# The TikTok Self: Music, Signaling, and Identity on Social Media

Jeffrey Sachs
Department of Sociology
Yale University

Rahshemah Wise
Yale College
Yale University

Daniel Karell
Department of Sociology
Yale University

*Corresponding author*

Jeffrey Sachs
jeffrey.sachs@yale.edu
Yale University
Department of Sociology
493 College Street
New Haven, Connecticut

## ABSTRACT

The COVID-19 pandemic has shifted many social worlds from in-person interaction to online activity. As a result, people increasingly have the freedom to choose more agreeable everyday environments. While such freedom has often been associated with negative outcomes – namely, the emergence of "echo-chambers" that corral insensitivity – this data visualization suggests a positive outcome: "flowering-chambers", where the freedom has enabled expressions of a truer self. Drawing on an original dataset of TikTok videos, the visualization charts a considerable increase in the number of "coming out" gender identity and sexual orientation videos during the latter three quarters of 2020. These results suggests that many TikTok users have publicly revealed private aspects of their identities, which we attribute to individuals becoming increasingly embedded in agreeable online spaces while quarantining or socially distancing. The visualization additionally introduces a publicly available dataset of 4.8 million TikToks to facilitate future research using data from the platform.

## KEY WORDS

TikTok; Covid; Gender identity; Sexual orientation; Dramaturgy; Music

**DATA VISUALIZATION**

Among the changes brought about by the Covid pandemic has been a shift in our social relationships. For many people, especially those living under prolonged quarantine or social distancing mandates, the balance of interpersonal connections has shifted from in-person towards online interactions. As a result, some of the societal and institutional expectations usually experienced through face-to-face interaction have likely dissipated and been replaced over the course of 2020. That is, with an increasingly online life, people can avoid or even block disapproving others, thereby mitigating unwanted social expectations and sanctioning while selecting and insulating within more agreeable communities.

The increased opportunities to select into agreeable interactions during quarantine has potentially brought about profound effects on our identities. Under typical circumstances, to use Goffman's (1959) dramaturgical metaphor, people usually invoke one identity, or performance, in public (the "frontstage"), while adopting another in private (the "backstage"). The terms of the former performance are defined by prevailing expectations and norms, and individuals are policed and shamed when their performance falls short. In the backstage, where individuals are largely unhampered by societal expectations, people tend to flout societal rules, nurture close relationships with similar others, and, in many cases, perform a more honest version of themselves. In short, the distinction between front and back stages – and the distinction between the identities we perform – rests on being exposed to external expectations and the sanctioning behavior of others.

Along this theoretical line, the metaphorical frontstage would be significantly weakened if individuals could freely choose a growing portion of the societal expectations and norms they encountered – as might happen through, say, long-term quarantine and a transition to a primarily online life. With such freedom, backstage life could move to the foreground, resulting in the public expression of individuals' oft-hidden, more true selves. In some ways, this process would be akin to

entering the frequently derided online "echo chambers" (see Bail et al. 2018), in which individuals can be radicalized by way of being exposed to only one worldview. We, however, conceptualize an alternative, liberating process: individuals become more comfortable expressing versions of themselves previously kept backstage. Indeed, the process could lead to more explicit expressions of suppressed hateful beliefs (Kilvington 2021), but it could also free individuals to express innocuous aspects of themselves as part of their everyday sociallife, such as gender identity and sexual orientation. Thus, we posit that the shift to online social life has created chambers allowing various self-identities to blossom and thrive.

In this data visualization, we offer evidence of backstages coming to the fore due to the quarantine and social distance regulations imposed during the pandemic. The visualization draws on an original dataset of 3,337,925 TikTok posts, sampled from 15,000 randomly chosen public TikTok users.[1] TikTok is a video sharing social media platform launched in 2017 with a large and growing user base.[2] The platform allows users to create short videos by compiling elements (e.g., videos shot with one's phone, text, and song clips) and editing the compilation with filters, voice effects, speed modulators, and other tools. Users can then share videos with a group of followers or other users they follow. Importantly, users can also select existing (posted) videos' music, or sound clips, for their own videos. Because of this feature, the user's choice of background music often serves as a salient indicator of the video's content and meaning: song clips become imbued with specific and widely recognized meaning as users pull music from other videos with the same message and incorporate it into their own videos. In this manner, users can draw on collective meaning (while contributing to it) to convey particular ideas (Serrano et al. 2020).

