Twitter analysis on CHUANG2021 contestants' popularity with diffusion networks

Jo Pan, tul02009@temple.edu
April 26, 2021

Abstract

Along with the increasingly high level of digital media consumption, numerous new buzz-out figures are created to fill our social media feeds. This project aims to study how new digital influencers' popularity diffuses in social media using the social network analysis perspective. I used two CHUANG2021's contestants, Liu Yu and Mika, as the main study targets. First, I created two networks with their related tweets scraped from Twitter during the competition period. Second, I analyzed the topological characteristics of their networks and how their networks progress during the competition. Lastly, through simulation with diffusion model, I discovered that with only 10 initial influential fans, contestants' popularity can easily diffuses throughout the network. GitHub: `https://github.com/Jo-Pan/Twitter_analysis_CHUANG2021`

1 INTRODUCTION

Nowadays, new buzz-out figures, including idols, celebrities and content creators, appear in social media everyday. How does these digital influencers' popularity start and diffuse? In this project, I try to answer this question from the social network perspective.

Two CHUANG2021's contestants are selected as the main study targets for this project. CHUANG2021 is a Chinese male group survival reality show streamed every Saturday from February 17, 2021 to April 24, 2021 [1]. The show brings 90 male trainees from different countries and agencies, in order to form an 11-member international boy group through global viewers' votes. Because most contestants started out as infamous trainees, the increase in their popularity during the show was apparent. As contestants' popularity increases, their fans voluntarily created and promoted the contestants in the social media,

including Twitter, Facebook and Weibo. Thus, by studying contestants' popularity in one of the social media platform, Twitter, I can obtain good estimation and understanding to their overall popularity and fan communities. Also, since the show just started not long ago, it reduces the challenges for data acquisition.

The contribution of this project is three-fold. First, this project will be the initial work of using Twitter and social network analysis to understand the progress of digital influencers' popularity. Existing works relating to social network analysis with Twitter are mostly for certain events, political figures or without specific themes. Second, for content creators and agencies, this project provides new ideas for formulating popularity measurements based on network statistics. These network-based measures can provide additional perspectives on top of current industry's standard measures, such as number of followers and average retweets. Network-based measures also allow advertisers to better estimate how their products will diffuse in the social media when choosing collaborators among digital influencers. Third, in terms of fan management, this project suggests ways to identify the "fan leader" in the digital space, whose opinions or posts have large diffusion rate in the fan community. Targeting these "fan leaders" can improve the efficiency of fan management and fan acquisition.

In overview, this project consists of three phases: 1) Data acquisition and preparation. I scrapped and pre-processed related tweets from Twitter. Then, for each contestant, I built a directed network representing their Twitter presence during the time period. 2) Topology analysis. I analysed and compared network characteristics, including connected components, diameters and density. "Fan leaders" are identified through node characteristics. Through time series analysis, I analysed the progress of their network characteristics as the competition progress. 3) Diffusion model simulation. Using diffusion models, I simulated how the contestant's popularity diffuses in their network, and analyzed the behaviors of these diffusion models.

2 Related Work

Twitter is a popular research domain for social network analysis. Pierri et al. did a topology comparison of Twitter diffusion networks about misleading news [2]. My network construction mechanism is highly inspired by their work, using users as nodes and "mention" as links.

Studying information propagation with diffusion models have been done by Pierri et al.[2] and Akrouf et. al.[3]. Akrouf et. al. used Linear Threshold Model, Independent Cascade Model and Weighted Cascade Model to experiment with Flick user connection network and YouTube user comments network. I borrowed the idea and implemented Linear Threshold Model and Independent Cascade model to simulate network for the two selected contestants.

