TikTok's Users Analysis in the Context of US Politics

Advanced Topics in Computer and Networks Security

Author: Marco Bellò
Author: Alberto Morini
Supervisor: Alessandro Galeazzi

18/06/2024

Abstract

Introduction

TikTok is one of the most prominent social networks currently available, boasting over 100 million users in the USA alone [14]. Approximately 60% of young adults (aged 18-24) and nearly all children (aged 5-15) use TikTok daily [20]. This significant influence attracts not only advertisers but also politicians: the Democratic Party began organizing paid influencers as early as the 2020 United States election (for instance, presidential candidate Michael Bloomberg engaged in paid partnerships on social media [10]), while Republican and Conservative hype houses campaigned on behalf of political candidates. In Germany, the political party CSU invited influencers to political events and has recently started creating influencer-like social media posts on platforms such as TikTok [10].

This led to the creation of influencer-driven marketing firms, which now claim to control vast, immediately-deployable stables of small-scale influencers for various campaigns. These "nano" and "micro" influencers differ from the conventional image of an influencer: they are everyday people with captive, intimate social media audiences who represent demographics particularly appealing to U.S. political campaigns, such as Latinos in South Florida, Black voters in Atlanta, and college-educated women in the Rust Belt [12]. Political influencers often do not have an institutional background, in fact most of the times their notoriety and fame is platform-built [11].

Despite this political engagement, TikTok has attempted to market itself as a platform for everything but politics: since 2019 the company has banned paid political advertising, stating that "the nature of paid political ads is not something we believe fits the TikTok platform experience." Nevertheless, many creators regularly use the platform to disseminate political messages and viewpoints without disclosing whether the content is sponsored or not [3, 13].

Given these dynamics, there is considerable value in studying user interactions on the platform. This work aims to do so, focusing on the context of the 2024 U.S. political election. The study will examine user movements using a social graph, analyze user similarity through cosine similarity, infer political affinity, measure engagement, and visualize the impact of publishing a video on followers and comments.

1 Data Gathering

Considering the context (US elections) it is imperative to have users' data divided between left and right-leaning, so a group of *super-users* was selected using various sources. *Super-users* are defined as follows:

- Influencers: people whose notoriety is platform-built, without a background in institutions of entertainment;
- Politicians;
- Newspapers or news sites (i.e., The Washington Post).

The complete followers list, with relative sources (missing if selected arbitrarily by the authors), is as follows:

- Left-Wing:
  - @aocinthehouse [15, 19]
  - @bernie [15, 19]
  - @chrisdmowrey [7]
  - @cnn [8, 2]
  - @democracynow.org [unofficial]
  - @genzforchange [9]
  - @ginadivittorio [6]
  - @harryjsisson [7]
  - @huffpost [8, 2]
  - @msnbc [8, 2]
  - @newyorker [8, 2]
  - @nytimes [8, 2]
  - @repbowman [13]
  - @rynnstar [20]
  - @teamkennedy2024 [15, 19]
  - @thedailybeast [8, 2]
  - @underthedesknews [7]
  - @vox [8, 2]

  - @washingtonpost [8, 2]

- Right-Wing:
  - @alynicolee1126 [3]
  - @babylonbee [4]
  - @clarksonlawson [7]
  - @dailymail [8, 2]
  - @dailywire [8, 2]
  - @itsthemandrew [3]
  - @notvictornieves [7]
  - @real.benshapiro [11]
  - @studentsforlife [4]
  - @theisabelbrown [3]
  - @thesun [8, 2]

Due to time and computational constraints, only five super-users had been selected from each group: *@alynicolee1126, @babylonbee, @real.benshapiro, @clarksonlawson, @notvictornieves* for right-leaning, and *@thedailybeast, @huffpost, @aocinthehouse, @repbowman, @newyorker* for left-leaning super-users.
All data has been gathered using TikTok's official APIs (https://developers.tiktok.com/doc/overview/)

1.1 TikTok's APIs

TikTok's APIs requires prior authentication using a Secret Key and a Client ID, both of which can be obtained by making a personal request to the platform's staff. Then, each call to the API must be authenticated with a Bearer token, previously obtained through the appropriate authentication endpoint. Each one has its own query string and body parameters to be included in the HTTP request.

