# Understanding and predicting the success lifecycle of an influencer

**Master Thesis**
**Date of Submission:** 02.03.2023


**Author**
Zehra Turgut

**Matriculation Number:** 20-743-076

**Email:** zehra.turgut@uzh.ch


Submitted in partial fulfillment of the requirements for the degree of Master of Science in Informatics, Major: Data Science


**Visualization and MultiMedia Lab**
**Department of Informatics**
&
**Department of Business Administration Chair for Marketing and Market Research**
**University of Zurich**

**Supervisors:**
Dr. Alexandra Diehl
&
Dr. Manuel S. Mariani


---

University of Zurich[UZH]

VISUALIZATIONANDMULTIMEDIALAB

## Acknowledgements

First, I would like to express my special thanks to Dr. Manuel S. Mariani for the enormous guidance throughout the thesis by introducing some methods used in the literature, providing useful feedback, and encouraging me to achieve better. My gratitude is extended to Dr. Alexandra Diehl who assisted in deciding on the interactive visualization tool, gave fruitful advice about how to present the results to the general audience, supplied valuable feedback, and motivated me to complete the thesis successfully. Similarly, I would like to sincerely thank Prof. Dr. Renato Pajarola for granting me the chance to work on my thesis in collaboration with his group-the Visualization and Multimedia Lab of the Department of Informatics- and Marketing and Market Research Group of the Department of Business Administration Chair.

Second, I could not have undertaken this journey without the continuous support and patience of my family abroad during my master's studies and especially my master's thesis.

Lastly, throughout my master's studies, I would like to mention that I feel quite lucky to have had moral support from the Gürel and Turhan families, my classmates at the University of Zurich (UZH) and Swiss Federal Institute of Technology Zurich (ETH), the Erasmus Student Network Zurich team, of which I have been an active member and from Adam whom I am also particularly grateful for his help with the language check of my thesis, and for providing some useful insights on writing and presenting a thesis.

# Abstract

Although influencer marketing has been experiencing tremendous growth, and many firms have been trying to find the right influencers for their marketing campaigns, there has been little research on identifying the timing of the high-impact work or analyzing the hot streak periods of influencers on social media. My thesis is among the first works in this area on the TikTok platform where I applied some methods used in previous literature. All the analyses were made using the video-level popularity metrics of TikTok: "number of likes", "number of shares", "number of plays", and "number of comments". In addition, unlike other studies, I created the visualization of the hot streak durations of TikTok authors to analyze information such as the start, and end time of hot streaks, or the length of hot streaks for each author. All plots for the analyses were built with bokeh- a Python library for interactive data visualizations. In conclusion, with the dataset in my thesis study, first I found that the timing of the most popular video is random in TikTok users' lifecycles. Second, I discovered that the timing of the biggest hit and the second hit are close to each other, so TikTok authors may experience average success close to their most popular videos. Third, I detected a pattern indicating that TikTok authors use more diverse hashtags during hot streak periods than before them. Lastly, I observed that the relative hot streak length -success duration- of an author is usually between 12% and 50%. As a result, marketers can use these observations in my thesis to identify successful influencers on the TikTok platform. Finally, the formulas and visualization methods used in my thesis can be applied to a larger TikTok dataset or other datasets from other social media platforms such as Instagram, or YouTube.

# Contents

# 1 Introduction

Social media platforms have been an alternative advertisement hub for brands to convince customers to buy their products. The popular users of social media who are taken as an example by other users in terms of their lifestyles, the fashion they follow, etc. are called influencers [3]. The brands tend to reach these influencers to shape consumers' purchase behavior because these popular users are evaluated as more sincere and reliable than famous people on social media [3].

Additionally, tremendous growth has been observed in influencer marketing -more than 100 percent- between the years 2019 and 2021 [9]. It is estimated that in 8 years, the capital outlay in influencer marketing will increase by about 70 million dollars [3]. As stated by Jacob Goldenberg et al., companies are trying to find impactful influencers to expand their customer reach and to affect the correct segment of people with their advertisements [9]. Although influencer marketing on TikTok was conducted by 42% of the US brands in 2021 and about 60% of the marketers try to get in touch with customers through social media, little research has been made about influencer marketing on TikTok (in terms of understanding the users' reactions to this type of advertisements) compared to other social media channels such as YouTube, Twitter, Facebook, and Instagram [3].