3 METHODOLOGY

3.1 DATA ACQUISITION AND PREPARATION

The two selected contestants are Liu Yu and Mika. I scrapped their related tweets with two Python libraries, Snscrape and Twython. Snscrape library is used to download sample historical tweets since the show starts (02/17/2021), without the need of official Twitter API. Twython is used to download sample tweets within a week from the official Twitter API. The official Twitter API provides access to a larger amount of tweets and retweets comparing to Snscrape. Thus, two libraries are used together. On Temple's high-performance computing server, I wrote a script to scraped using Snscrape once every two hours and using Twython once every 15 minutes since 03/19/2021.

In terms of query keywords, I constructed a set of keywords. Each keyword is a combination of the show name and the contestant name (eg: chuang2021 liuyu). For show name, I used chuang2021 and chuang. For contestant name, I used liuyu, liu yu, and mika. For all keywords, there is also a Chinese version included, because the show is in Chinese.

For network construction, all unique authors are mapped to a node ID. When user A mentioned, retweeted or replied to user B in A's tweets, a directed link is constructed pointing from B to A. This direction of link allows me to simulate how information diffuses from the author to audiences.

3.2 TOPOLOGY ANALYSIS

Two separate networks are created for Liu Yu and Mika. In terms of topology analysis, the following network characteristics are calculated: number of vertices, number of edges, graph density, number of connected components, maximum number of vertices/edges in a connected component, and clustering coefficient. Since the network is too large for betweenness calculation for all nodes, I calculated for the top-1000 nodes with the highest degree for each network. I also evaluated overall degree distribution.

In order to understand how Liu Yu and Mika gained popularity during the show. I did a time series analysis on some of their network characteristics. First, I separated the tweet dataset based on dates. The dates chosen are all Saturdays, which are the scheduled day when most of the episodes were released. For time series analysis, I analyzed the progress of number of tweets, number of edges, number of vertices, density and clustering coefficient across the time period.

3.3 DIFFUSION MODEL SIMULATION

For diffusion model simulation, I did experiments on both contestants' networks. There are two types of node status: active and not active. Active can be interpreted as favor toward a contestant. I simulated with Threshold Model (TM) and Independent Cascade Model (ICM). For TM, I set each node's threshold as 0.5. For ICM, I set each edge's threshold as 0.5.

Next, I will describe how I chose the initial active set or the initial adopters using a similar process as [3]. The problem is posed as follows: "if you want to trigger a large cascade of

favor toward a contestant, you need first convince a subset of social media users to like the contestant. In this case, which set of users (nodes) should you target initially?". In this project, since I focus only on the structure of the network, the initial active set is chosen based on the structural characteristics of nodes.

The first initial active set is chosen based on the degree, which is the count of total number of connections linked to a node. In my experiment setting, the degree measures how many users a user can reach directly in the network. Also, since the networks are all directed, in-degree and out-degree for each node are also computed. In-degree represents how many links point inward to a node, which equals to the number of unique users that the user have retweeted from. Out-degree is the number of links point outward to other nodes, which equals to the number of unique users that have retweeted from the user. The sum of in-degree and out-degree is overall degree. I choose the initial adopters based on the high out-degree, high in-degree and the high overall degree values.

The second initial active set is basing on random. The random set allows me to compare and prove the important role of choosing the initial adopters for spreading the popularity widely in Twitter.

The size of initial active set depends on the size of network, its type, and the domain of the experiments. For example: in marketing, targeting a large size of initial adopters set can be costly since one has to pay for each initial adopter. On the other side, spreading a post in Twitter, doesn't require much cost. Then targeting a large initial adopters set will be better. Thus, in my experiments, I tested various size of initial active nodes $N_0 \in \{10, 100, 200, 500, 1000\}$.

The selected diffusion models, TM and ICM, were implemented on both contestants' networks. For each diffusion model, the propagation process was run for five times for every initial set, then the size of active nodes after each run was recorded. This number is considered to be the influence of the initial set.