There is a daily limit of 100,000 records (which resets at 12 AM UTC) for videos and comments while, for the followers/following endpoint, the limit is set up to 2 million records (https://developers.tiktok.com/doc/research-api-faq/).

In this project, every call is parameterized to retrieve the maximum allowed data, typically 100 records. However, the APIs do not always provide the exact data requested, possibly due to a lack of content or other unknown issues.

Download Component

A wrapper for the APIs has been realized for data gathering, which simply makes WebAPI calls and stores the results. Almost every public endpoint provided has been totally covered by the script, specifically: users' followers, users' videos, videos' comments, following users, liked videos, and users' information.

In the first step, the script authenticates to TikTok's endpoint, gaining the token, which will be refreshed a few minutes before its 2-hour lifespan. Then it starts retrieving the data batch required, storing each response in a JSON file for later analysis.

For the specified number of videos, in this case 100 per month for each super-user (see second code block of section 4), the script will download their metadata and store them, before concluding its job by retrieving super-users' followers.

```
# @query {JSON} the video query as specified in the TikTok
    docs
# @nrVideo {int} the number of videos which we want to
    download
# @startDate {string} in Unix format
# @endDate {string} in Unix format -- NB: can't be greater
    than a month
# @filename {string} of JSON where data will be stored
def processVideo(query, nrVideo, nrComments, startDate,
    endDate, filename):
    ....

processVideo(videoQuery,100,200,'20240301','20240330','
    influencer-month')
```

Amount of Data Downloaded

For each video, followers have been downloaded (100 per request) for up to 5 days, within 3-hour intervals, allowing for the retrieval of a theoretical number of followers equal to 4,000:

$$(100 \text{ users} \times ((24\text{h}/3\text{h}) \times 5 = 40 \text{ requests})) = 4000$$

However, such numbers are reached only if all new followers are distinct in every call, which is difficult to achieve within such a short time span.

Since TikTok's APIs return the user who has started following the influencer from the date (in Unix format) declared in the body of the request (called cursor), the problem of duplicated accounts arises.

To clarify: if *JohnDoe* follows *MrWhite* on 31/12/2023 at 10:00:00 and *MrWhite* does not gain 100 new followers in the next 3 hours, *JohnDoe* will also be included in the request at 31/12/2023 at 13:00:00.

To solve this problem, a Python script (*DataCleaner.py*) was created, which keeps only the first occurrence (sorted by ascending date) of each username found in the total downloaded data.

```
for i in range(0,len(total)-1):
for j in range(i+1, len(total)):
    if(total[i].get("influencer")==total[j].get("
        influencer")): ##check if the same influencer (we
        don't want to remove common followers)
        total[i]["followerList"] = [elem for elem in total
            [i].get("followerList") if elem not in total[j
            ].get("followerList")]
```

Additionally, video metadata (such as views, likes, number of comments, etc) and super-users' public information (with a single call for each one) are stored for a later analysis.

Ultimately, the total amount of data downloaded, divided among 10 super-users, is as follows:

- 35,798 distinct followers;
- 182 videos.

2 Privacy and Echo Chambers

To help visualize all the gathered data, identify the presence of echo chambers, and infer the political views of users, super-users' followers were used to create a social network graph and a cosine similarity matrix.

