

TikTok Unwrapped Final Project Report

Oscar Chan & Angela Liu
Info 247, Spring 2022

Visualization: <https://ochan1.github.io/info247-sp22-tiktok-unwrapped/website/>

TikTok Unwrapped Final Project Report	1
Project Goals	4
Discussion of Related Work	5
Visualization Design Inspiration	5
InfraNodus	5
Digital Tracking Data Visualization	6
Event Log Visualizations	7
Narrative Inspiration	8
ProPublica's Breaking the Black Box	8
The Verge: GDPR makes it easier to get your data, but that doesn't mean you'll understand it	9
Investigation: How TikTok's Algorithm Figures Out Your Deepest Desires	10
How does TikTok use machine learning?	11
Description of Visualizations	12
Introduction	12
The Data Download	12
Data Download Interactive Tree	12
Data Scraping Process Infographic	14
Analyzing Oski's Data	15
Top Hashtags Tableau Dashboard	15
Visualizing the relationship between Categories of Hashtags and Ads	19
Advertising Trends over Time	20
Advertisement Cadence	21
Approach	24
Data Used	24
Tools Used	24
Data Scraping	24
Data Pre-Processing	25
Data Analysis / EDA	25
Visualizations	26
Website	26
Usability Testing Results	27
Method	27
Participants	28
Scenarios / Tasks	29
Pre-Study Questions	29
Study Tasks	30

Task 1: Overall Site	30
Task 2: Interact with the D3 Tree Viz	30
Task 3: Flow-Chart Visualization on Getting Video Metadata	30
Task 4: Interact with Hashtags Rank Chart Story Vizzes	31
Task 5: Interact with Hashtags-Ads Category Relationship Viz	31
Task 6: Interact with Ads Viz	31
Task 7: Interact with Gantt Chart Story on Ads Vizzes	32
Post-Study Survey	32
Results	33
Pre-Study	33
Study Tasks	33
Post-Survey	36
Usability Study Results Discussion	38
Data Download Tree Hierarchy Chart	38
Hashtag Chart Changes	38
% Ads and % Advertisers Line Chart	38
Ads-Hashtags Prototype	39
Links	40
Work Distribution	40

Project Goals

This project complements the work being done by the MIMS Capstone Project team, [Algorithms Unwrapped](#). Algorithms Unwrapped is a machine learning education tool to help social media users better understand their own algorithmic content. The project is focused on a single platform, TikTok, for the centrality of algorithm-generated content in its product.

Our information visualization project contributes to the overall Algorithms Unwrapped project by conducting a deep dive on the user data download made available by TikTok. Specifically, our final product is a website that introduces users to the TikTok User Data Download. The website walks users through how we identified advertisements and topics from the data and leverages a single user's data to produce case study visualizations of advertisement and content trends.

We focus our visualizations on the following overall goals:

- Describing the information available in the TikTok User Download
- Effectiveness of using one user's data as a case study to
 - Demonstrate how we identified ads and video content on the platform
 - Describe what ads and content TikTok served to this user
 - Describe trends and relationships in the ads and content served to this user
- Spark users' interest in accessing their user data and critically considering algorithmic activity

Discussion of Related Work

Related-work fell into two categories: visualization design inspiration and narrative inspiration from past data privacy/data download analyses and TikTok analyses.

Visualization Design Inspiration

1. InfraNodus

<https://infranodus.com/>



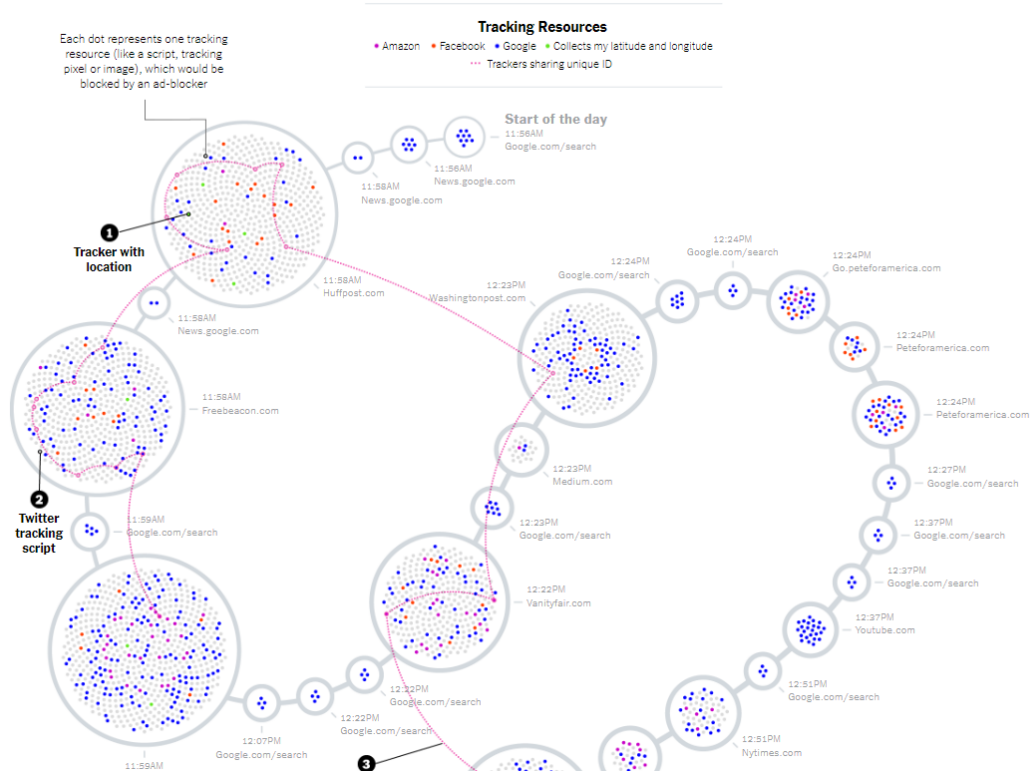
The complexity of this visualization would be a great way to view how one word can lead to or is related to other words. The visualization already has a way to take text data files and create a graph representing those words. This visualization tool also uses Machine Learning to provide a network analysis using OpenAI's GPT-3 network. However, we believed that having a visualization like this would be overwhelming.

We take inspiration from this visualization to see if we could find a way to create a relationship between hashtags to see if TikTok is learning and creating a profile for a user, or between hashtags and ads to see if there is some learning on what the user watches and then serve ads based on the learned interests.

2. Digital Tracking Data Visualization

www.visualcinnamon.com/portfolio/new-york-times-digital-trackers/

www.nytimes.com/interactive/2019/08/23/opinion/data-internet-privacy-tracking.html

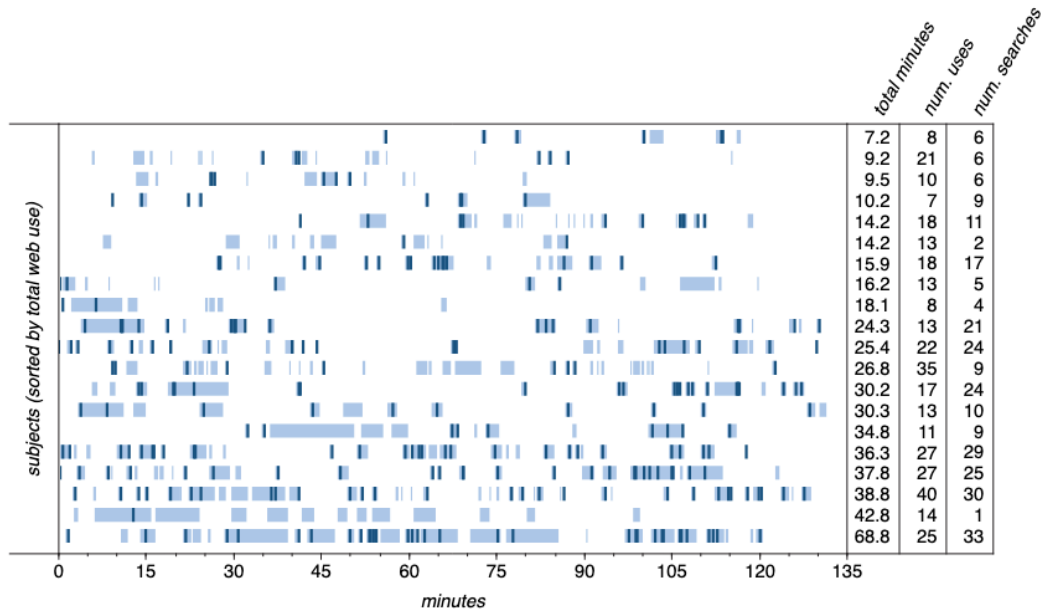


This visualization represents the number of trackers on each site over time using tracker cookies from Amazon, Facebook, Google, and Location Information. Identical trackers using the same unique ID are denoted by color codes.

This work inspired us to look to see if trends over time influenced the increased serving of certain hashtags or ads. While we are not able to use data from tracking cookies specifically, would the appearance or increasing trends of certain hashtags influence the appearance of new categories of ads.

3. Event Log Visualizations

https://hci.stanford.edu/publications/2009/webUseStudy/brandt_chi09_webuse.pdf



We also drew inspiration for our Tableau-based Gantt chart visualizations from this HCI study around event logging conducted by Brandt et al. (2009). In the chart below, Brandt et al. logged different activity types by research subject and by length of time. We modeled our advertiser chart after this: instead of research subjects, we had advertisers, and the time was represented in days. While this study looked at continuous instances of an activity, we instead logged discrete instances (i.e. each time an ad appeared).

Narrative Inspiration

4. ProPublica's Breaking the Black Box

www.propublica.org/article/breaking-the-black-box-what-facebook-knows-about-you

Liberal Party of Canada | Parti libéral du Canada
Sponsored · 14 hrs ·

It's going to take every one of us to ensure Justin Trudeau and the Liberal team will earn another mandate from Canadians in 2019 and continue building on the hope and hard work that brought us here.

