.

TABLE I. TWITTER DATASET REPLY STATISTICS

Replies	Low infl leader	High infl leader
Unique follower replying	1239	774
Unique leader replied to	4966	515
Ratio	1239/4966	774/515
Total replies	**5555**	**2766**
Total replies from top 40 follower repliers	4089	1956
Count of top 40 repliers who replied to a high / low leader	40	39
Total replies from bottom 1554 follower repliers (with only 1 reply)	992	562
Count of bottom 1554 repliers (with only 1 reply) who replied to a high / low leader	992	562

For retweets we removed the redundant lines, as there were twice more redundant than unique retweets. After removing redundant links, the number of retweets was almost halved from 32345+15567 to 8334+16593 (high+low) with 1503 of the high influence users re-classified as low influence by our definition. After removing redundant links there were 14818 low influence leaders getting retweeted once or twice and 1602 high influence leaders getting retweeted three or more times for a total of 16593 and 8334 retweets, respectively.

For retweets one of the most interesting findings is that for both the most retweeted (highly influential) leaders and the less retweeted (lower influential) leaders there are significantly fewer retweeting followers doing the retweets than there are leaders who get retweeted. There are only hundreds of retweeting followers, while there are 14818 and 1602 low and highly influential retweeted leaders. This shows that follower users are unfocused and distribute their attention across many other low influence leaders who get retweeted.

Retweets gave a similar result as replies: all 40 top retweeters reply to both high and low influence leaders. Though they reply twice more to low than high influence leaders (12071 vs 6092), that is probably because there are significantly more low influence than high influence leaders (14818 vs. 1602).

The bottom 60 retweeters retweet much less than the top 40 retweeters above: 1460 vs. 6092 in the high influencer network, and 2467 vs. 12071 in the low influencer network. The bottom 60, like the top 40, also retweet significantly more to low influence (2467) than to high influence (1460) leaders. The top 40 and bottom 60 retweeters retweet about the same for leaders who are either high (40 and 54) or low influencers (40 and 59), despite bottom retweeters producing significantly fewer retweets.

Therefore, it is plausible that a few hundred retweeting users can give rise to a large network of influencers. When comparing retweets with replies we can see that the bottom retweeters tend to retweet more than once, while the bottom repliers mostly reply just once. This might be a result of the ease of retweeting as compared to typing.

TABLE II. TWITTER DATASET RETWEET STATISTICS

Retweets	Low infl leader	High infl leader
Unique follower retweeting	113	108
Unique leader retweeted	14818	1602
Ratio	113/14818	108/1602
Total retweets	**16593**	**8334**
Total retweets from top 40 follower retweeters	12071	6092
Count of top 40 retweeters who retweeted a high / low leader	40	40
Total retweets from bottom 60 follower retweeters	2467	1460
Count of bottom 60 retweeters who retweeted a high / low leader	59	54

A subset of 174 of the follower repliers to low influence leaders also replied to high influence leaders, while just 70 of the low influence leaders replied to high influence leaders. Similarly, in retweets a subset of 107 of the follower retweeters of low influence leaders also retweeted high influence leaders, while just 1 low influence leader replied to high influence leaders. Therefore there is little overlap between the categories.

In addition, the top repliers and retweeters appear to reply in "bursts" on various days. Below is a list of anonymized top 10 repliers and retweeters with the number of replies given on various days.

```
Replies:
165 user1 2022-03-04
98 user2 2022-02-23
79 user3 2022-02-04
77 user4 2022-04-23
74 user5 2022-04-08
66 user6 2022-03-11
63 user7 2022-02-18
55 user8 2022-04-02
52 user9 2022-03-25
40 user0 2022-03-18
```

```
Retweets:
165 user1 2022-02-27
157 user2 2022-03-02
102 user3 2022-04-20
94 user4 2022-02-28
92 user5 2022-02-26
92 user6 2022-04-19
88 user7 2022-01-27
76 user8 2022-03-10
70 user9 2022-02-28
65 user0 2022-03-01
```

V. Discussion

Our first observation is that there are twice as many users who reply or retweet just once or twice, than there are users who reply or retweet many times (3 or more times). Thus, there are many users who reply once or twice, and few users who reply 3 or more times in both low and high influencers' networks.