2.1 Data Preparation

First, a list of super-users' followers was gathered via TikTok's APIs in the form of a JSON file, and then processed with a script to get the following structure (we can ignore `"videoID"` and `"videoDate"`):

```
[
    {
        "influencer": "ith-super-user Name",
        "videoID": "videoID",
        "videoDate": "videoDate",
        "followerList": [
            "follower1",
            "follower2",
            "follower3",
            "follower4",
            "follower5",
            "follower6",
            "follower-k"
        ]
    }
]
```


To clarify, here follows a partition of the real JSON data used:


```
    {
        "influencer": "huffpost",
        "videoID": "7354208741996186911",
        "videoDate": "2024-04-05 11:46:08",
        "followerList": [
            "mathieucambet",
            "raphclp",
            "jennet153"
        ]
    },
    {
        "influencer": "huffpost",
        "videoID": "7354208741996186911",
        "videoDate": "2024-04-05 20:46:08",
        "followerList": [
            "doodlegolden0",
            "evanroyalaug",
            "cshanebritt",
            "kabed70"
        ]
    },
```


JSON data is then imported to *Social_Graph.r* for analysis:

```
data <- fromJSON(paste(readLines("data.json")))

left_influencer_names <-  # vector of strings with left
                          # super-user names
right_influencer_names <- # vector of strings with right
                          # super-user names

# data.frame used to calculate all the graphs and tables
full_total <- data.frame(
  influencer = data$influencer,
  followerList = I(data$followerList)
)

full_influencer_names <- union(left_influencer_names,
                               right_influencer_names)
```

Now we have three data structures to work with: two vectors with all super-users' names and a `data.frame` that stores all super-users and their gathered followers, as shown below:

influencer	followerList
alynicolee1126	character(0)
alynicolee1126	c("elsee30", "renetheriot171")
alynicolee1126	character(0)
alynicolee1126	markbutler5636
alynicolee1126	c("frazierbklima", "electricmaster57", "ravirquaddnos", "gabbertour", "darrenricks87", "kory_sprinkle")
alynicolee1126	character(0)
alynicolee1126	c("addie.thorp", "caliwren", "truth.be.told", "idgafboutya501", "rowdy_chav", "thesalvatoresandmore._", "monahuluta", "recoveringhedonist", "pureblood_k")
alynicolee1126	c("_christian_l11", "robertdonald98", "mesa11964", "red_tacoma_4.0", "sladethekoolaid")

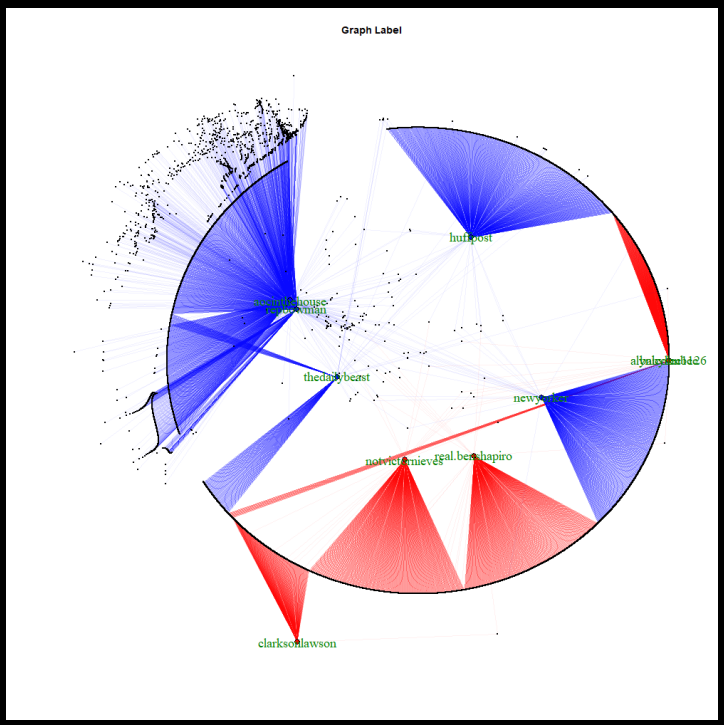
2.2 Social Graph

Broadly speaking, a social graph is a graph that represents social relations between entities, were vertices (or nodes) represents users and edges represents relations between such users. It is a model of representation of a social network, and has been referred to as "the global mapping of everybody and how they're related".