---

[1] See Appendices A and B in the supplementary materials for details of the data collection and dataset.

[2] In early 2020, TikTok had 26.5 million active users in the United States. Sixty percent were between the ages of 16 and 24 (Serrano et al. 2020).

Our visualization displays the distribution of three kinds of TikTok videos from January 1, 2020 to March 10, 2021. One set of these videos uses four songs with specific and widely recognized meanings in the world of TikTok. These songs are used to signal gender identity or sexual orientation, and are integral to a "coming out" genre of videos on TikTok.[3] The other two kinds of videos feature songs that do not convey a particular meaning within TikTok: these are the top four most popular songs of 2020 or two of the most popularsongs on TikTok in 2020.[4] If frontstages have been weakened – and if people have been entering "flowering chambers" due to the replacement of in-person social life with the increasing predominance of online life during the pandemic – we should observe an increase in the usage of the "coming out" songs in the latter three quarters of 2020. In addition, to qualify this observation as independent of systemic trends, we should *not* see greater usage of the songs from the latter two kinds of videos over the same period.

[3] See Appendix A for validation of the interpretation of these songs' meaning.
[4] The two popular TikTok songs of 2020 are two of thetop 10 ranked songs on TikTok for 2020 that were also released prior to March 2020. The eight songs released during or after March 2020 would necessarily have a growing rate of usage during 2020. See Appendix A in the supplementary materials for further details.