There are fewer followers doing the retweets than there are getting retweeted, which means followers retweet many different leaders. It holds for both high and less influential leaders. Moreover, high influence leaders are associated with fewer followers replying or retweeting them and producing about half the total replies or retweets than low influence leaders in aggregate. Both replies and retweets show the low influence users get less focus from the many users retweeting or replying to them, meaning that followers are still very active with low influence users, but just don't focus on any particular user. Whereas low influence leaders get retweeted much more in aggregate and by more followers than high influence leaders, users retweeting the low influence do not focus on any one individual, thus they all remain low influence. On the other hand, the high influence leaders are retweeted or replied to by a smaller number of followers than low influence leaders, but they also attract more focused attention per leader.

VI. Conclusion

We studied a "follower-repliesto/retweets-leader" twitter network. As conclusion, there are way fewer high influence users overall, produced by a smaller number of followers who overall make less noise than the low influencers' followers in aggregate. The cause of high-influence leaders could be the focus of a relatively modest or small number of followers, rather than a general wave propagating throughout everyone in the network. The characteristics of the group of followers that elevated a highly influential leader are unclear and remain to be studied to find any frequent patterns.

References

[1] A. Mitchell, E. Shearer, and G. Stocking, "News on Twitter: Consumed by Most Users and Trusted by Many," PewResearch.org. https://www.pewresearch.org/journalism/2021/11/15/news-on-twitter-consumed-by-most-users-and-trusted-by-many/ (accessed Feb. 3, 2023).

[2] Guilbeault, D., Centola, D. Topological measures for identifying and predicting the spread of complex contagions. *Nat Commun* **12**, 4430 (2021). https://doi.org/10.1038/s41467-021-24704-6

[3] Yao Lu, Peng Zhang, Yanan Cao, Yue Hu, Li Guo, On the Frequency Distribution of Retweets, Procedia Computer Science, Volume 31, 2014, Pages 747-753, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2014.05.323.

[4] X. Zhao, Z. Niu, and W. Chen, "Opinion-based collaborative filtering to solve popularity bias in recommender systems," in Database and Expert Systems Applications: 24th Int. Conf., DEXA 2013, Prague, Czech Republic, August 26-29, 2013. Proc., Part II 24,2013,pp.426–433.

[5] Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line Trend Analysis with Topic Models: #twitter Trends Detection Topic Model Online. In Proceedings of COLING 2012, pages 1519–1534, Mumbai, India. The COLING 2012 Organizing Committee.

[6] M. Tong, A. Sanzgiri, D. Koutsonikolas and S. Upadhyaya, "Twitter structure as a composition of two distinct networks," *2016 International Conference on Computing, Networking and Communications (ICNC)*, Kauai, HI, USA, 2016, pp. 1-5, doi: 10.1109/ICCNC.2016.7440622.

[7] Ediger, David & Jiang, Karl & Riedy, Jason & Bader, David & Corley, Courtney & Farber, Robert & Reynolds, William. (2010). Massive Social Network Analysis: Mining Twitter for Social Good. Parallel Processing, International Conference on. 583-593. 10.1109/ICPP.2010.66.

[8] Ramponi, Giorgia & Brambilla, Marco & Daniel, Florian & Di Giovanni, Marco. (2019). Content-based characterization of online social communities. Information Processing & Management. 57. 102133. 10.1016/j.ipm.2019.102133.

[9] Paudel, Pujan & Hatua, Amartya & Nguyen, Trung & Sung, Andrew. (2019). User Level Multi-feed Weighted Topic Embeddings for Studying Network Interaction in Twitter. 10.1007/978-3-030-23551-2_6.

[10] Grandjean, Martin. (2016). A social network analysis of Twitter: Mapping the digital humanities community. Cogent Arts & Humanities. 3. 1171458. 10.1080/23311983.2016.1171458.