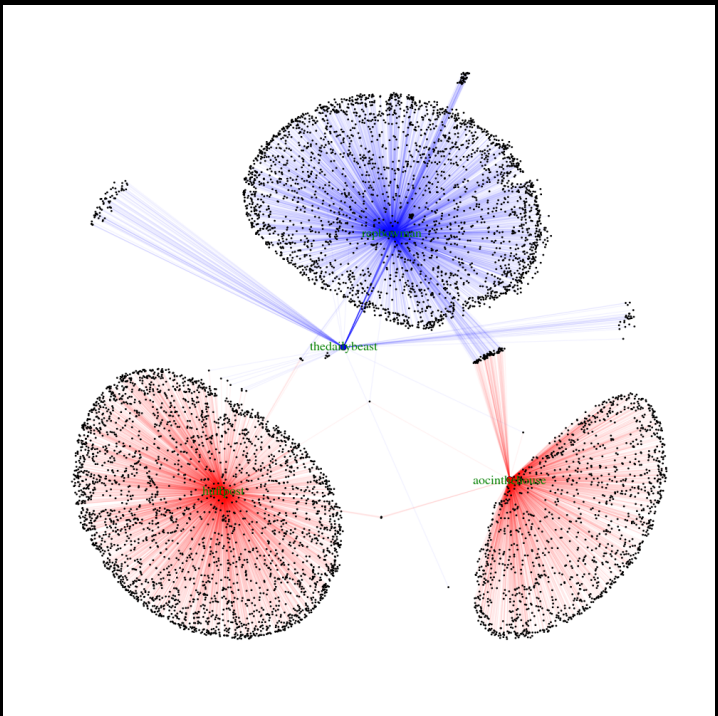
To give a brief example: if Alice and Bob are friends on a social network, in a social graph they would be represented each as a node, and there would be an edge between them.

The term was popularized at the Facebook F8 conference on May 24, 2007, when it was used to explain how the newly introduced Facebook Platform would take advantage of the relationships between individuals to offer a richer online experience [17].

Employing a social graph has numerous advantages: it helps visualize all gathered data (all users and their relations), identify the presence of echo chambers and provide insights to analyze the network as a whole.
The graph that follows clearly demonstrate that, considering the gathered data, users do not interact with each other outside their communities, thus forming cliques that can be interpreted as echo chambers:



Unfortunately the above graph could be difficult to understand due to the exceptional amount of data points, so here follows a mock-up graph made with a reduced dataset to help clarify how it works:



Super-users can be easily identified: their nodes are larger, labeled and, most notably, they are at the center of their respective sub-graphs.

All black vertices represents followers of the super-users, unlabeled for improved readability, and the edge color represent the political orientation of the super-user they are connected to(red for right-leaning and blue for left-leaning super-users).

To clarify: if *@user-Alice* follows super-user *@aocinthehouse* (official account of Congress member Alexandria Ocasio-Cortez), which is classified as a left-leaning super-user, the edge connecting them will be blue.

Conversely, if *@user-Bob* follows super-user *@thesun* (official account of UK's tabloid "The Sun"), which is classified as a right-leaning super-user, the edge connecting them will be red.

2.3 Cosine Similarity

Having a graphical representation of a network is valuable: images are easier to recognize, process, and recall. When words enter long-term memory, they do so with a single code. Pictures, on the other hand, contain two codes: one visual and one verbal, each stored in different brain regions (Paivio). The dual-coding nature of images allows two independent ways of accessing visual memories, increasing the odds of remembering at least one. Adding illustrations to text aids comprehension and learning [5]. However, it is also advisable to measure the similarity between users numerically.

Broadly speaking, there are two types of similarity measures between nodes in a network: edge similarity, which provides the index of intersection of parent nodes, and global structure similarity, that aims to evaluate the similarity between two nodes in the context of the whole network. Regarding the latter, Salton Index, Jaccard Index, and Sorensen Index always have good performance, while cosine similarity's computational complexity is very high when applied to large volumes of data [1]. When the data is dense, structure-based indices like Salton's can perform as well as the cosine index but with lower computational complexity. Furthermore, when the data is sparse, structure-based indices outperform the cosine index [21].