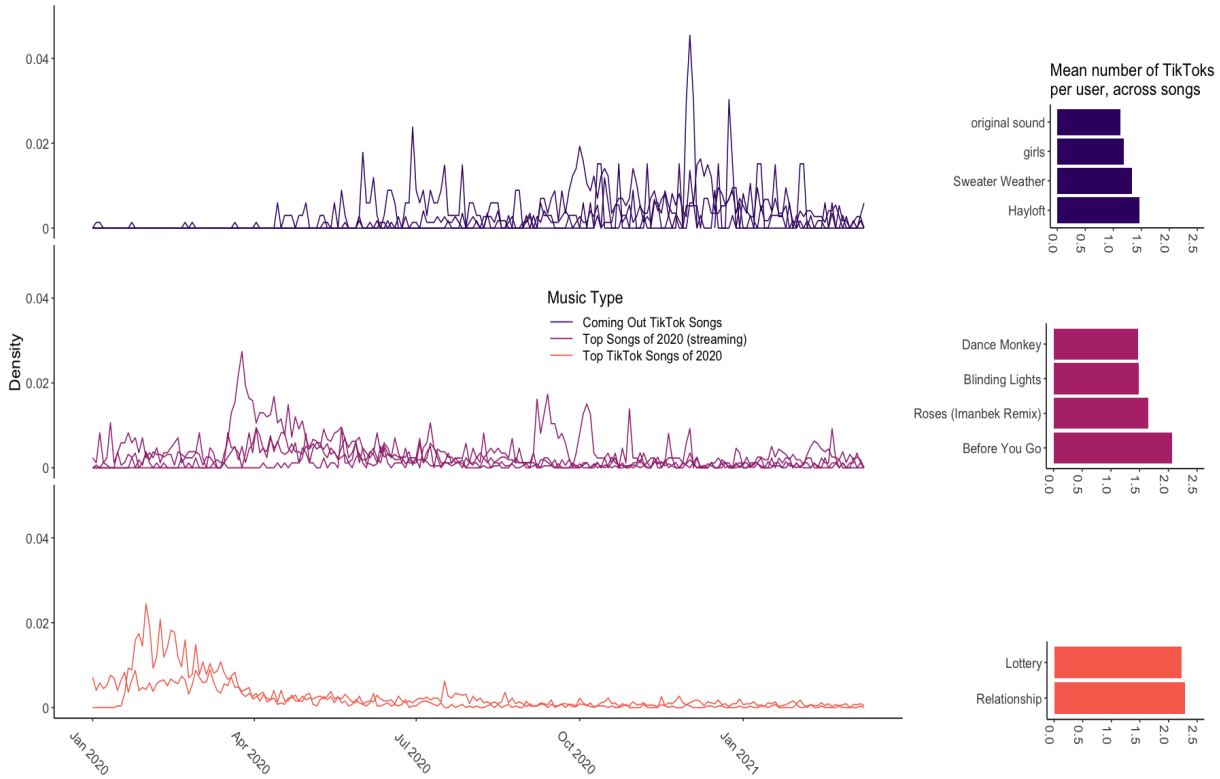


**Figure 1**. The prevalence of different music types on TikTok during 2020. The line plots show the distribution of TikTok posts for each song between January 1st, 2020 and March 10th, 2021. Observations are binned into two-day spans. The bar plots show the average number of TikTok posts per user featuring one of these songs. The visualization depicts "coming out" songs (n = 854), top general "pop" songs of 2020 (n = 2,688), and top "TikTok pop" songs of 2020 (n = 3,579), identified in a sample of 3,337,925 TikTok videos.

The uppermost plot of Figure 1 shows that, as quarantine wore on, the number of TikTok videos featuring one of the four recognized "coming out" songs increased considerably. In contrast, the middle and bottom plots show that the videos using the generally popular songs or the popular TikTok songs generally decreased or were otherwise stable over the quarantine period. The somewhat stable rates of the latter two groups suggest that the growth in "coming out" songs was not simply a function of TikTok's growing user base. The bar plots along the right-hand side of

Figure 1 show the mean number of TikTok videos per user across each song choice and support the interpretation of the "coming out" songs. If thesesongs are in fact predominantly used to signal gender identity or sexual orientation, we should typically see these songs used only once by each user – that is, to "come out" – rather than multiple times per user, as would be more common among general fans of the music. The bar plots confirm this logic, showing the average number of TikTok videos per user for the "coming out" songs to be considerably closer to one than for the 2020 popular songs or popular on TikTok songs.

The visualization indicates that numerous people leveraged music to signal their gender identity or sexual orientation on TikTok over the pandemic period. This finding advances our understanding of the broader effects of society'sresponse to Covid. Along with consequences related to health, education, politics, labor and inequality, the pandemic appears to have also transformed many people's expression of identities. Recognizing that the replacement of in-person social life with online social life can foster freedoms and flexibility with respect to personal identity,[5] we should further consider the dynamics underlying "echo chambers". Namely, "echo chambers" could realize both negative and positive outcomes. In addition, we hope that our visualization encourages more research using TikTok as a fruitful source of data. After all, TikTok serves as a leading channel of expression among young people. To support such research efforts, we make publicly available a version of our dataset that includes additional features not used in this visualization.[6]

[5] Of course, online life could also entail more careful management of one's backstage life, including the manufacturing of mock "backstages" (see Stuart 2020). We do not see such processes as mutually exclusive to the one we describe; both can occur across a society of community simultaneously.

[6] See Appendix C for details.

**REFERENCES**


Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization", *Proceedings of the National Academy of Sciences* 115(37):9216-9221.

Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. New York: Doubleday.

Kilvington, Daniel. 2021. "The Virtual Stages of hate: Using Goffman's Work to Conceptualize the Motivations for Online Hate", *Media, Culture & Society* 43(2): 256-272.

Serrano, Juan Carlos Medina, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. "Dancing to the Partisan Beat: A First Analysis of Political Communication on TikTok", *WebSci '20* (Southampton, UK). DOI:10.1145/3394231.3397916.

Stuart, Forest. 2020. "Code of the Tweet: Urban Gang Violence in the Social Media Age", *Social Forces* 67(2): 191-207.

Supplementary Material

*for*

# The TikTok Self: Music, Signaling and Identity on Social Media

April 2021

**Contents**

Appendix A. Data collection procedures

Appendix B. Details of the data

Appendix C. Information for future use of the dataset

**Appendix A. Data collection procedures**

Our main data collection procedures helped us approximate a random sample of TikTok posts with which we could measure the prevalence of particular songs (henceforth, "focal songs"). This data gathering was preceded by an initial round of data collection. Specifically, we initially collected `musicID` codes from official TikTok pages for each focal song, then used the "Unofficial TikTok API" to collect metadata on posts that used the songs.[7] For each query, we attempted to collect as many posts as we could. However, TikTok limited the amount of metadata that could be collected to about four thousand posts per song request. Little documentation exists on the sampling process for these posts, but our analysis of the metadata revealed that the posts were from public accounts and were ordered with respect to `digg_coun`t (the number of likes a video creator has received). Thus, our first round of data collection yielded observations of the four thousand most popular videos using each of the 10 focal songs, conditional on the video being posted to a public account. Analysis on this dataset confirmed the results found in the main visualization (Figure 1 in the main text), which was produced with data gathered in a second round of data collection.