## 1.1 Emergence and main features of TikTok as a popular social network

Jacob Goldenberg et al.'s study points out that outbound activities such as musical performances make user engagement possible and therefore are useful to increase the follower amount [9]. Also, sharing videos and generating them has been popular in social media, from short to long ones broadcasted live [1]. For instance, stories feature on Instagram and Facebook. Another instance was Musical.ly, which received the highest download amount in 2015 among other apps in the App store. Musical.ly is a Chinese social media platform that allows users to create short lip-syncing or dance videos with options to select among various sounds and change the pace of the videos [1].

Subsequently, another app called TikTok with similar features emerged. TikTok, which was originally introduced as Douyin in China in 2016 by a Chinese firm called ByteDance, spread to the rest of the world after Musical.ly was acquired by ByteDance and joined TikTok in 2018 [1, 5]. TikTok received the second-highest download amount among other apps in 2019 -176 million downloads [1], and in the following years, it received the highest downloads among other apps in 2020, 2021, and 2022 [6, 11]. According to the research of Cervi, by 2021, TikTok had 800 million users, most of whom were below 24, and was becoming more popular among kids from 10 to 12 [5]. Also, it was stated that the majority of the most popular users on the app are below 25 years old [6]. That is, the TikTok users mostly represent generation Z, which corresponds to the birth year range of 1996 to 2010 and which is expected to be the majority population in the world [5]. A recent analysis showed that while most users (47%) are below 29, an almost equal percentage of users (41%) are in the age range 30–49

[11]. Therefore, it has reached a broad audience in less than a year. The highest number of users belongs to the countries United States, Indonesia, and Brazil, from high to low, respectively [11].

Additionally, Cervi's study states that neurobiologists have discovered that the visual cortex of generation Z is developed in a way that enables them to be more responsive to visuals and that allows a short period of concentration [5]. TikTok appeals to generation Z's mental ability because, besides videos with lengths up to 10 min [11], it provides users to create short videos of 3–15 s, or "looping videos of 3 - 60 s" [5]. It also inspired other platforms to offer short video features. For instance, YouTube has Shorts and Instagram has Reels [11].

Research on using smartphones in the US discovered that American people spend short periods on their mobile phones, but the total period is 2 h 40 min per day [6]. Facca et al.'s study claims that applications and media channels that offer fun besides informative content are more popular among young people [6]. They add that in this regard, TikTok provides an easy-to-use platform to watch enjoyable and informative videos, generate these videos, and establish a network through digital tools within easy reach like mobile phones, tablets, or computers [6].

TikTok launched by ByteDance is an artificial intelligence (AI) product[6], which examines the interactions of users on the platform and makes estimations based on these to offer individualized social media experience and relevant advertisements [6]. Consequently, the underlying machine learning algorithm enables any user to become popular as it offers an individualized "For You" page where users can access videos according to their previous activities such as plays, likes, and comments on other videos on the platform [5]. Moreover, the videos are not listed on the "For You" page only because of their content's fame or how much they have been played [6]. Namely, even if users don't have many followers, they can reach many users and obtain followers because their videos can be seen on the "For You" page of other users [1]. These characteristics of the "For You" page are unique compared to other social media apps [6]. More specifically, TikTok's algorithm is different from those used on YouTube and Facebook because it evaluates various data about users and creates a result as if the application understands your behavior as a person [6].

Furthermore, the authors on TikTok can make their videos more prominent to appear on the "For You" page by using hashtags that connect them to other authors, other categories, challenges, and popular content [6]. In my thesis study, knowing the power of hashtags on TikTok to reach a wider audience, I investigated the usage of diverse hashtags before and during hot streak periods to understand whether there is a notable difference in the number of unique hashtags used in these durations so that the onset of a hot streak can be predicted.

As it is claimed in Jacob Goldenberg et al.'s study that user engagement is crucial to increase the follower amount and reaching potential customers through influencers, TikTok platform's duet[1], stitch[2], livestream (if the user has minimum of 1000 followers[11]) features, and challenge facility encourage user interaction. Especially other platforms don't have a duet feature like TikTok, and this is a characteristic that made the app more popular, as users could generate videos side by side with influencers or celebrities of TikTok [1]. Another characteristic

[1]Duets in TikTok are videos that users show another user's video beside their video with similar content or style.
[2]Stitch is creating a video by combining parts of other videos of other users.

that made this app gain popularity is that the video pace can be adjusted [1].