4 Results

4.1 Data overview

A set of anonymized sample data scrapped from Twitter is shown In Figure. 1. I scrapped the user id, tweet date, tweet content and user name. The keyword column shows the keyword used for querying the tweet. The mentioned column is derived from content, by locating the user name after "RT @". I created a map from username to node id and used that to convert user name and mentioned user to node and m node column respectively. The Liu Yu column and Mika column indicates which candidate the tweet is relating to. There is a small amount of tweets mentioning both Liu Yu and Mika.

Overall, I scrapped 505,814 tweets from 2021-02-17 (competition start date) to 2021-04-24 (competition end date). There are 460,732 (93%) retweets. For Liu Yu, the total number of related tweets is 190,015, which is 38% of total number of tweets. For Mika, the total number of related tweets is 370,153, which is 75% of total number of tweets. In terms of users, in total, there are 53,413 unique users in the whole dataset.

Figure 1: Anonymized sample data scrapped from Twitter.

id	date	content	username	keyword	Liu Yu	Mika	mentioned	node	m_node
13710----	2021-03-14 10:32:48+00:00	RT @CHUANG_Official: #ChuangThemeSong #CHUANG2...	zydea-----	liuyu #CHUANG2021	True	False	[CHUANG_Official]	33618	[33854]
13803----	2021-04-09 01:48:25+00:00	RT @Mika_THOfficial: ชุ้มของมิกะสำหรับงานวันนี...	bkpmk-----	mika #CHUANG2021	False	True	[Mika_THOfficial]	34935	[34285]
13790----	2021-04-05 10:37:59+00:00	RT @LiuyuThailand: 210405 | 今朝_刘宇 Weibo Update...	vivin-----	liu yu #创造营 2021	True	False	[LiuyuThailand]	31636	[34213]
13705----	2021-03-13 03:14:27+00:00	RT @liuyuarchive: — liuyu dancing to Dose 🕺 ye...	itskm-----	liuyu #CHUANG2021	True	False	[liuyuarchive]	35487	[35769]

Table 1: Structure characteristics of the contestant networks

	Liu Yu	Mika
Number of vertices	28,284	45,889
Number of edges	92,399	171,365
Graph density	0.00012	0.00008
Number of connected components	81	199
Maximum number of vertices in a connected component	28,108	45,448
Maximum number of edges in a connected component	92,299	171,113
Clustering coefficient	0.0790	0.1252
Maximum betweenness	0.2097	0.2989