N, q	CN	Sal	Jac	Sor	AA	RA	Cos	PCC
N=10, q=0.2	48	45	46	46	47	45	44	46
N=10, q=0.4	108	99	99	100	107	102	97	99
N=10, q=0.6	169	153	153	153	169	163	152	147
N=10, q=0.8	227	202	201	201	226	220	200	194
N=20, q=0.2	54	50	50	50	53	51	48	54
N=20, q=0.4	119	110	109	110	118	114	108	112
N=20, q=0.6	185	167	167	167	184	179	166	166
N=20, q=0.8	245	216	215	215	244	238	215	217
N=50, q=0.2	62	58	58	58	61	58	58	63
N=50, q=0.4	131	123	123	123	131	127	122	128
N=50, q=0.6	200	182	181	181	199	195	181	186
N=50, q=0.8	261	232	231	231	259	254	231	243

The above table shows values regarding precision in inferring similarity between users: Salton index (*Sal* column) seems to perform the best [21], therefore it has been used in this work to measure similarity between super-users.

Index formula is as follows:

$$s_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_y}}.$$

All index values are calculated for each couple of super-users and shown in the following table where, as mentioned before, blue represent left-leaning and red represent right-leaning super users:

	thedailybeast	huffpost	aocinthehouse	repbowman	newyorker	alynicolee1126	babylonbee	real.benshapiro	clarksonlawson	notvictornieves
thedailybeast	1.0000000000	0.0088749628	0.003280670	0.0030531062	0.0028821675	0.003546902	0.0023624420	0.0008080255	0.0000000000	0.0007953238
huffpost	0.0088749628	1.0000000000	0.002755189	0.0019942806	0.0023532799	0.001654876	0.0003674144	0.0007540000	0.0000000000	0.0007421476
aocinthehouse	0.0032806704	0.0027551892	1.000000000	0.0728348570	0.0038275628	0.001712851	0.0003802860	0.0019510368	0.0000000000	0.0007681470
repbowman	0.0030531062	0.0019942806	0.072834857	1.0000000000	0.0019429436	0.001594039	0.0003539074	0.0010894217	0.0000000000	0.0007148644
newyorker	0.0028821675	0.0023532799	0.003827563	0.0019429436	1.0000000000	0.001880989	0.0008352317	0.0008570222	0.0000000000	0.0004217752
alynicolee1126	0.0035469022	0.0016548763	0.001712851	0.0015940390	0.0018809890	1.000000000	0.0020557341	0.0021093667	0.0000000000	0.0020762086
babylonbee	0.0023624420	0.0003674144	0.000380286	0.0003539074	0.0008352317	0.002055734	1.0000000000	0.0014049602	0.0000000000	0.0036876667
real.benshapiro	0.0008080255	0.0007540000	0.001951037	0.0010894217	0.0008570222	0.002109367	0.0014049602	1.0000000000	0.0006124765	0.0052028285
clarksonlawson	0.0000000000	0.0000000000	0.000000000	0.0000000000	0.0000000000	0.000000000	0.0000000000	0.0006124765	1.0000000000	0.0036170919
notvictornieves	0.0007953238	0.0007421476	0.000768147	0.0007148644	0.0004217752	0.002076209	0.0036876667	0.0052028285	0.0036170919	1.0000000000

Salton index values range between 0 and 1, with the diagonal of the matrix showing all values equal to 1 because a super-user is always identical to itself. The table confirms numerically what could be seen in the social graph: super-users share very few followers, which means that each community is an echo chamber.

2.4 Privacy Inference

Another way to leverage social network information, specifically the network structure (edges and their respective nodes), is to infer users' attributes. This falls under classification problems (identifying which of a set of categories an observation belongs to, such as classifying an email as "spam" or "not spam") [18], and such techniques have been employed in this context (e.g., k-nearest neighbors classification [16]).

The objective here is to infer the political orientation of users based on the super-users they follow. Given the scarcity of super-users, the presence of only two sets (left and right-leaning), and the high degree of isolation between communities, complex statistical tools are deemed unnecessary. The idea is straightforward: if *user-Alice* follows mainly left-leaning super-users (e.g., *aocinthehouse*, *bernie*, *repbowman*), that user is classified as left-leaning. Conversely, if *user-Bob*

follows mainly right-leaning super-users, he is classified as right-leaning.