Can we count on you?
CHIP IN ->

SECURE.LIBERAL.CA
Chip in now to help elect Team Trudeau candidates
Donate Now

WHAT IS COLLECTED

- 1 Text and links in the ad.
- 2 The picture in the ad.
- 3 Information Facebook provides about the ad's target audience.

- The time and date the ad was seen.
- The number of times the ad has been seen.
- The ad's language of origin

WHAT ISN'T COLLECTED
Anything else, including...

A classmate shared Angwin, Parris, & Mattu's excellent reporting on Facebook's algorithmic practices with us. This article discusses a tool ProPublica created to open up the "black box" by reflecting back to users the data collected by Facebook about them.

This article helped us get ideas about what pieces of data to highlight and reminded us to make obvious both explicit and implicit details in our data. For example, the fact that TikTok collects your browsing history isn't a surprise. What is enlightening is the specificity of that data (URL, date and time watched, whether you shared it), as well as the details that are conspicuously *not* included but factor into the recommendations [per TikTok's own account](#) – such as video interaction time, length of time watched, whether the video was skipped.

5. The Verge: GDPR makes it easier to get your data, but that doesn't mean you'll understand it

<https://www.theverge.com/2019/1/27/18195630/gdpr-right-of-access-data-download-facebook-google-amazon-apple>

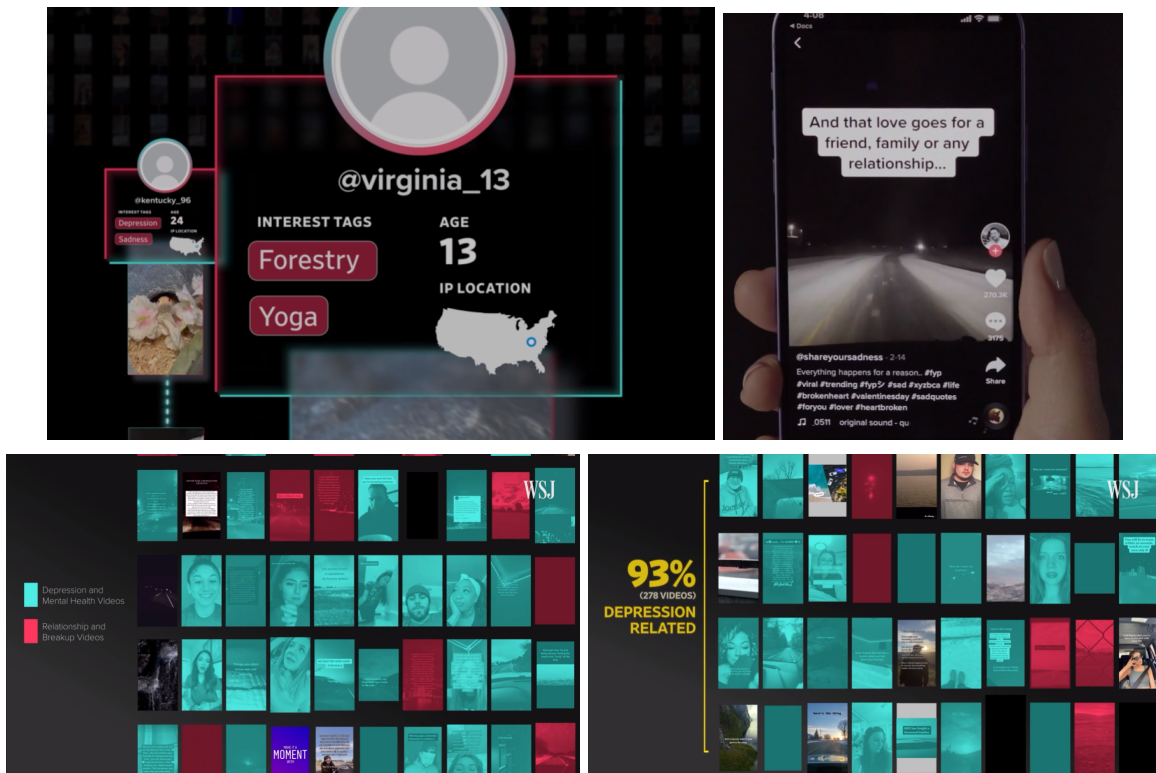
This Verge article provided the inspiration for our narrative. This reporter conducted a similar data download analysis (albeit from a textual, journalistic perspective) with popular platforms including Facebook, Google, Amazon, and Apple. We knew we wanted to explore the TikTok data download file but Porter's analysis reminded us of the broader policy landscape framing the existence of these data downloads. A quote from Porter's article resonated with our initial impressions after conducting exploratory data analysis on the TikTok dataset:

*"At the end of my experiment, I'm left with just under 138GB of data across the four services I contacted...After attempting to sift through and understand it all, it's clear that these companies, and the GDPR regulations that govern them, have a long way to go if they want to give us real control over our data. **Being able to download it is one thing, but making it useful means working harder to ensure that what's downloaded is easier for the average person to understand.**"*

- Jon Porter for the Verge (emphasis ours)

6. Investigation: How TikTok's Algorithm Figures Out Your Deepest Desires

<https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/>

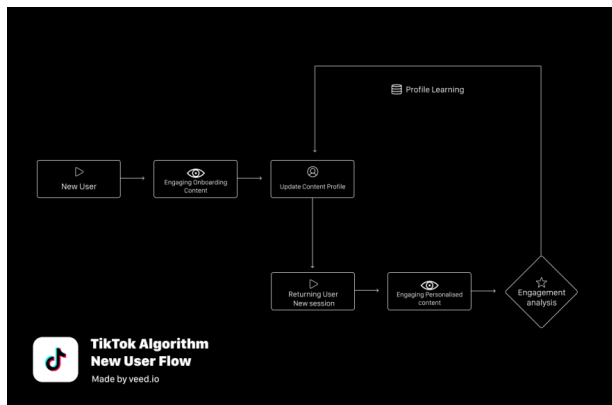


This video served as one of our biggest inspirations on the stressed need to visualize social media data. The Wall Street Journal created multiple bots based on its own tagged interests, but not initially supplied to TikTok, and TikTok was able to guess in a matter of hours the interests tagged by the bots. Of concern is a bot named kentucky_96, whose tags of focus were “Depression” and “Sadness”, whose content was a majority of depression-related content, 93% to be exact. Some of the videos obviously should have been filtered by some kind of moderator which could be considered dangerous. The Wall Street Journal also presented a visualization where the bots at first had very general videos of varying interests delivered to them and then diverges based on their interests in politics, dancing, sexual content, and depressing content.

We aim to see if we can find similar trends in the case study user’s data that might lead us to find and create a digital profile for the user. If such a digital profile can be found, we know TikTok is creating one for each user as we could see in the Wall Street Journal’s multiple bots. In the spirit of fairness and also innocent, until proven guilty, we also hope to see if the interviewed TikTok representative who mentioned that the algorithm tries to provide as much of a diverse set of videos as possible were true.

7. How does TikTok use machine learning?

https://dev.to/mage_ai/how-does-tiktok-use-machine-learning-5b7i



Computer vision

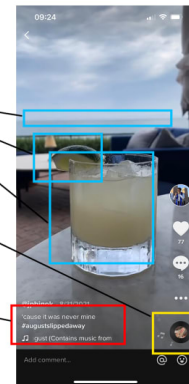
Ocean, lime, drink

NLP

Sound or additional audio

Metadata

Caption and hashtags



The article discusses some of the ways the TikTok algorithm uses videos the user interacts with to identify some of their interests. While we won't be using machine learning in our project, we took inspiration from the article to identify the features that we could quickly analyze. For starters, using a combination of the author, hashtags, and proper name helped identify advertisers. At the start, we wanted to watch each video and categorize them by hand to mimic this, but we had to narrow down our scope to only look at hashtags due to the sheer size of the data file.

The article also points out that interest in some videos could lead to an interest in other videos based on the viewing habits of a similar user. As a result, we went ahead to see if there was an increase in the same hashtags delivered to the user. We were only interested in a particular hashtag that appears in videos very often and not a few, per day.

Description of Visualizations

We divided our visualization website into four sections:

- Introduction to the website,
- Description of the TikTok Data Download,
- Analysis of the video history content to determine the algorithm's activity, and
- Next steps to explore one's own data

The last three sections roughly align with our overall project goals.

Introduction

This section sets up the narrative. While it doesn't include any visualizations, we discuss the background of why data downloads exist and provide the rationale for specifically exploring TikTok. Finally, we introduce Oski as a data subject persona. Oski's persona serves two purposes: (1) to preserve the privacy of the actual data subject while reminding readers that there is a real person generating the data, and (2) to emphasize that this is a case study exploration into one subject's data.

The Data Download

This section primarily addresses the first overall goal of our project – describing the information available in the TikTok User Download. The subgoals for this section included:

- 1) Teaching our audience how to retrieve a TikTok data download
- 2) Helping our audience understand the format of the raw data export from TikTok and explore its contents
- 3) Describing the additional steps we took to gather information of interest about the videos and advertisements.

For the first goal, we relied on text and video to demonstrate this background information.

Data Download Interactive Tree





For our second goal, we implemented a hierarchical visualization displaying the text of the raw data export file. We found the hierarchy appropriate to reflect the nested nature of both the JSON and zipped TXT formats.

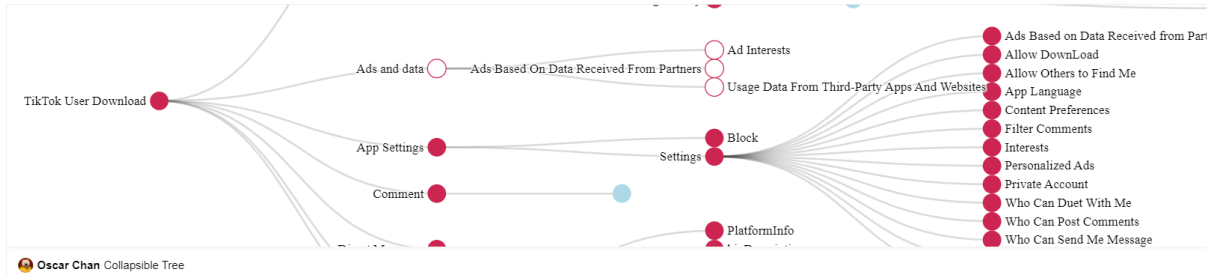
Nodes in the hierarchy represent dictionaries and dictionary entries in the JSON file. The "fill" of the node represents whether entries were available in that dictionary. Notably, in our file, the Advertisements dictionary was available but had no entries.