4.2 Topology analysis

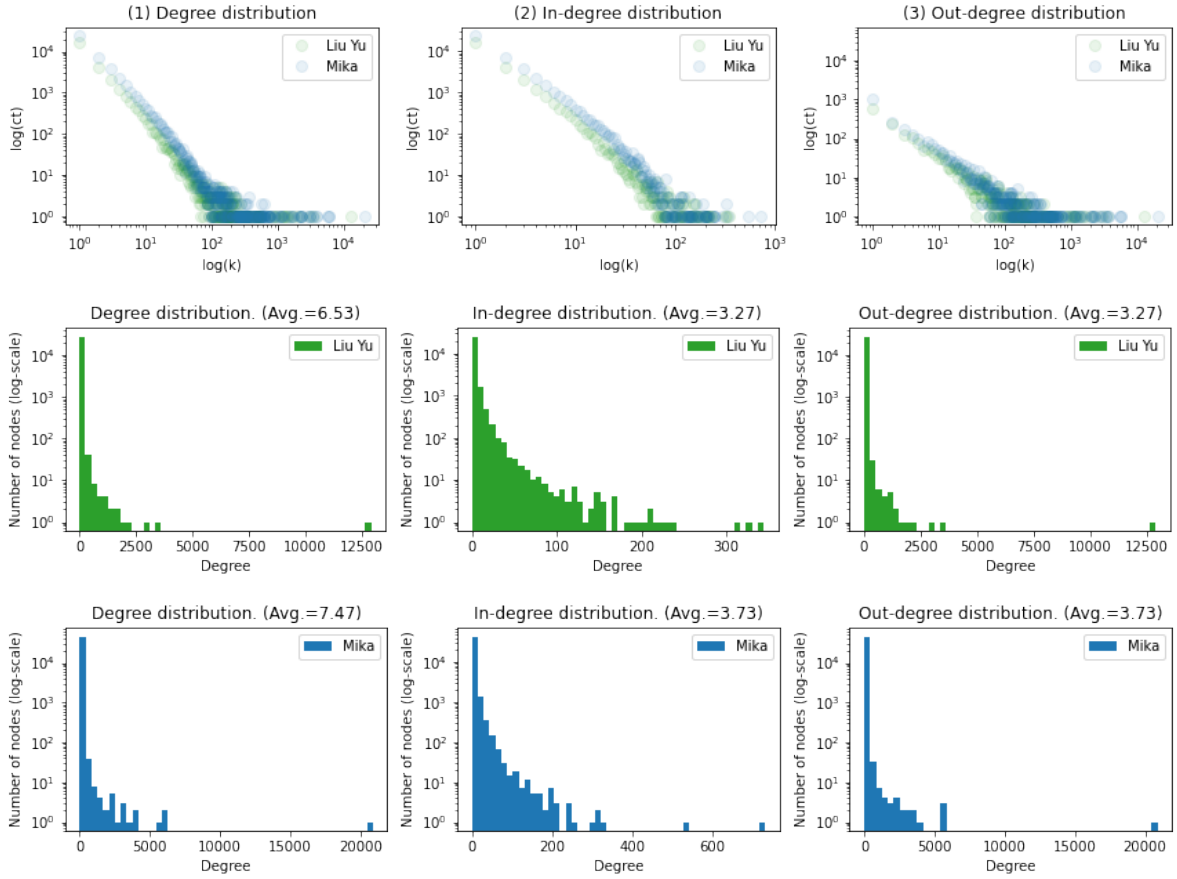
4.2.1 Overall network characteristics

Table. 1 summarizes the structural characteristics for both networks. Comparing to Mika's network, Liu Yu's network has a smaller number of nodes, connected components, clustering coefficient and betweenness. These characteristics may all suggest that Mika has a higher popularity than Liu Yu. The node with maximum betweenness centrality in both network is the official account, "CHUANG_Official".

In terms of degree distribution, both networks are highly similar. Figure. 2 shows the degree (k) distribution. Overall, both networks follow the power law distribution. Noticeably, the in-degree for both networks is significantly lower than the out-degree. This can be explained by the nature of Twitter, as it is easier to mention many users in tweets than writing a popular tweet which is retweeted by many.

I visualized the two networks with Gephi, which is an open-source visualization and exploration software for networks. The resulting network visualization and details can be seen in Figure. 3. For Figure. 3(a-b), the visualization used the ForceAtlas2 graph layout algorithm [4] and used the Dissuade Hubs mode. For Figure. 3(c-d), the visualization used the Fruchterman-Reingold graph layout algorithm for nodes [5]. In order to focus on the important nodes, I used a threshold on node degree. For Figure. 3(c-d), only node degree greater than 400 are visualized.

Figure 2: Degree distribution. The first row is in log-log scale.



The top 10 nodes with maximum out-degrees are shown in Figure. 5, with their longest shortest path annotated at the end of each bar. These nodes can be treated as the fan leaders in each networks. For both networks, "CHUANG_Official" is the node with the maximum out-degree. The out-degree of "CHUANG_Official" is 40% higher in Mika's network than in Liu Yu's network. In addition, all the remaining fan leaders in Mika's network have a higher out-degree than Liu Yu's. These may be a result of the higher popularity of Mika in Twitter. However, in terms of longest shortest path, which represents the depth of diffusion, fan-leaders in both sides have similar numbers ranging from 4 to 8. This means that the length of discussion chain in both fan groups are about the same.