Moreover, Facca et al.'s research indicates that an author should have some followers to make their videos appear on the "For You" pages [6]. Therefore, in the TikTok dataset harnessed in my thesis, authors without very few or no followers are less likely to experience the biggest hit or a hot streak period compared with the other authors.

## 1.2 Harnessing TikTok's impact on users' or consumers' behavior

Although TikTok has the aforementioned user participation features, and almost half of the users assert that TikTok affected their decisions of buying a product, as indicated by the hashtag $\#TikTokMadeMeBuyIt$, brand owners don't use this social media platform much [11]. It is advised that brands can generate their hashtags and challenges to attract potential customers [11]. In this way, brands can increase their visibility by having more followers through these challenges. Since generating "challenges" on the platform via a business agreement with TikTok is expensive, brands are suggested to reach influencers for their advertising campaigns [11]. For instance, Universal Pictures utilized influencers to make their movie "The House with a Clock in its Walls" popular by getting them to use a specific hashtag in a specific style in their stories. Within 6 days, this film production company acquired more than 11 000 followers, and by 2022, the videos with that hashtag were viewed more than 80 million times.

Furthermore, TikTok can be effective for users' actions outside social media. An example of this effect is mentioned in news about a high school graduate American teenage girl who caused the database of an online survey system for academic research to change substantially in terms of gender and age of users in a very short time [13]. In her video, she promoted this survey platform-Prolific, where she could earn money by filling out surveys. That video received 4.1 million views in a month making a vast number of users to be a member of Prolific and utilizing the system to earn money. Since the user profiles have been altered to a great extent on this survey platform Prolific, they couldn't represent a selection of the US population that they could before this teenage's TikTok video became popular [13].

## 1.3 Fun factor on TikTok videos

Moreover, Barta et al.'s research examines if funny content and the joy followers undergo can affect them to be influenced by certain users on TikTok by seeing them as "opinion leaders" and therefore by agreeing with their ideas and pursuing their recommendations. They conclude that unique content make users follow influencers' accounts and recommendations, so long as they provide users a fun time they make them think these accounts belong to "opinion leaders" [3].

However, they have discovered that the number of videos or more sophisticated content doesn't have much effect on the joyful time users undergo. Also, it is discovered that, unlike other social media platforms, if users post a huge number of videos on TikTok, they receive less interaction because it decreases the joy obtained from these accounts' media contents. This result is also found to be similar to the advertisements' impact when they are presented in high quantity. If the number is high, the consumers can get distracted or annoyed by a high number

of inputs [3].

Lastly, they indicate that the joy experienced by users from the influencers' posts affects users' follow decisions indirectly by perceiving the account owners as "opinion leaders" [3]. Thereby, uniqueness should be the aim instead of quality while producing content, and both fun factor and "opinion leadership" should be used in the media to increase and keep the follower amount and to boost the impact of their ideas on the viewers [3].

## 1.4 Social media marketing

In Appel et al.'s research, it is stated the simple "number of likes" on social media content has the possibility of not leveraging users' actions of purchasing products [2].

Another point is that it is much less costly for brands that are not big to use micro-influencers than famous people such as Selena Gomez [2]. They define micro-influencers as possessing a follower amount of a minimum of a few thousand up to 1 million, but whose followers are more engaged and more specific. Micro-influencers are evaluated as more reliable, and genuine compared to famous people on TikTok. Also, they are considered more sincere and to have expertise in the topics that they generate videos. These characteristics of them make the users interact with the content more, and therefore they make the brands choose micro-influencers for advertising their products [2].

It has also been discovered that micro-influencers who have fewer followers than macro-influencers are less risky choices because they are more responsive to activation by companies such as messages, and reposts. The advertising of these micro-influencers is also observed to be more efficient such that they can make the number of customers for the advertised product increase in two years [9].