Interactivity also helped encode information we wanted to convey about the raw data download. Hovering over nodes that contained nested data would change the transparency of the node, encouraging users to click to uncollapse that node.

The experience of collapsing and uncollapsing the nodes helped replicate the experience of exploring the data export files without actually having to download and open a file.

Legend

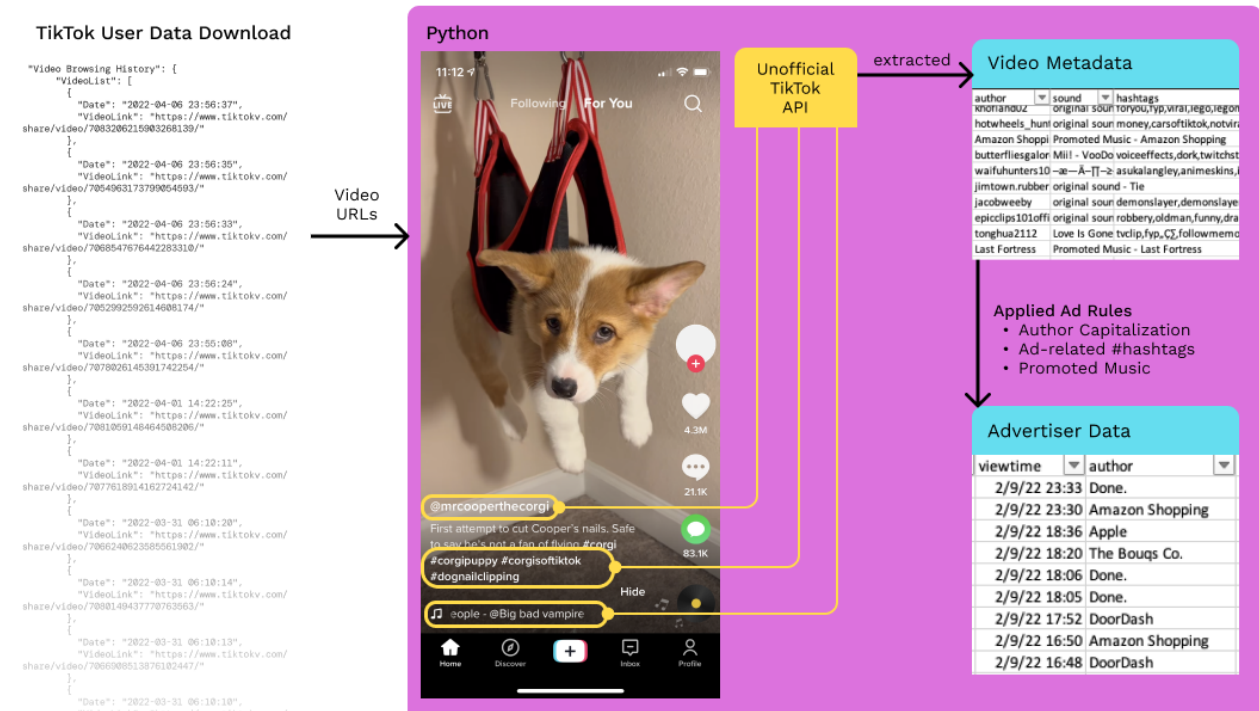
	Multiple Data, each entry share the same fields in child nodes
	Data Node (can represent a group of data or a value)
	Data Node like above, but specifically marked for the Ads data
	(Fading dots on mouseover) More sub-data available, click to reveal more of the data tree



One feature we attempted to have was to have a button to automatically expand a set of nodes based on the story we wanted to present. Unfortunately, we faced challenges in implementation so this remained more of an exploratory than explanatory visualization. In lieu of that, since there were only two sets of nodes to expand, the storytelling below the visualization does point out what nodes to expand in order to prove the two points we wanted to mention: no advertisement data and having the user's viewing history.

Data Scraping Process Infographic

Finally, for goal 3, we summarize the data scraping process we underwent to retrieve additional metadata about videos and advertisers. We present this process as an infographic. The use of imagery in the infographic – the screenshot of the initial JSON file and TikTok video layout – is meant to visually engage the reader and make the technical details more concrete for a lay, non-technical audience.



Analyzing Oski's Data

In this section, we dive deeper into the detailed analysis of Oski's data to address these components of our overall project goals:

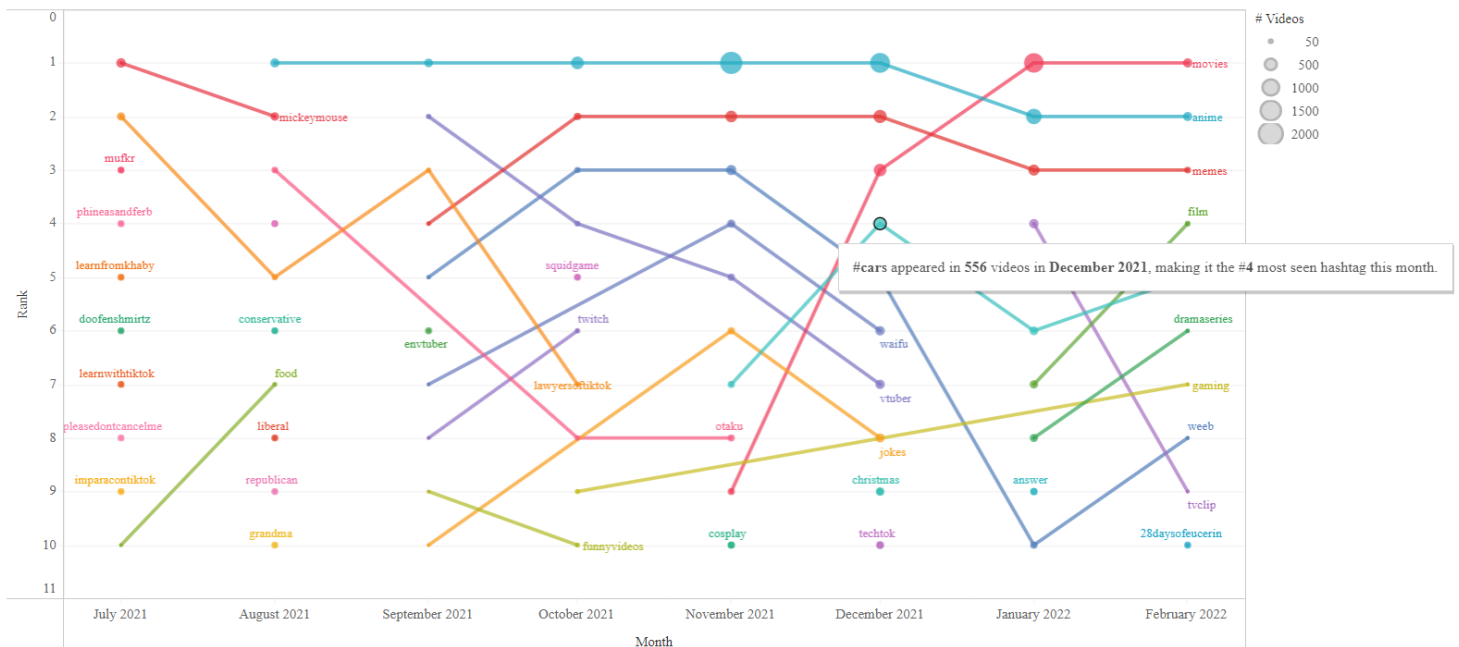
- Describe what ads and content TikTok served to this user
- Describe trends and relationships in the ads and content served to this user

Top Hashtags Tableau Dashboard

For content analysis, we decided to use hashtags as a proxy for categories. We realize this is an imperfect method: many videos lack hashtags altogether; further, through early manual categorization of hashtags, we noted that some tags don't reflect the true content of the video. However, since the majority (~80%) of videos did contain hashtags and absent other practical options for content analysis, we decided to move ahead with leveraging tags for content analysis while acknowledging these constraints.

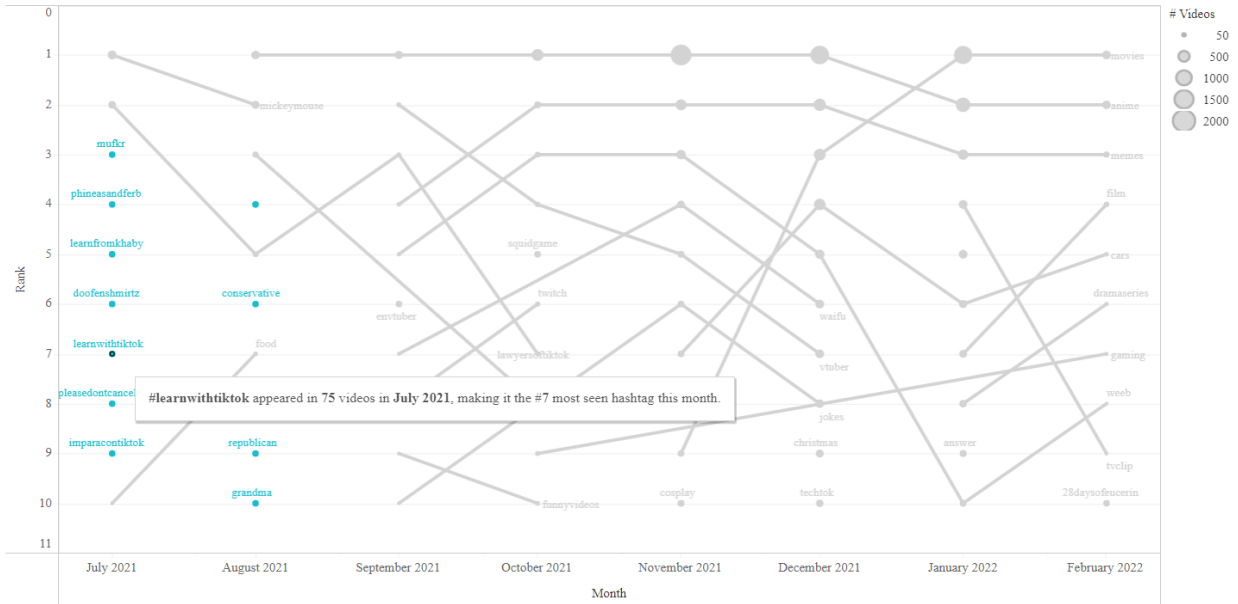
From over 40,000 unique hashtags, we focused on the top 10 monthly hashtags by frequency and plotted them in a rank chart. The rank chart shows the top 10 hashtags each month in Oski's feed; if a hashtag remains in the top 10 in consecutive months, a line connects those months for that hashtag. The size of each point denotes the number of videos the hashtag appeared in each month. The rank chart replaced the somewhat confusing and misleading hashtag gantt chart (see usability study) we had originally created.