Overall, if agencies for Liu Yu and Mika want to promote their contestants in Twitter, agencies can take care of the ones who has high impacts, indicated by the high out-degree value. Secondly, since most contestant's related content is from CHUANG2021's official team, to improve the diversity, they can generate related contents, such as clips, and distribute to those fan leaders. This will provide a more diverse source of contents being used for contestant related-discussion. Last but not least, they should be thankful for their contestants being included in many official tweets.

Figure 3: Network visualization

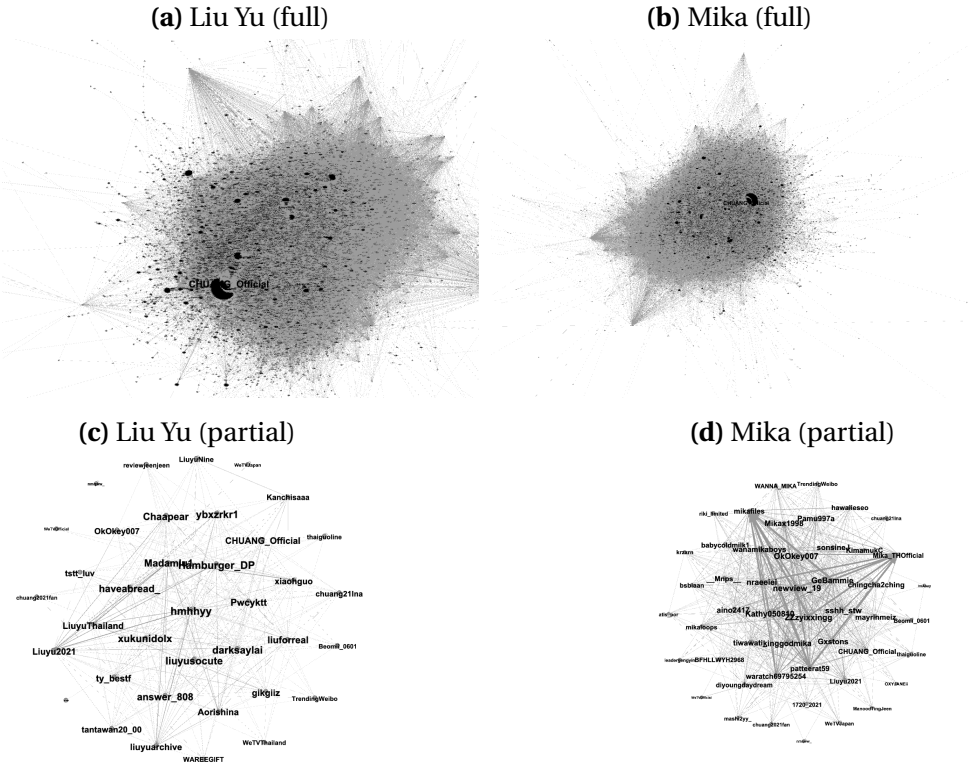
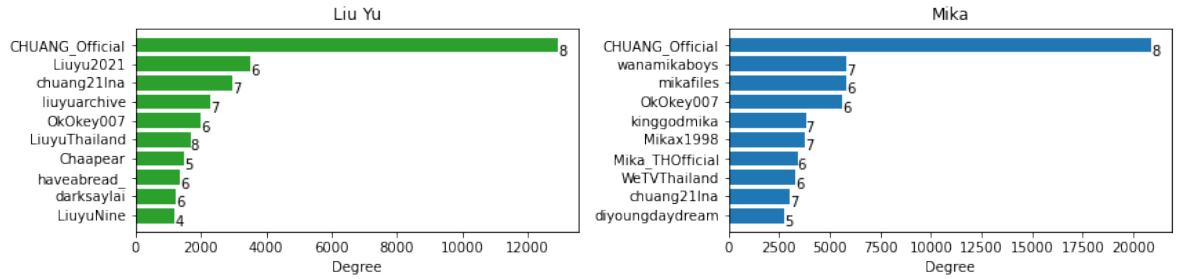


Figure 5: Top-10 nodes with maximum out-degree. The number next to each bar is each node's longest shortest path in each network.