The following table shows users following more than one super-user from both lists of left and right-leaning super-users, with results ordered by the number of super-users followed.

User	Influencers
mdmcgraw1	huffpost, thedailybeast, newyorker
tukutanelimuru3	huffpost, aocinthehouse, repbowman
maazch_official	aocinthehouse, repbowman, newyorker
britanny_leeann	repbowman, thedailybeast, aocinthehouse
benjamin__bitboy	repbowman, thedailybeast, newyorker
audicharon	repbowman, notvictornieves, real.benshapiro
sgib527	babylonbee, real.benshapiro
chellen8er3.0	babylonbee, real.benshapiro
canyouhearmewaving	babylonbee, thedailybeast

As we can see, only six users follows more than three super-users and, generally speaking, most users follows super-users of the same political affiliation. Considering the scarcity of data, if a user follows less than 100% of super-users from the same political spectrum, it is difficult to say anything about their political beliefs.

Let's make some examples: taken into consideration the table shown above (which is a small sample of the original, available a the following link: https://github.com/albertomorini/CNS/blob/main/privacy_table.html) we can say the following:

- User *mdmcgraw1* follows three super-users, all left-leaning, therefore is classified as left-leaning;
- User *audicharon* follows three super-users, one left-leaning, two right-leaning, therefore nothing can be said about its political orientation;
- User *chellen8er3.0* follows two super-users, both right-leaning, therefore is classified as right-leaning.

In future extensions, this classification could be improved by giving weights to super-users: *repbowman* could weight more than *notvictornieves* because the latter is an influencer, while the former is a politician.

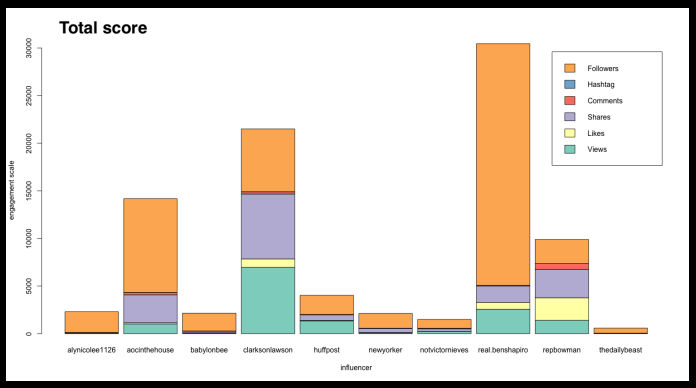
3 Engagement

Studying engagement on TikTok can reveal how many people are reached and the approval rate of content.

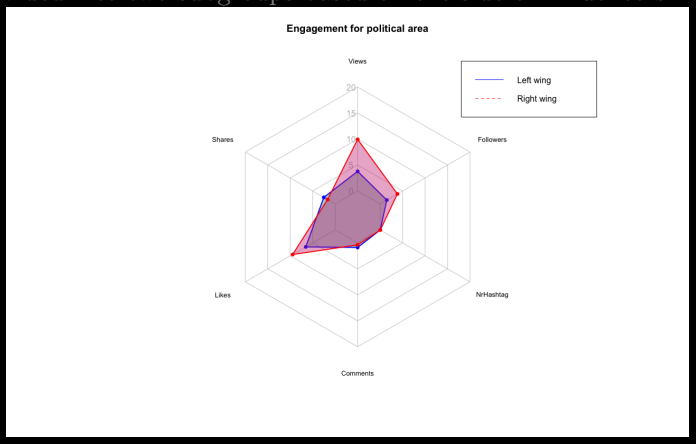
Each post (in this case, only videos) includes several pieces of information that can be used to analyze the content's impact. For example, the API returns data such as the number of views, comments, likes, shares, and more.

Naturally, views have the highest numbers, followed by likes and comments. For these video topics, there is a very low "repost" rate (called shares), which refers to users reposting the influencer's video.