The rank chart contains several different views that the user can click through. The views highlight and lowlight different hashtags based on the trends we were emphasizing. The initial view (below) is intentionally a bit chaotic, echoing the dynamic nature of the hashtags.



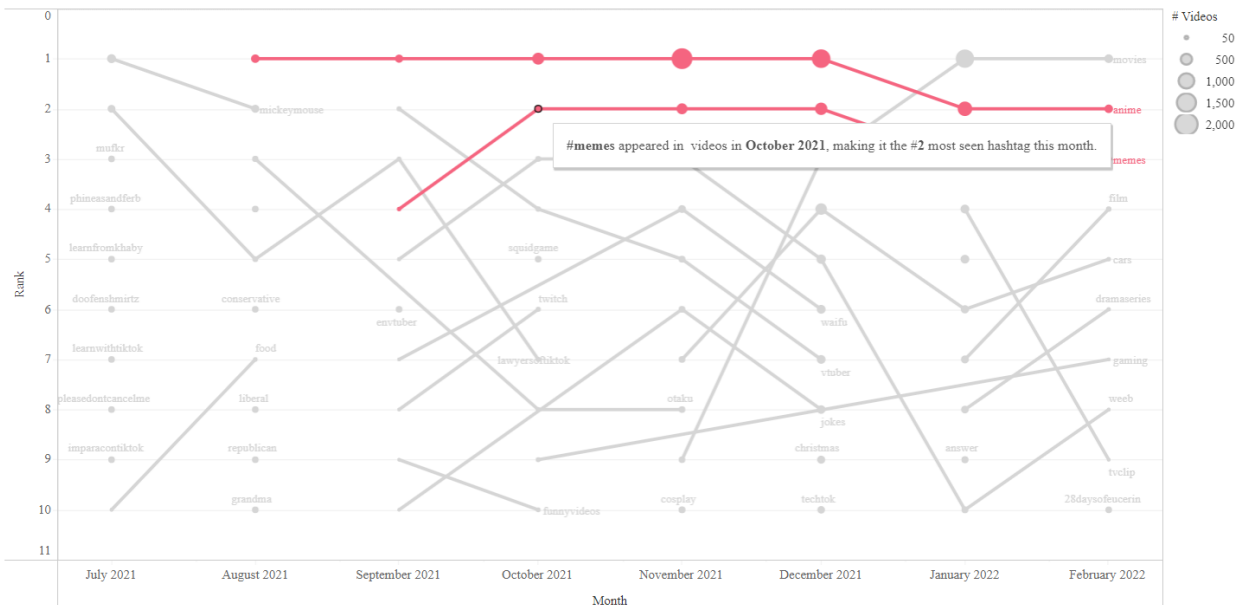
View #2 focuses on the early hashtags, specifically those that appeared in the top 10 once within the first two months. These tended to be thematically unrelated to later popular hashtags, and show the early types of content that dominated Oski's feed.

Many early hashtags quickly grew irrelevant.
We interpret this as the signs of the recommender algorithm figuring out your interests.



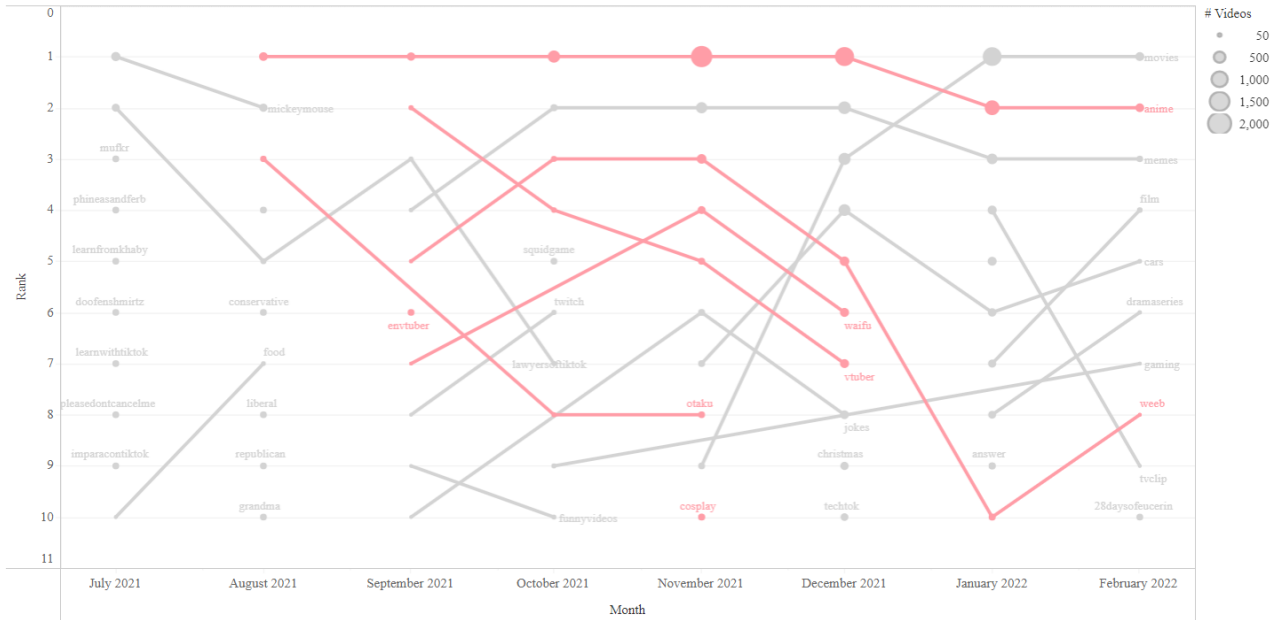
View #3 shows that a few of the top hashtags that stayed dominant over time. From this we inferred that by Oski's 2nd month of use, TikTok had pinpointed some engrossing key interests.

Other hashtags showed up early and remained popular.
For example, anime and memes topped Oski's "hashtag charts" within the second month and have remained dominant ever since.



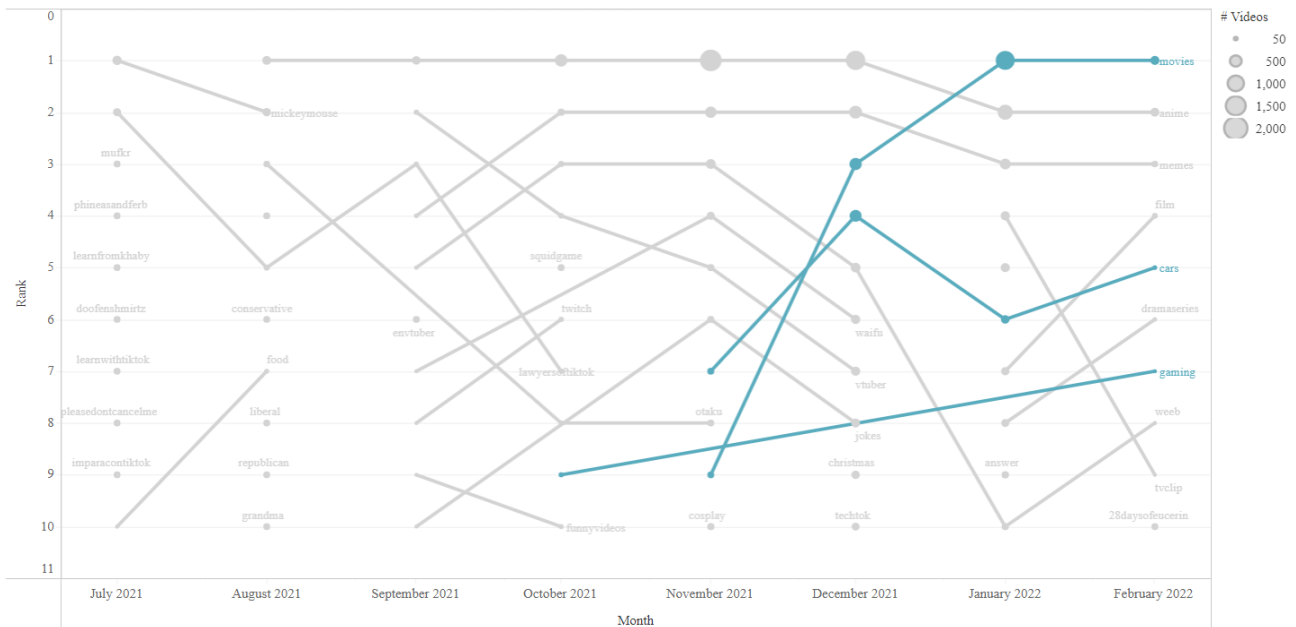
View #4 shows that many related hashtags co-occur. There are a few theories behind these co-occurrences. One theory is user behavior: creators tend to tag videos with many related tags to reach broad audiences. A second theory is that this was a sign of the recommender algorithm at work – serving hashtags related to top ones (like #anime) that Oski found interesting

Many of top hashtags are related.
Often videos are tagged with many duplicative or similarly themed tags to reach broader audiences.