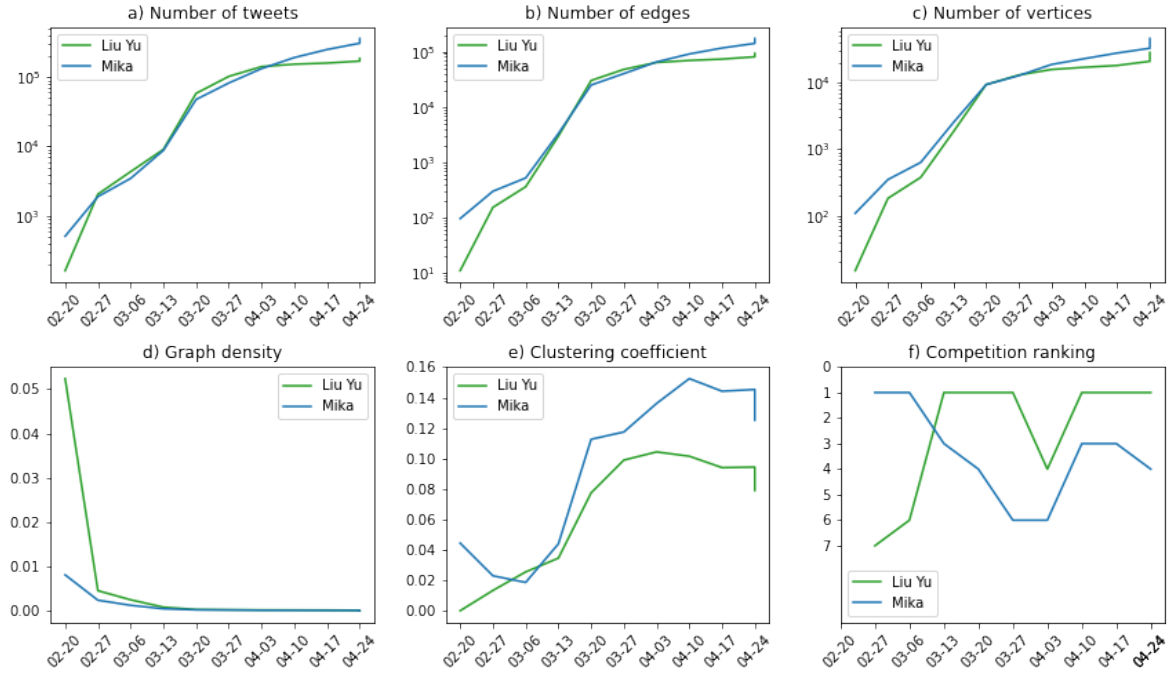


4.2.2 TIME SERIES ANALYSIS

Figure. 6 shows, as each episode released, the progress of network characteristics and competition ranking, which is the psudo ground truth for popularity. For total number of tweets, number of edges and number of vertices, referred as "statistics" in following discussions, Liu Yu was behind Mika at the first 1-2 weeks. The reason was that Mika has been debuted in Japan for years and has significant international fan foundation. Thus, at the beginning of the competition, Mika's statistics were always better than Liu Yu's

In March, Liu Yu's statistics got over Mika's. The potential reason was that Liu Yu ranked