In the analysis, some attributes has been normalized thus to obtain a graph well formed. In detail: views, likes, comments are reported in thousands



To compare the two political wings, the data has been divided into two subgroups based on the side of influencers.



4 Content Impact

One aim of this research is to analyze the impact posting new content has on super-users' followers counts.

To this end, all videos posted by a super-user were downloaded over the span of five months. For each video, all new followers within five days from the upload date were registered.


```
{
    "data": {
        "cursor": 1717588518,
        "has_more": true,
        "user_followers": [
            {
                "username": "rev6luv",
                "display_name": "t."
            },
            {
                "display_name": "Caden M. Flanagan",
                "username": "therealcadenflanagan"
            },
            ...
    }
}
```

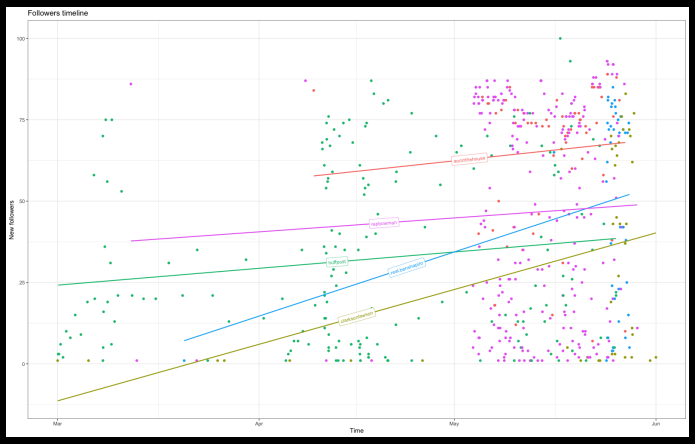

There is a potential bias in this approach, as new followers can be gained independently of posted content. However, in the current social network landscape, content either goes viral almost immediately or not at all. For this reason, this approach has been considered valid.

The query included in the body of the request contains only the username, without filtering other parameters (such as hashtag, or region). This decision has been made since a super-user, while eclectic in the topics discussed, still belongs to a specific political orientation.


```
{
    "and": [
        {
            "operation": "IN",
            "field_name": "username",
```

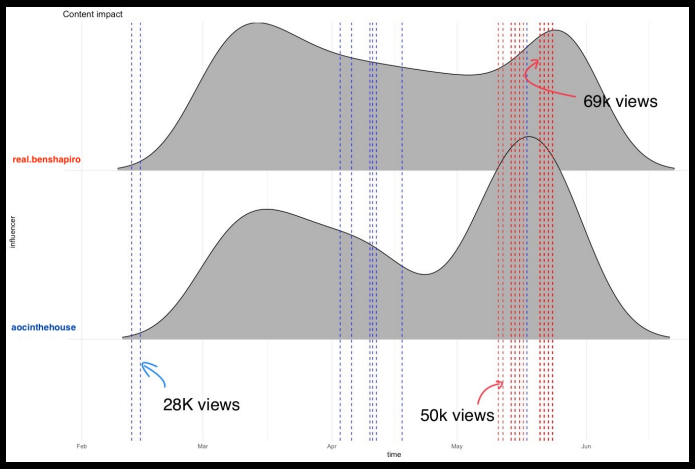
```
            "field_values": [
                "$influencer"
            ]
        }
    ]
}
```

In this analysis, only the most influential super-users have been considered (those with more followers and higher engagement values):



As can be seen, the total number of followers increases steadily over time for almost all super-users.

A more specific analysis has been done on two super-users of particular interest: *@aocinthehouse* and *@real.benshapiro*:



This analysis consisted of the average trend of new followers, creating a detailed curve enriched with timestamps of new post creations.

In the graph shown above, there is a noticeable increase in new followers for *@aocinthehouse* after the content posted on 15 February was viewed 28,000 times.

The aforementioned bias is evident in the increase in *@real.benshapiro*'s follower count in mid-March despite not being active on TikTok. It can be inferred that influence on followers' count extends beyond TikTok, supported by the release of a YouTube video on 9 March 2024, which received nearly 300,000 views.