View #5 shows some hashtags (such as cars, gaming) that appeared later and rose significantly in ranking, indicating the discovery of new interests or “niches” of content growing in dominance.

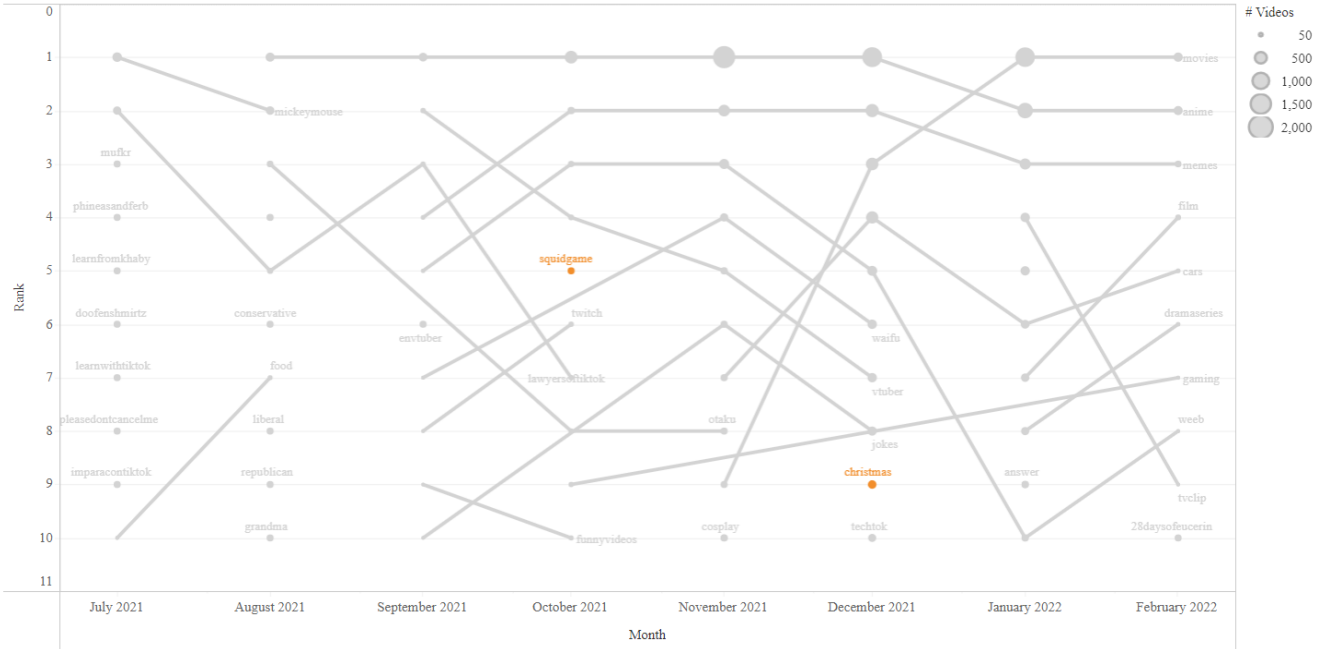
From hashtags trends, we can also see when new video interests grew dominant.
Cars and Gaming didn't top the ranks until later in Oski's TikTok tenure.



View #6 shows that some top hashtags are influenced by seasonal or current pop culture trends, such as holidays (#christmas) or popular TV shows (#squidgame).

Still other hashtags reflect seasonal or popular trends.

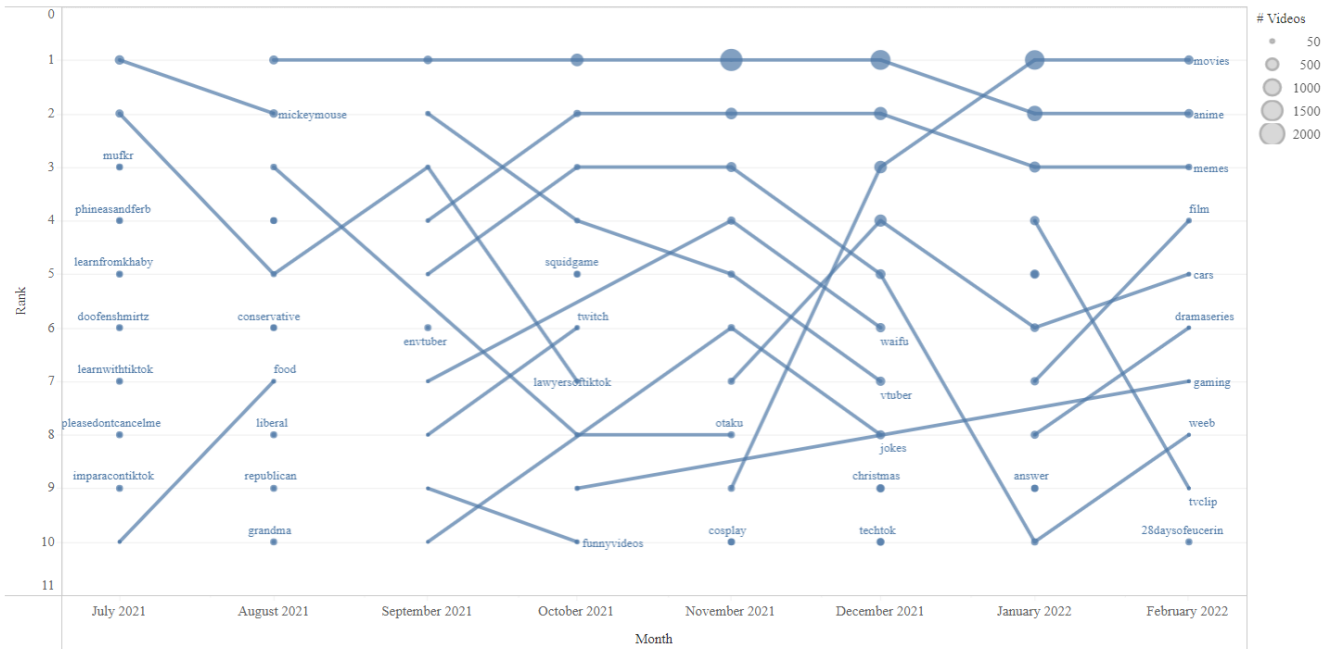
For example, #squidgame remained relevant for the month its namesake TV show was released, as did #christmas for the month of December.



View #7 summarizes the findings and how the different trends identified contribute to the overall turbulent nature of the hashtags represented.

Ultimately, the messiness of these rankings are reflective of the ebb and flow of TikTok's video recommendations.

The dynamic shifts in content are likely part of how TikTok keeps content fresh and users hooked.

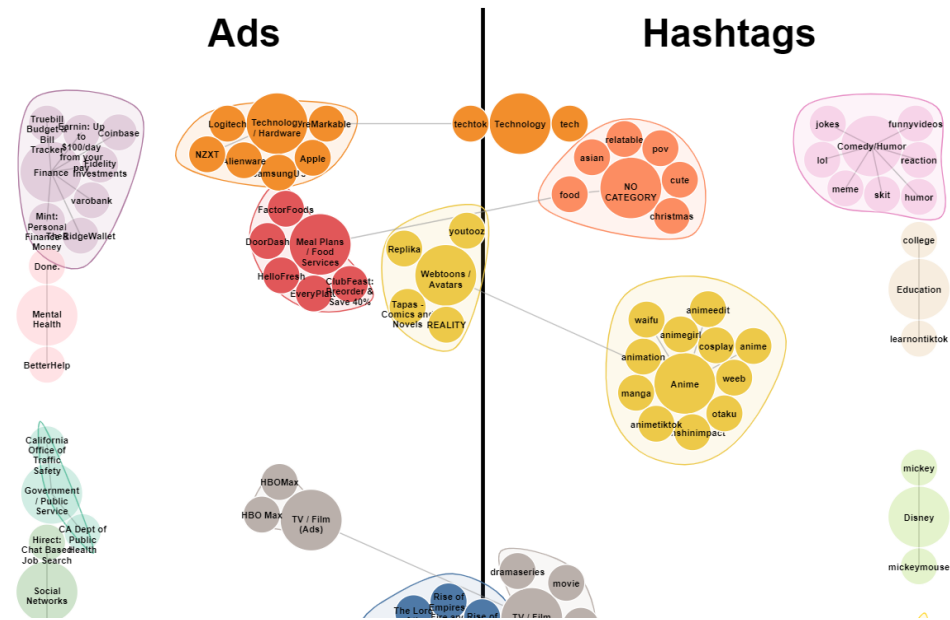


Visualizing the relationship between Categories of Hashtags and Ads

We wanted to explore and visualize whether there was a relationship between Content (Hashtags) and Ads. For this visualization, we took the top 50 hashtags and advertisers by frequency and manually grouped them by theme (If an advertiser or hashtag was unfamiliar, we conducted a web search to understand the meaning of the hashtag or the advertiser's industry).

Then, we used a line to connect the ad/hashtag pairs that were thematically related. We wanted to demonstrate that there seemed to be some relationships between the ad and video content, but not all ad/hashtag groups had a clear pair. We interpreted this to mean that while there is some content or interest based targeting going on, it was not as extensive as we initially expected. Ads seemed targeted based on location (for example, CA Department of Public Health seems based on Oski's CA-based location) and demographics (for example, mental health, hiring apps for someone college-aged like Oski). Still others are likely shown to broad audiences, such as the Retail/Advertising group.

Relationship of the top Hashtags and Ads Categories



We attempted to engage users with the interactivity of the chart executed in D3. Design principles we leveraged included using the gestalt principle of enclosure to indicate related individual hashtags/ads; the gestalt principle of connection to indicate relationships between groups, and used color/saturation to reinforce these connections. The only exception is the “food” hashtag, which was placed in the NO CATEGORY group, paired with the “Meal Plan / Food Services” category. Groups with no pair are lighter in saturation than those with a pair. We also use a border right at the middle to create an enclosure of the ads and hashtags in their own areas. Their links between each side with the same category are the only thing creating a relationship across the border.

