



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 the following:

- Left-Wing:
 - @aocinthehouse [15, 19]
 - @bernie [15, 19]
 - @chrismowrey [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]

Only five of each groups had been selected to data gathering, due to time constraints (@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

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, with the following structure (we can ignore "videoID" and "videoDate"):

```
1 [
2   {
3     "influencer": "ith-super-user Name",
4     "videoID": "videoID",
5     "videoDate": "videoDate",
6     "followerList": [
7       "follower1",
8       "follower2",
9       "follower3",
10      "follower4",
11      "follower5",
12      "follower6",
13      "follower-k"
14    ]
15  }
16 ]
```

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

```
1 {
2   "influencer": "huffpost",
3   "videoID": "7354208741996186911",
4   "videoDate": "2024-04-05 11:46:08",
5   "followerList": [
6     "mathieucambet",
7     "raphclp",
8     "jennet153"
9   ]
10 },
11 {
12   "influencer": "huffpost",
13   "videoID": "7354208741996186911",
```

```
14   "videoDate": "2024-04-05 20:46:08",
15   "followerList": [
16     "doodlegolden0",
17     "evanroyalaug",
18     "cshanebritt",
19     "kabed70"
20   ]
21 },
```

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

```
1 data <- fromJSON(paste(readLines("data.json")))
2
3 left_influencer_names <- # vector of strings with left
4                           # super-user names
5 right_influencer_names <- # vector of strings with right
6                           # super-user names
7
8 # data.frame used to calculate all the graphs and tables
9 full_total <- data.frame(
10   influencer = data$influencer,
11   followerList = I(data$followerList)
12 )
13
14 full_influencer_names <- union(left_influencer_names,
15                               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", "caliwn", "truth.be.told", "idgaboutya501", "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, where 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.

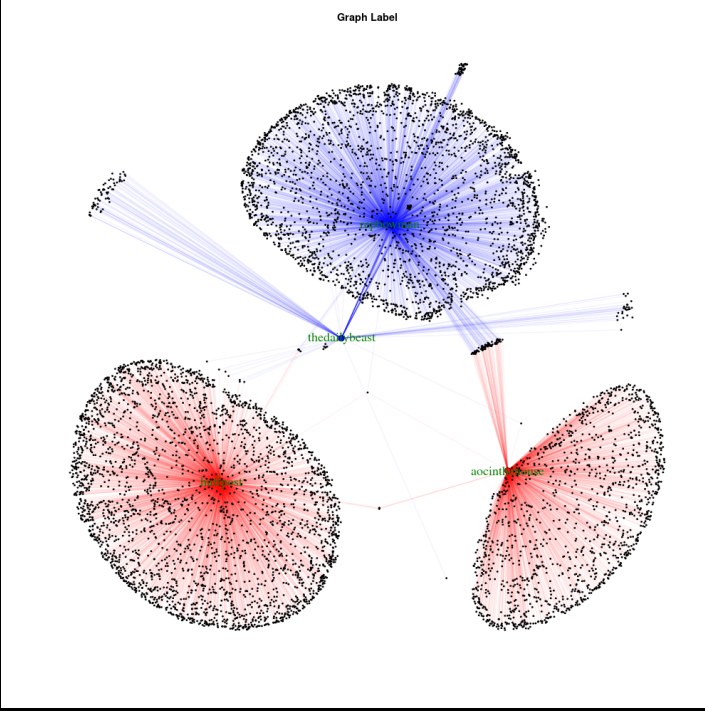
The term was popularized at the Facebook F8 conference on May 24, 2007, when it was used to explain how the newly introduced Facebook Platform would take advantage of the relationships between individuals to offer a richer online experience [17].



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:



The aforementioned graph is actually a mockUp without all the complete data: due to hardware constraints, graph compilation is impossible at the moment of writing (04/06/2024).

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 *@aocinthefirst* (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	aocinthefirst	repbowman	newyorker	alynicollee1126	babyonbee	real.benshapiro	clarksonlawson	noticvictories
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
aocinthefirst	0.0032806704	0.0027551892	1.0000000000	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
alynicollee1126	0.0035469022	0.0016548763	0.001712851	0.0015940390	0.0018809890	1.0000000000	0.0020557341	0.0021093667	0.0000000000	0.0020762086
babyonbee	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	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0006124765	0.0006124765	1.0000000000	0.0036170919
noticvictories	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 at 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

4 Content Impact

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.