We undertook a second round of data collection because in the initial dataset the aggregate number of TikTok videos featuring each song varied dramatically while the number of observations for each song in our dataset was relatively similar (see Table B1). This indicated that the initial dataset potentially over-represented the prevalence of the songs of primary interest (*i.e.*, the "coming out" songs). To circumvent this problem, we produced the second dataset and measured the prevalence of the focal songs among TikTok posts from users that were sampled randomly from a much larger population of users. To construct the population of users, we gathered a list of 10,000 songs from 2010 to 2019 from Spotify.[8] We also added to this list songs from four other playlists on Spotify that related to popular music and popular TikTok music for 2019, 2020, and 2021.[9] After removing redundancies, our list comprised 10,251 songs from 2010 to 2021. Importantly, the "coming out" songs of interest were not included in this list of songs. We then used a search

[7] https://github.com/davidteather/TikTok-Api
[8] https://open.spotify.com/playlist/69VOdzkx8II6pi0sUvcu65
[9] https://open.spotify.com/playlist/0UEXxMFJNARHB3XEfUhgEd;
https://open.spotify.com/playlist/3lxZEjjwYHtls9aD9zU65A;
https://open.spotify.com/playlist/65LdqYCLcsV0lJoxpeQ6fW;
https://open.spotify.com/playlist/5AU9HfecQkzlu52erbiyjK

algorithm to collect the music_id codes from each song's official TikTok music page. Altogether, we identified TikTok musicID codes for 5,882 songs.

Next, we collected TikTok post metadata from the official music pages for each of the 10,224 songs using the "TikTok Scraper & Downloader."[10] This collection procedure resulted in 1,558,858 TikTok posts from 1,064,170 unique public users. From the list of unique public users, we then randomly sampled 15,000 users and collected the available metadata on their TikTok posts. The result was data on 3,337,925 TikTok posts, the number of observations in our main (second) dataset. Collection of data from each user was limited to 2000 posts, so if a user had produced more than 2000 TikTok, we collected data on their 2000 most popular TikToks in terms of diggCount. This dataset contains 8,168 TikTok videos that featured the focal songs (see Table B1).

The focal songs meet three different criteria. First, some focal songs are those that were widely used on TikTok to announce a gender identification or sexual orientation. These four songs were: "Original Sound" by Lucy, "girls" by girl in red, "Hayloft" by Mother Mother, and "Sweater Weather" by The Neighbourhood. Second, we chose four currently popular "pop" songs to serve as a comparison group. To identify these songs, we used the *Official Charts* website[11] and chose four songs from the top-40 most-streamed songs of 2020 list. In selecting these songs, we identified four songs that did not map saliently onto any particular meaning in TikTok. We also required that the songs were released prior to March 2020 (which we qualified as the beginning of the general United States quarantine period). Ultimately, we selected the top four songs on the top-40 list, as each of these songs satisfied the requirements. The songs were: "Before You Go" by Lewis Capaldi, "Blinding Lights" by the Weeknd, "Roses" by SAINt JHN, and "Dance Monkey" by Tones and I. Third, to serve as another comparison group, we chose two TikTok songs from a list of the top 10 most popular songs on TikTok during 2020.[12] As with the "pop" songs, we required that the popular TikTok songs were released prior to the start of the general quarantine period so that the effect of popularity and the potential effect of quarantine could be more easily disentangled and identified in the visualization. The two selected songs, "Lottery" by K Camp and "Relationship" by Young Thug, were the only two on the top-10 list meeting this criterion.