Advertising Trends over Time

One of the clearest trends we noted in the data was that the concentration of ads served increased significantly with use. We calculated the total number of ads / total number of videos and plotted a line chart with this proportion.

Between Oski's first month of use and the last month of data we had available, the proportion of ads more than tripled. In our user studies, participants were really drawn to this simple but effective chart. Again, none of them were surprised that this trend existed – one participant said “it makes sense – they draw you in then make money off of you” – but it was engaging to see the evidence and extent of this trend.

Advertising made up more and more of Oski's feed over time

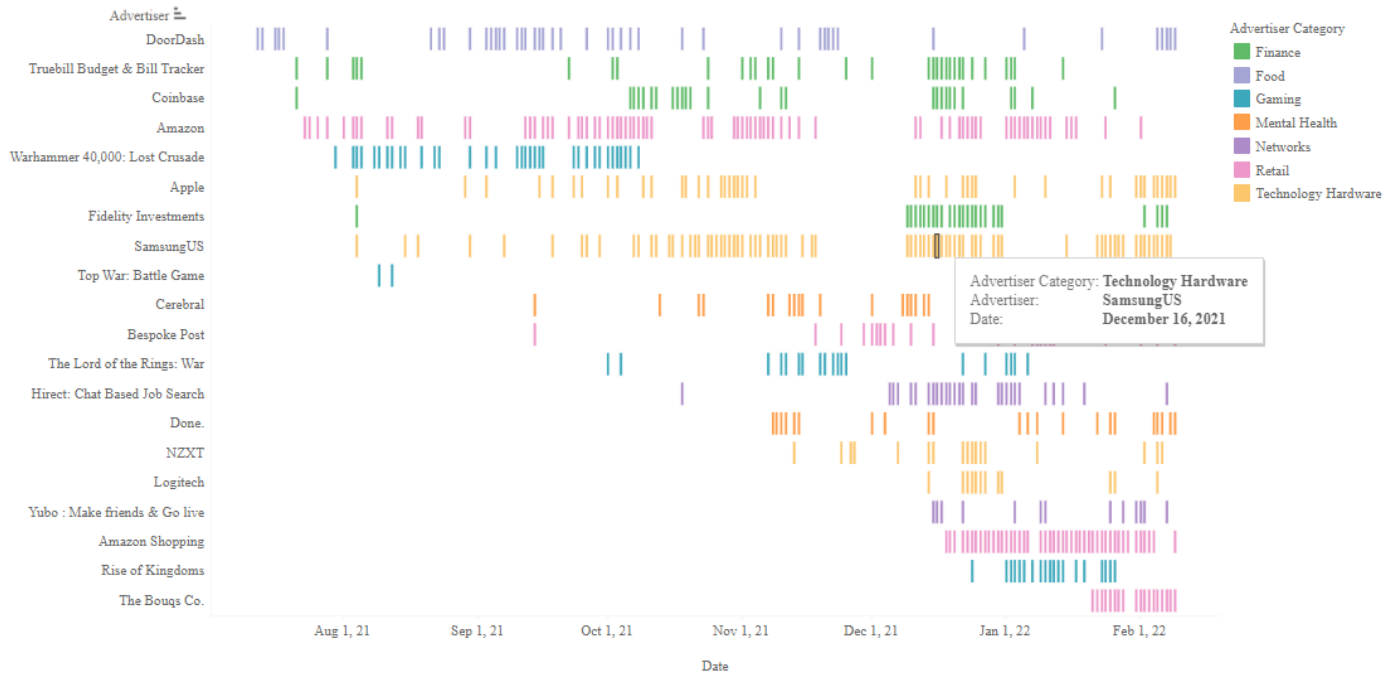


Advertisement Cadence

Finally, we created the advertiser chart inspired by Brandt et al. (2009)'s event log chart. Our version shows the top 20 advertisers by frequency as rows, and instances of ads from that advertiser represented as small vertical bar marks. In this first view, color indicates the advertiser's industry category. The advertisers in this view are ordered by first ad instance (i.e. Doordash showed up in Oski's feed in July 2021, while the Bouqs Co did not appear until February 2022). The primary purpose of this view is to introduce the reader to this kind of chart and show the varying advertising patterns at a high level. For example, some advertisers appear in concentrated bursts (like Warhammer 40,000), while others are more consistent (like Amazon).

Oski's Top Advertisers' Advertising Patterns

Like hashtags, it looks like there is a lot of noise and randomness to advertising.

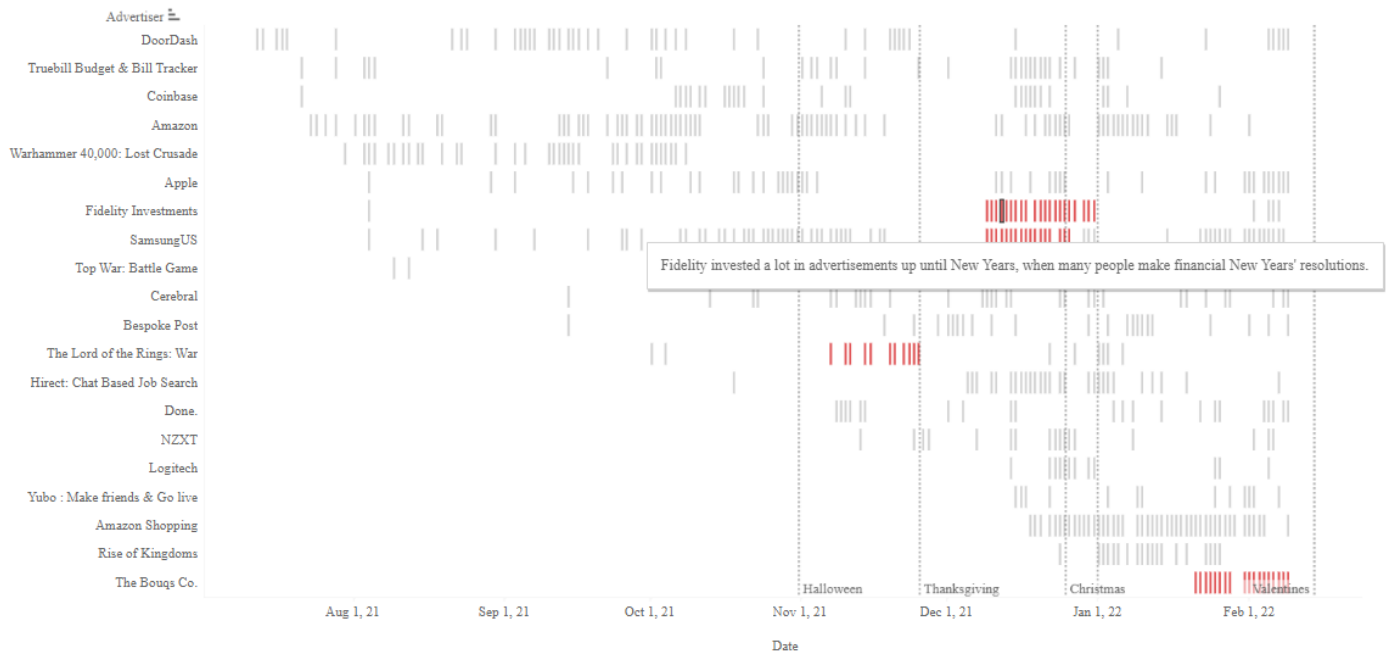


View #2 highlights a trend we noticed that is common in traditional advertising: increased advertising around the holidays. Out of the top 20 advertisers, about 25% of them appeared to have concentrated campaigns immediately before or after a holiday.

In hindsight, we noted that since we didn't normalize for viewing activity, these patterns could be an artifact of Oski's viewing patterns. For example, as a college student, it is likely Oski spent more time on TikTok during breaks, which often align with holidays. Since this is just observational, we noted instances where the advertiser had a strong interest in advertising during a specific holiday (such as Bouqs Co., a bouquet seller, advertising in the lead up to Valentine's day). We would want to repeat this analysis with more data from more individuals to make stronger claims about advertising trends.

For some advertisers, ad appearances tracked closely with holidays.

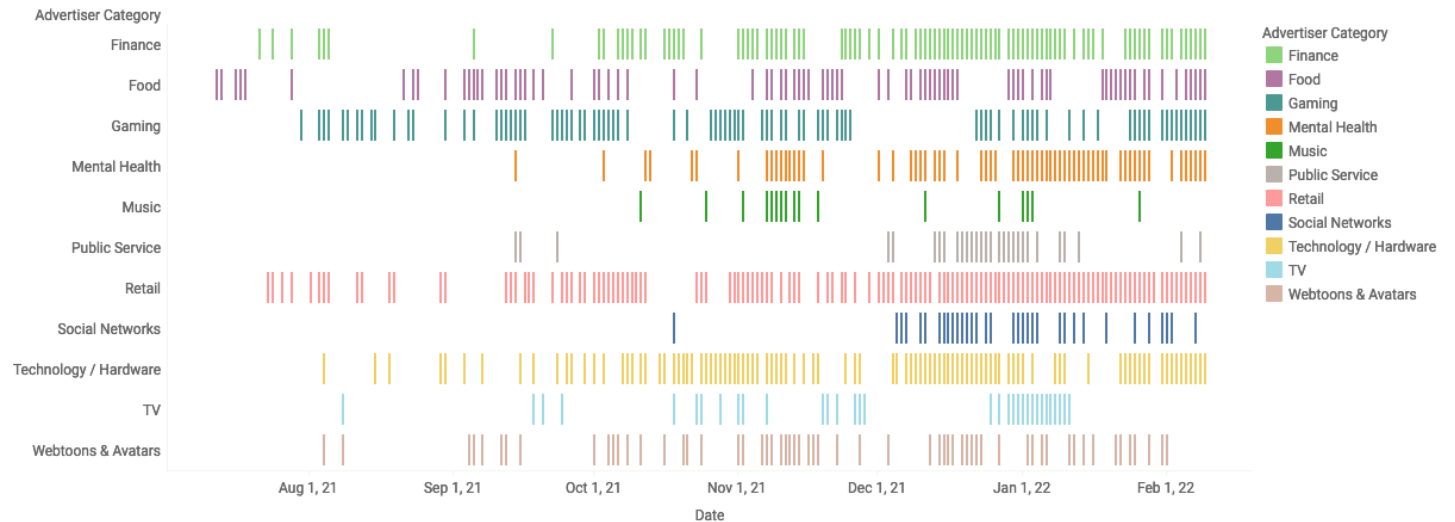
When we layer on major holidays, there are some hard stops in ad incidences suggesting advertising campaigns targeted at holidays. Some (such as Bouqs and Fidelity) make more sense than others.