Figure 6: Time series analysis in network characteristics and ranking



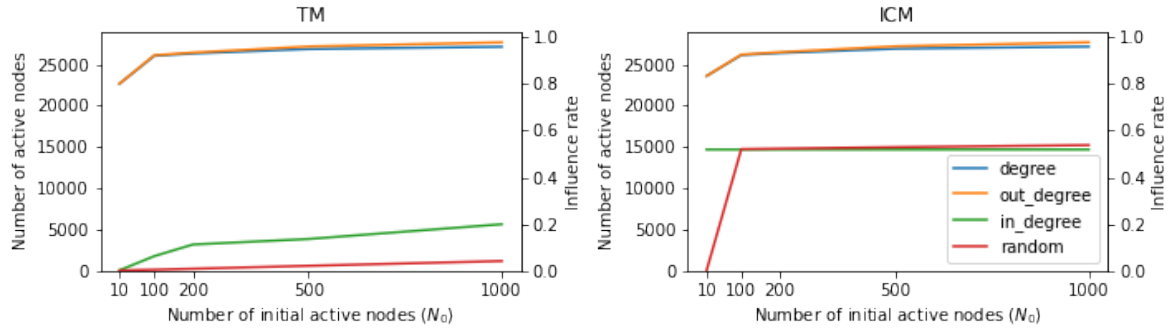
the first throughout this month. Also, due to his novel Chinese performance style, he caught significant attention from the international audiences. During this time period, the "CHUANG_official" account mentioned Liu Yu much more times than Mika. As the result, though Liu Yu started with a lower popularity in Twitter, he quickly gained favor as he kept ranking high. Looking in Mika's network during this period of time, though "CHUANG_official" does not promote him that much, but Mika's fan leaders have a stronger influence than Liu Yu's, indicated by the higher average degree in the top-10 out-degree nodes. In other words, though with less discussions, Mika's popularity might not be lower than Liu Yu's.

After April, Mika's statistics got over Liu Yu again. At the beginning of April, there were social media attacks targeted toward Liu Yu, partially because he ranked at first for many weeks. The other potential reason is that, in Chinese social media, there were too many online fans commenting Liu Yu would always be the first and thus making Liu Yu less favorable to fans for all other contestants. As the results, the Liu Yu's ranking dropped, "CHUANG_official" started to promote less about Liu Yu and Liu Yu's statistics dropped behind Mika's.

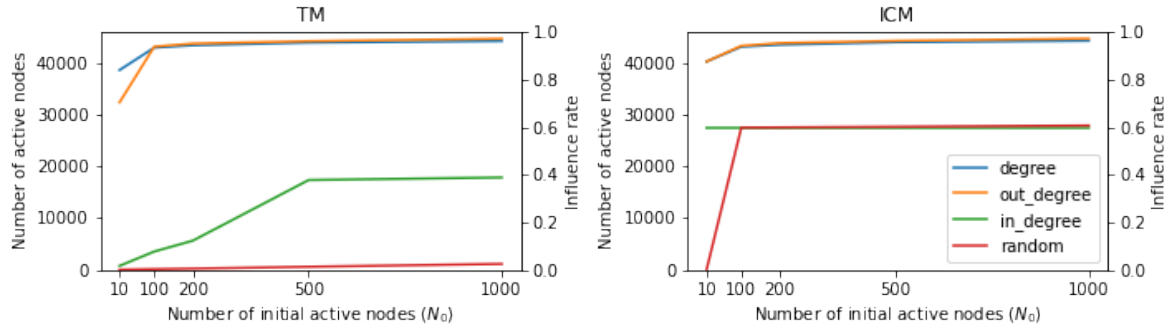
Besides all above temporal progress in network characteristics, the clustering coefficient of Mika was almost always higher than Liu Yu. This again suggests a higher popularity of Mika than Liu Yu. In terms of graph density, Liu Yu is always higher than Mika. This may simply means that Liu Yu's fans are less inter-connected than Mika's fans. This is reasonable as Liu Yu's presence in international stage only started from February. So Liu Yu's fans had less time for connection.

Figure 7: Simulation results with diffusion models

(a) Liu Yu



(b) Mika



4.3 Diffusion models evaluation

The experiments using Threshold Model (TM)[6] and Independent Cascade Model (ICM)[7] are done with 4 types of degree initialization and 5 numbers of initial initial actove nodes (N_0). Results are shown in Figure. 7. All experiments are done 5 times and the average results are reported. Since both Liu Yu's results and Mika's results are similar, thus I will discuss them together in the following discussion.