---

[10] https://github.com/drawrowfly/tiktok-scraper
[11] https://www.officialcharts.com/chart-news/the-official-top-40-biggest-songs-of-2020__29264/
[12] https://www.popbuzz.com/internet/viral/top-tiktok-songs-2020/

The "coming out" songs were identified based on the expert knowledge of a member of the authorship team. To validate the identification, we conducted word frequency analyses using the desc variable for each video type. Table A1 shows the top 10 most frequent words for each song type. We see in this table that words associated with gender identity and sexual orientation are not only frequent, but also more frequent than in the descriptions of songs in the two control groups. We additionally validated the interpretation of "coming out" songs by measuring the average number of times a user posted a TikTok featuring the song. The logic was that "coming out" songs would be used considerably closer to one time per user than the popular comparison songs, as "coming out" would often be a single occurrence. The right-hand bar plots of the visualization (Figure 1 in the main text) confirm a considerably lower average per user use for the "coming out" songs than for the songs in the control groups.

| | **Coming Out** (n = 854) | N | **Pop** (n = 2,688) | N | **TikTok** (n = 3,579) | N |
|---|---|---|---|---|---|---|
| 1 | シ | 63 | la | 98 | renegade | 239 |
| 2 | i'm | 35 | suite | 85 | day | 81 |
| 3 | love | 32 | dans | 83 | parati | 69 |
| 4 | bi | 30 | instant | 83 | charlidamelio | 64 |
| 5 | xyzbca | 30 | art | 81 | i'm | 64 |
| 6 | lgbt | 28 | parati | 77 | xyzbca | 63 |
| 7 | greenscreen | 27 | pourtoi | 70 | de | 60 |
| 8 | gay | 25 | love | 69 | fy | 52 |
| 9 | lesbian | 24 | poutoi | 63 | lol | 50 |
| 10 | parati | 24 | de | 59 | love | 47 |

**Table A1.** Word frequencies on TikTok posts by song type

**Appendix B. Details of the data**

The dataset used for the visualization contains metadata on 8,186 TikTok videos, narrowed to 7,121 TikToks posted between January 1, 2020 and March 10, 2021. The observations are instances in which one of our songs of interest were used in aTikTok video contained in our sample of 3,337,925 videos shared by 15,000 randomly sampled users.

Table B1 shows the number of observations per song, as well as whether the song is part of the "coming out", generally popular, or the popular on TikTok set. The table also shows the total number of TikTok videos featuring the song as of March10, 2021.[13]

The information we collected covers a range of video attributes (see Appendix C), but the visualization makes use only of `date` (the time a video was uploaded), `music_id` (the identifier of the song), and `type`, a classified we assigned based on whether the song is part of the "coming out", generally popular, or popular on TikTok group. The date variable ranges from January 1, 2020 to March 10, 2021, the day we collected the data.

| Song title | Music type | N | Total TikToks |
|---|---|---|---|
| Girls | Coming out | 161 | 119.6K |
| Hayloft | Coming out | 285 | 222K |
| Original Sound | Coming out | 33 | 12.6K |
| Sweater Weather | Coming out | 368 | 457.7K |
| Before You Go | Generally popular | 432 | 831.1K |
| Blinding Lights | Generally popular | 792 | 1.3MM |
| Dance Monkey | Generally popular | 424 | 6.7MM |
| Roses (Imanbek remix) | Generally popular | 1040 | 2.6MM |
| Lottery | TikTok popular | 2230 | 28.8MM |
| Relationship (Young Thug, ft. Future) | TikTok popular | 1349 | 27.6MM |

**Table B1.** Summary of the data

[13] The count of total TikToks featuring each song is publicly available and not part of our dataset.

**Appendix C. Information on the publicly available dataset**

The original data structure of our dataset was a JSON object that included 74 variables. We converted this data structure into a "tidy" data frame comprising 23 of the more substantive variables. Some of the variables that could prove to be interesting to researchers are:

- `music_id`: Unique ID for TikTok song.
- `tiktok_id`: Unique ID for each post.
- `desc`: caption for TikTok (often also includes hashtags, which also is a variable).
- `user_avatar`: A user's profile avatar (often a cartoon).
- `user_verified`: Boolean, denoting whether a user profile is verified.
- `digg_count`: number of likes for the post.
- `share_count`: number of shares of the post.
- `comment_count`: number of comments on post.
- `play_count`: number of plays of post.
- `following_user`: number of user followers.
- `user_hearts`: number of hearts accumulated by user.
- `user_tiktoks`: number of TikToks created by user.
- `user_fans`: number of fans of user.

Researchers interested in some of the 74 variables not included in the publicly available dataset (but available from public TikToks) should contact the corresponding author. The `user_id` and `s_user_id` (secret user ID) variables have been removed in the publicly available dataset to preserve some privacy (though the user accounts are all public). Please note that all 74 variables for a TikTok can also be recovered using the unique `tiktok_id`.