In View #3, we attempted to visualize by category. Overall this visualization shows the frequency and periodicity of different advertising categories. However, to draw clearer conclusions, an analysis like this would benefit from normalization by the number of advertisers and more users' data.

When we look at Advertiser Categories instead, some subtle trends appear.

We grouped the top 50 advertisers by category and plotted all of their advertising instances. We see that retail and food advertisers tend to advertise regularly, while other industries are more periodic. We'd need to look at data from more than just Oski's to verify these patterns.



Approach

Data Used

We used a single TikTok user's data (with permission, of course) to perform our analysis. The person is referred to as Oski on our website.

Using the Unofficial TikTok API¹ Python library, we took the user's viewing history and also obtained the TikTok Author, Sound name + author, and hashtags. From the original viewing history data, we also got the time the TikTok video was seen and the TikTok URL.

Tools Used

You can find the code for data scraping and pre-processing that we used here:

<https://github.com/ochan1/info247-sp22-tiktok-unwrapped/tree/master/code>

Data Scraping

After getting the data from TikTok, we use the Unofficial TikTok API¹ Python library to get additional data.

We then create another Python script to read the TikTok JSON file and pass it to the library in order to get the information we desire and then output the result in a CSV.

However, because for some technical reason the scraping times out all the time, we wrap it in a custom Shell script to run the code repeatedly in batches, allowing us to bypass the timeout issue that we encountered. This also allows us to have a batch checkpoint system to only rerun a portion of the data in case something happens during the scraping process rather than the whole thing again.

For the hashtag or ads categorization, we manually used our own knowledge to assign the top 50 hashtags and ad companies that appear on the user's feed to a common category.

¹ <https://github.com/davidteather/TikTok-API>

Data Pre-Processing

Taking inspiration on how databases work with multiple data related to one other entry, or in other words One-to-Many relations, we used Pandas to create two CSV databases of the video metadata and hashtags with the metadata entry ID number. To do so, we split the hashtags by commas, as the rules of hashtags are that one can't use spaces and punctuation (including commas) in hashtags, using Panda's "str.split(",")² and then separate the list into individual entries on a separate table via Panda's "explode"³ function.

Data for some of the visualizations like the d3 charts for the file tree and hashtag-ad category relationship were also created from the original and processed data, respectively. The file tree was manually created using the original TikTok given JSON and just stripped of the data. It was as simple as adding "children" to the sub-data of each level alongside other identifying markers. For the hashtag-ad category, we did a bit of EDA beforehand and manually categorized the top 50 hashtags and ads each into their own respective categories. It was then we found out that there was such a relationship, so we added an interactive visualization to best illustrate that.

A Jupyter Notebook was also created to create another CSV marking videos that we believe are ads. The criteria to be an ad has to be any one of the following: capitalized names (regular usernames are required to be lowercase), having the hashtags #ad or #sponsored, or having "Sponsored Sound" in the sound name or sound author.

Finally, we further cleaned the data by identifying and combining similar hashtags or advertisers where applicable (for example, instances of "memes" and "meme" were added together) to reduce the noisiness of the hashtag dataset.

Data Analysis / EDA

We used Python and Jupyter Notebooks to look at the data in closer detail before finally showing visualizations on d3 or Tableau. Sometimes we chose to use Python for our EDA and Data Analysis because Pandas allowed us to more easily work with table operations and view the data table more easily at the same time than Tableau. Tableau does have the ability to automatically create informative charts and visualizations on many simple operations. We use a combination of both.

² <https://pandas.pydata.org/docs/reference/api/pandas.Series.str.split.html>

³ <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.explode.html>

Visualizations

We used a combination of d3.js on Observable, Tableau (via Tableau Public), and standard HTML and CSS.

In Section 2 where we analyzed the user data download, the base code inspired to create the d3 hierarchy was from one of Observable's d3 tutorials on it.⁴ The fading nodes were inspired from Spring 2022 Info 247's Lab 12 on Animating Nodes by Ian Wu.⁵

The grouping of nodes in the hashtag-ads visualization was based off sample code from Xavi Gimenez on Grouping nodes in a Force-Directed Graph,⁶ where we then grouped ads and hashtags by their respective categories.

Some of our diagrams like the data processing pipeline and initial design of the website was done on Figma.

We used a feature Tableau called "Story"⁷ that allows us to create a visualization that allows the user to flip through different Tableau tables with different parts of the chart highlighted and the visualization described for each one.

Website

Our website's base template leveraged the Start Bootstrap's Scrolling Nav template⁸. We changed the fonts but much of everything else stayed the same.

⁴ <https://observablehq.com/@d3/d3-hierarchy>

⁵ <https://observablehq.com/@iwle/i247-lab-12-2>

⁶ <https://bl.ocks.org/XavierGimenez/a8e8c5e9aed71ba96bd52332682c0399>

⁷ https://help.tableau.com/current/pro/desktop/en-us/story_create.htm

⁸ <https://startbootstrap.com/template/scrolling-nav> and
<https://github.com/StartBootstrap/startbootstrap-scrolling-nav>

Usability Testing Results

We conducted a usability study on our project's website, where we wanted to make sure our goals aligned with what the users saw.

Again, the goals as mentioned in the **Project Goals** section were:

- Describing the information available in the TikTok User Download
- Effectiveness of using one user's data as a case study to
 - Demonstrate how we identified ads and video content on the platform
 - Describe what ads and content TikTok served to this user
 - Describe trends and relationships in the ads and content served to this user
- Spark users' interest in accessing their user data and critically considering algorithmic activity

Method

Our Usability Study uses a virtual Zoom call where participants interact with the visualizations online while we view the user and take notes of their interaction.

The methods used here were pre-post surveys, interviews, contextual inquiry, and session recordings.

Participants

We used a convenience sample of 3 people who use video-based social media apps, all of whom use TikTok and one using YouTube Shorts primarily. We purposefully selected one TikTok non-user, since our target audience is social media users in general. We also invited back two people from a previous usability test to see if their comments were addressed and their data from the pre-interview survey then was copied here. All post-interview entries are brand new to see if the new visualizations do change their mind.

Self-identified demographics from our intake survey are listed below:

Participant #	Age	Gender	Education	App Usage	Returning
Participant 1	21	Male	Undergraduate	Occasional TikTok and Daily YouTube Shorts user for a little over a year	Yes
Participant 2	29	Male	Completed Undergraduate	Not a TikTok User	No
Participant 3	30	Male	PhD	TikTok User for > 1 year, daily user	Yes
Participant 4	29	Female	Masters	Occasional TikTok user for more than a year	No

Scenarios / Tasks

Descriptions of all survey and task scenarios conducted via interview are listed.

Pre-Study Questions

We asked a few questions about ads and understanding of data download tools to gain a baseline of participants' familiarity with these concepts.

The questions are:

1. How long have you been using TikTok (or similar video-based social media app like Youtube Shorts or Instagram Reels)?
 - a. > 1 year
 - b. 6-12 mos
 - c. < 6 mos
 - a. Not a user of any of these kinds of apps
2. How often do you use the TikTok app (or similar video-based social media apps like Youtube Shorts or Instagram Reels)?
 - a. Several times a day
 - b. Once a day
 - c. A few times a week
 - d. Once a week or less
 - a. Not a user of any of these kinds of apps
3. How interested are you in understanding the data collected about you by social media sites (including TikTok, Twitter, Facebook, Google, Instagram, Facebook, Snapchat)?
 - a. Extremely interested
 - b. Very interested
 - c. Moderately interested
 - d. Slightly interested
 - e. Not at all interested
4. Are you aware of the data download tools available from social media sites (including TikTok, Twitter, Facebook, Google, Instagram, Facebook, Snapchat)?
 - a. Yes
 - b. No
5. **If yes:** Have you ever accessed any user data downloads?
 - a. Yes, from more than one of these sites
 - b. Yes, from one of these sites
 - c. No, never
6. Out of 100%, please estimate how much of your TikTok feed contains ads. (Another way to consider this question: out of every 100 videos, how many videos would you estimate are ads?)
____ %

Study Tasks

Task 1: Overall Site

We ask the user to interact with the site for a few minutes on their own. We were interested in seeing what parts of the site they spent the most time with, were there any website features they skipped over, as well as what open-ended questions and comments they had – in the absence of direction from the interviewer.

Task 2: Interact with the D3 Tree Viz

We assess the effectiveness of a hierarchical tree visualization displaying the structure of the user data download in communicating the completeness and content of the data provided by TikTok. We ask the interviewees questions about the contents of this file.

In particular, we paid attention to see if the interviewees:

- Used the fading dots feature to quickly see what nodes can be expanded
- Do they question the data available in the Activity or Ads and data sections of the file hierarchy?
- Any areas they interact with that we didn't consider looking at?

Questions:

1. What kind of information is available in the TikTok user data file?
2. What kind of information is missing from the TikTok user data file?
3. What are your takeaways, if any, from viewing this chart?

Task 3: Flow-Chart Visualization on Getting Video Metadata

Because we also wanted to educate users on understanding their data, we created a flow chart on the pipeline to get more information about a user's data.

We pay attention to see if the user follows their cursor along the path we wanted them to follow and understand the process we did to get the data we have.

Questions:

1. Explain the data flow of getting the data and processing the resulting data.
2. How useful is knowing the process of getting the extra data, regardless if you plan on in the end using it or not for your own data?