4.3.1 Threshold Model (TM)

For TM, initializing using out-degree gives the highest influence rate for all N_0. When N_0 equals to 1000, both Liu Yu's network and Mika's network reached 95% influence rate. Initializing with overall-degree has a similar behavior as the out-degree.

For both degree and out-degree, there was no significant increase in the additional number of active nodes when increasing N_0 from 100 to 1000. Also, when N_0 is as low as 10, initializing with top degree can still have an influence rate of 79% for Liu Yu and 84% for Mika. When N_0 is 10 and initializing with out-degree, the influence rate is 79% for Liu Yu and 68% for Mika.

Initializing with in-degree has a lower influence rate comparing to overall-degree and out-degree. This is because the nodes with high in-degree does not means that they have

high out-degree for influencing other nodes. Some of the top-5 in-degree nodes even have 0 out-degree. For Liu Yu, when N_0 equals to 1000, the final number of active nodes is only 5618, which is 20% influence rate. Unlike overall-degree and out-degree, the number of active nodes only becomes steady when N_0 >= 200.

Initializing with random has the lowest influence rate, as expected. For Liu Yu, when N_0 equals to 1000, the final number of active nodes is only 1149, which is 4% influence rate. For Mika, when N_0 equals to 1000, the influence rate is also low at 2%.

4.3.2 Independent Cascade Model (ICM)

For Independent Cascade Model, using out-degree has the highest influence rate, closely followed by using degree. Similar to TM, the increase in number of active nodes is small when N_0 increases from 100 to 1000. Unlike TM, the random policy has a slightly higher influence rate than the in-degree policy when N_0 >= 100. When N_0 is 10, Liu Yu's influence rate is 52% and Mika's influence rate is 59%. When N_0 is 1000, Liu Yu's influence rate increased only 2% to 54% and Mika's influence rate increased 1% to 60%.

Generally both models, TM and ICM simulated with close to 100% diffusion rate when using more than 100 initial nodes selected based on either overall degree or out-degree.

5 Discussion and Conclusion

In this project, I analyzed the Twitter network characteristics for two new digital influencers, Liu Yu and Mika from CHUANG2021. In general, Mika's network has "better" network characteristics, which indicate for higher popularity than Liu Yu. The time series in network characteristics and competition ranking, suggests that the change in network characteristics do relate to the change in popularity.

From diffusion model simulation, I found that using degree and out-degree for choosing initial active nodes performs well for both Threshold Model and Independent Cascade Model. Also, as the results suggest, agencies for Liu Yu and Mika can focus on just 10 highly influential fan leaders to achieve high influence rate in their social media networks.

In terms of limitation, this project only studied Twitter but not Weibo. Since CHUANG2021 is a Chinese show, popularity on Weibo, a Chinese social media platform, should have a higher correlation with the competition ranking and contestants' popularity. Also, the scrapped dataset is not a complete set of Tweets as I don't have premier Twitter API access. Last but not least, for advertisement agencies, in order to have accurate quantitative estimation on digital influencers' commercial values, a larger amount of digital influencers should be studied.

6 Acknowledgements

This project partially builds based on the data from Twitter's official API.

REFERENCES

[1] Wikipedia Contributors. Produce camp 2021, 03 2021.

[2] Francesco Pierri, Carlo Piccardi, and Stefano Ceri. Topology comparison of twitter diffusion networks effectively reveals misleading information. *Scientific Reports*, 10, 01 2020.

[3] Samir Akrouf, Laifa Meriem, Belayadi Yahia, and Mouhoub Eddine. Social network analysis and information propagation: A case study using flickr and youtube networks - volume 2 number 3 (jun. 2013) - ijfcc, 2013.

[4] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE*, 9:e98679, 06 2014.

[5] Derek Hansen. Reingold layout - an overview | sciencedirect topics, 2015.

[6] Mark Granovetter. Threshold models of collective behavior | american journal of sociology: Vol 83, no 6, 2020.

[7] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network | proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining, 2003.