5 Conclusions

While gathering data through TikTok's official APIs has proven difficult, the amount of total information ultimately obtained was satisfying: 35,798 distinct followers and 182 videos were downloaded among 10 selected super-users.

This dataset was used to visualize the network and identify the presence of echo chambers, numerically confirmed by the low cosine similarity between super-users, measure users' engagement, revealing higher participation by right-leaning accounts, and to demonstrate the impact a posted video has on both engagement and the number of followers.

Finally the information was utilized to develop a prototype for inferring political orientation, although the high degree of isolation mentioned before made it difficult to draw clear conclusions.

6 Extensions

Regarding echo chambers, privacy inference, engagement, and content impact, better results could be achieved by significantly increasing the number of gathered data points. For example, one could circumvent the limitations imposed by TikTok's APIs by using third-party alternatives such as https://github.com/davidteather/TikTok-Api.

Another approach to enhance the social graph could involve collecting not only the super-users' followers but also the accounts they follow. This strategy, while powerful, would exponentially increase the number of data points, potentially making the processing phase rather computationally onerous.

Incorporating users' comments into the dataset would allow for sentiment analysis using large language models (LLMs), facilitating the study of polarization. These observations could be intersected with prior findings. Additionally, utilizing language recognition tools and analyzing users' pinned, shared, and liked videos (obtainable through official APIs) could help infer geo-location information, which would be valuable in our geopolitical context.

References

[1] A. L. R. Ahmad Rawashdeh. Similarity measure for social networks – a brief survey. In J. H. K. Michael Glass, editor, MAICS 2015 Modern AI and Cognitive Science Conference, volume 1353, pages 153–159, 2015.

[2] allsides. Media bias ratings, 2024.

[3] M. F. Becca Ricks, Brandi Geurkink. These are not political ads: How partisan influencers are evading tiktok's weak political ad policies, 2021.

[4] F. N. Brian Flood. Tiktok has silenced 11 pro-free speech organizations while 'muzzling conservatives,' study finds, 2022.

[5] P. Dewan. Words versus pictures: Leveraging the research on visual communication. Partnership: The Canadian Journal of Library and Information Practice and Research, 10, 06 2015.

[6] S. Lai. Campaigns pay influencers to carry their messages, skirting political ad rules, 2022.

[7] F. N. Matteo Cina. Conservative tiktok influencers claim platform discriminates against right-leaning voices, 2023.

[8] mediabiasfactcheck. Media bias fact check, 2024.

[9] A. Nafees. Political tiktok, explained, 2024.

[10] M. Riedl, C. Schwemmer, S. Ziewiecki, and L. M. Ross. The rise of political influencers—perspectives on a trend towards meaningful content. Frontiers in Communication, 6, 2021.

[11] M. J. Riedl, J. Lukito, and S. C. Woolley. Political influencers on social media: An introduction. Social Media + Society, 9(2):20563051231177938, 2023.

[12] J. L. Samuel Woolley, Martin Riedl. Tiktok has a political influencer problem targeted at gen z voters, 2023.

[13] N. Scola. Will politicians ban their best way of reaching young voters?, 2023.

[14] Statista. Countries with the largest tiktok audience as of april 2024, 2024.

[15] Wikipedia. 2024 democratic party presidential primaries, 2024.

[16] Wikipedia. k-nearest neighbors algorithm, 2024.

[17] Wikipedia. Social graph, 2024.

[18] Wikipedia. Statistical classification, 2024.

[19] Yougov. The most famous democrats (q1 2024), 2024.

[20] J. Zeng and D. V. Kaye. From content moderation to visibility moderation: A case study of platform governance on tiktok. Policy & Internet, 14(1):79–95, 2022.

[21] Q.-M. Zhang, M. S. Shang, W. Zeng, Y. Chen, and L. Lü. Empirical comparison of local structural similarity indices for collaborative-filtering-based recommender systems. Physics Procedia, 3:1887–1896, 08 2010.