Task 4: Interact with Hashtags Rank Chart Story Vizzes

We let the user view a case study of someone's TikTok user data for their top 10 hashtags. We present this as a way to let the user see their interest over time and see if the user is able to use this tool to see their digital profile of their interest and when TikTok will keep serving that hashtag to the user once they find out the user is interested in that particular hashtag. Additionally, we only show the top 10 to avoid overwhelming the user on the vast amount of hashtags in the data.

In particular, we paid attention to see if the interviewees:

- We able to understand the chart based on the storytelling and emphasizing some parts of the visualization

Questions:

1. When were the most trending hashtags for Oski?
2. What did you not understand about the charts?
 - a. Side note: We hope to see if the storytelling was adequate enough to explain the charts and was easy enough to spot
3. What are your takeaways, if any, from viewing this chart?

Task 5: Interact with Hashtags-Ads Category Relationship Viz

Sampling only the top 50 hashtags and ad companies, we look and visualize if there is some correlation between a category of ads to a category of hashtags.

We paid attention to see if the interviewees:

- Had any general feeling about the chart and it's interactivity
- Did the distinguishing between more clear nodes and regular nodes help show the difference between categories with ads-hashtag relationships?

Questions:

1. What is the relationship between hashtags and ads?
2. What are your takeaways, if any, from viewing this chart?

Task 6: Interact with Ads Viz

Using the user data case study, we present if it is possible to visualize the statistics around how much of their feed is advertisements.

Questions:

1. What % of ads made up Oski's feed the first month he joined TikTok?
2. What % of ads made up Oski's feed the last month he joined TikTok?
3. What are your takeaways, if any, from viewing this chart?

Task 7: Interact with Gantt Chart Story on Ads Vizzes

TikTok serves certain ads to users and how they determine what ads to serve is a mystery. We display the top 20 ads from the case study of someone's TikTok user data. We can then see what ads TikTok decided would be interesting for the user.

In particular, we paid attention to see if the interviewees:

- We able to understand the chart based on the storytelling and emphasizing some parts of the visualization

Questions:

1. When were the most trending ads for Oski?
2. What did you not understand about the charts?
 - a. Side note: We hope to see if the storytelling was adequate enough to explain the charts and was easy enough to spot
3. Anything you find interesting for the ads visualization?
4. What are your takeaways, if any, from viewing this chart?

Post-Study Survey

1. Which visualization did you like most? Why?
2. Were there any parts that were especially difficult to understand? Why?
3. What other suggestions do you have for improving the visualizations?
4. How interested are you in understanding the data collected about you by social media sites?
 - a. Extremely interested
 - b. Very interested
 - c. Moderately interested
 - d. Slightly interested
 - e. Not at all interested
5. Please share what you learned, if anything, about the data download tools available from TikTok?
6. Please share what you learned, if anything, about ads and video content on TikTok.

Results

Pre-Study

	User 1	User 2	User 3	User 4
Aware you can download your data, and have you done so and where?	Yes, and Yes from multiple sites	Yes, and Yes from only one site	Yes, and Yes from only one site	Yes, and Yes from only one site
Percentage estimate of ads in feed	25%	8%	5%	10%

Study Tasks

Task	Task Completion	Qualitative Results
Task 1: Overall Site	Average Time spent on site: 10 mins	<ul style="list-style-type: none"> Surprised on the complexity to get information to get user data, especially for returning participants Skimmed through the storytelling of tableau Quick use of sections that don't have another dimension of date Returning participants enjoyed the extra details available for them to read!
Task 2: Interact with the D3 Tree Viz	All able to use the storytelling description to find the more important data	<ul style="list-style-type: none"> Clicking on all the bubbles that fade to see children Storytelling helped navigate the more important parts of the hierarchy viz
Task 3: Flow-Chart Visualization on Getting Video Metadata	All able to describe the process of getting the data	<ul style="list-style-type: none"> Easy to understand and see the expected data that result Still not interested in the daunting task of processing data

<p>Task 4: Interact with Hashtags Rank Chart Story Vizzes</p>	<p>All able to navigate through the story.</p> <p>Because of one of the storytelling charts, all were able to identify the top trending hashtags for the user.</p>	<ul style="list-style-type: none"> ● Can quickly see the more focused parts of the visualization stories ● The breaking of the rank and links were weird. One participant noticed there were more hashtags than ranks. ● Enjoyed the ever changing digital profile over time ● All realized how much Squid Games fell out of favor, in general
<p>Task 5: Interact with Hashtags-Ads Category Relationship Viz</p>	<p>All users are able to find the relationship of categories via lines and/or coloring.</p> <p>It took a bit to realize for one that there was a dividing line between Ads and Hashtags.</p>	<ul style="list-style-type: none"> ● Enjoyed the playful interactivity ● Large was category, small was hashtags / ads ● Two users enjoyed moving the nodes around to focus on certain category relationships ● Although a bit messy at first, need to self-organize it
<p>Task 6: Interact with Ads Viz</p>	<p>All able to identify the first month and last month percentage of ads in the user feed via mouseover on the chart</p>	<ul style="list-style-type: none"> ● Easy to find the percent of ads relative to TikTok content ● Appreciated the storytelling on some trending parts of the graph ● Returning participants pointed out removing the line was probably a good step since it was not interesting

<p>Task 7: Interact with Gantt Chart Story on Ads Vizzes</p>	<p>All able to navigate through the story</p> <p>Returning user, User 1, still found the first story section of the visualization cluttered</p>	<ul style="list-style-type: none">• Can quickly see the more focused parts of the visualization stories• Very interesting to see an upper-right triangle pattern in the ads• In the third storytelling, users agreed that there was some difficulty in being able to identify ads vs hashtag relationships with the charts. The interactive category relationship chart helped better.
--	---	--

Post-Survey

	User 1	User 2	User 3	User 4
Visualization liked the most	Hashtags Rank Chart Story Visualization	Ads-Hashtag Category interactive visualization Also liked the Data Scraping process description	Gantt Chart Story on Ads Visualization	Gantt Chart Story on Ads Visualization
Parts difficult to understand	Gantt Chart Story on Ads Visualization was cluttered and unsure besides the increase concentration at some points what else to see	The bars on the Gantt Chart Story on Ads Visualization were a bit confusing at first	The bars on the Gantt Chart Story on Ads Visualization were a bit confusing at first	None
Other suggestions do you have for improving the visualizations	Nope, really liked how there is now a description on many of the visualizations now!	Nope, really like the website visualization	Nope, liked the new descriptions and highlighting to target certain parts of the chart	Nope, really like the website visualization
Interested in understanding social media data	Very interested	Very interested	Extremely interested	Extremely interested
Data download tools comments	Know how to access them and the steps needed Would perform scraping given code	Know how to access them and the steps needed Unlikely to perform scraping given code	Know how to access them and the steps needed Unlikely to perform scraping given code	Know how to access them and the steps needed Would perform scraping given code

<p>Ads and Video content comments</p>	<p>Look back at interested in the particular topic</p> <p>Good understanding of interests better</p> <p>Expect lots of interest about their own data, we can already see the interests of the case study person</p>	<p>Good understanding of interests better and how the patterns TikTok is using that to serve ads for user</p>	<p>Interested to see what kind of content is targeting a user</p>	<p>Good understanding of interests better</p> <p>Interesting to see certain ads are targeted based on interests and times</p>
---------------------------------------	---	---	---	---

Usability Study Results Discussion

We considered the feedback seriously and made improvements based on both qualitative feedback and suggestions. Overall, we also noted the importance of the background data and process information to understanding the charts so we focused on enhancing the surrounding narrative.

Otherwise, general sentiment was overall positive and returning users really enjoyed the new storytelling and descriptions we provided to the visualizations.

Data Download Tree Hierarchy Chart

From last time's usability study, participants noted it was pure guesswork to find out what could be expanded and what could not. As a result, we decided to make the data nodes with children fade on hover. We also added an additional multiple data light blue node to emphasize that there is multiple data.

It was confusing on what the audience was supposed to get from the visualization. The best thing we could do to remedy this was to guide in the website what to click and explain to the user the data we could use and what was omitted. The ads nodes are highlighted white to show that there is a field for ads, but are blank.

Hashtag Chart Changes

We found that the initial hashtag gantt chart we tested with users was far less engaging than the advertiser gantt chart, even though it was presented after the advertiser chart. Participants found this relatively cluttered and confusing to follow. Taking this feedback into consideration, we decided to limit the number of hashtags and facet out top hashtags rankings by month (rather than rankings over the entire time span of the data). The result was the rank chart, which shows similar takeaways in a relatively less cluttered way.

We also implemented progressive disclosure to both this new rank chart and the Advertiser Gantt Chart with Tableau's Stories feature. This added an element of interactivity while limiting the amount of information users had to take in at once.

% Ads and % Advertisers Line Chart

We repeatedly got feedback that the % advertisers to content authors ratio was confusing; participants often fixated and spent several minutes trying to think aloud about what that implied. Since we did not think this added much to the visualization, we decided to omit that statistic from the line chart to let the user focus in on that main takeaway. This created a more simple visualization for both returning and new participants.

Ads-Hashtags Prototype

The initial prototype implemented in Miro was executed as a d3 visualization in Observable. The d3 added elements of color and made the categories and linkages more obvious. Participants were really happy with the interactivity and playfulness of the visualization, and could sort on their own what category relationship they wanted to focus on more. However, they did note it was a bit messy, but better than nothing and a static image.

Links

- Visualization: <https://ochan1.github.io/info247-sp22-tiktok-unwrapped/website/>
- Pre-Processing Code: <https://github.com/ochan1/info247-sp22-tiktok-unwrapped/tree/master/code>

Work Distribution

Tasks Completed	Oscar Chan	Angela Liu
Data Sourcing	100%	0%
Data Scraping	100%	0%
EDA	50%	50%
Data Processing & Analysis in Python	20%	80%
Website Implementation	80%	20%
Tableau Visualizations	10%	90%
Observable D3 Visualizations	80%	20%
Website Narrative	30%	70%
Usability Study Design	50%	50%
Usability Study Interviews	60%	40%
Final Report Writeup	50%	50%