Moreover, Appel et al. claim that since users like to spend more time viewing live streaming on social media than a taped video, the use of live streaming will increase and will be more common across a wide area of topics. However, people's basic needs such as rest and their life out of work may not meet the demand for live experiences on social media. Therefore, it is asserted that in the future, more sophisticated virtual influencers (in terms of the underlying AI structure) will be more popular because they would be able to respond whenever a user interacts [2].

While many marketers harness social media channels for supporting consumers to solve their issues or answering their information requests, this usage of social media is expected to be more common, individualized, and offer more preventive support in the next years [2].

Finally, one of the future predictions of Apple et al.'s research is that since podcasts and search engines working with audio have become favorites of users and the demand is increasing, the usage of voice-only content that allows interaction without hands and eyes will be more prevalent in the future [2].

## 1.5 Motivation for the thesis

With the previously described features and effects on user's purchasing behavior, since TikTok has emerged as a popular social media platform that attracted marketers to promote their brands through influencers, my study aimed to investigate whether the influencers on TikTok who are close to making their biggest hit video, or who are about to enter their hot streak periods, can be identified.

# 2 Related Work

There has been little research on identifying the timing of the high-impact work or analyzing the hot streak periods of influencers on TikTok. A recent detailed study by Garimella et al. was on examining hot streaks on social media using the Twitter dataset [8].

Garimella et al.'s study found that on Twitter the timing of the users' tweets that have the highest impact has little distance between each other[8]. This finding is similar to my study's finding which has discovered that the timing of the biggest hit video is close to the timing of the second-biggest hit video. They also investigated the existence of hot streaks for Twitter users and concluded that most of the users in their dataset experienced hot streaks [8]. In my thesis, the number of hot streaks and their lengths show that these periods of clustered success are also common for TikTok authors in the dataset used.

Moreover, Garimella et al.'s study has a finding about a regular rise of impact in small steps until the highest-impact tweet and a regular fall of impact in small steps after the highest-impact tweet. However, they do not observe this pattern in the shuffled dataset [8]. I could also examine whether there is such a pattern in the dataset of my study for TikTok users in the future. Currently, the result obtained from the plots mainly denotes the timings of the biggest and the second-biggest hit videos. The other results and comparisons to my study are stated in the following paragraphs.

**Popularity metric to calculate impact** While in this thesis study, the "number of likes", the "number of shares", the "number of comments", and the "number of plays" are used to calculate the impact of TikTok videos, Garimella et al.'s study chose retweet counts to gauge the impact of tweets [8]. Additionally, the difference between my study and their research is that they also analyze the effects of "user's network" on the impact of the tweets [8].

**Interactions from followers and non-followers** Another difference is that they discovered that hot streaks on Twitter are caused by retweeters' interactions [8]. In my thesis, we have the number of interactions of all users whether they are followers or non-followers. In future work of this thesis study on TikTok, users who interacted with videos by liking, sharing, playing, or commenting can be differentiated as followers and non-followers to understand which type of users are attracted by the videos.

**User's network effect** Garimella et al.'s study tried examining the effect of a user's network on tweets' impacts by including the follower amount information of each year. They discovered that there is a parallelism between the follower amount and the retweet amount. As a result, they assert that the users obtain the highest retweets in late careers [8]. In the TikTok dataset, there is not a follower amount information that corresponds to the timing of each video of an author, so in future work, this data can be obtained to investigate the network

effect on videos' impacts. Therefore, when this information is added to the dataset, the existence of a correlation between follower amount and number of likes, number of shares, number of plays, and number of comments can be inspected. Also, with the TikTok dataset, my study identified that the biggest impact video can occur randomly within an author's career.

**Timing of the biggest hits** Regarding the timing of the tweets, in addition to the tweet's index information in the lifecycle of a user, Garimella et al. gauged the week a tweet is published as a week distance from the timing of the first tweet of that user [8]. In my study, I used only the video index of the TikTok video of an author as the timing of the video. In future work, the week information can also be added to represent the occurrence of the video in terms of calendar timing.

When Garimella et al. inspect the strength of their results by shuffling the dataset, they observe that for the real careers, the timing of the biggest hits is close to each other unlike the ones in randomized careers [8]. Similar to my study's results, they observed that the biggest hit can occur before or after the second biggest hit [8]. However, they also noticed that this possibility is equally likely [8], unlike the possibility obtained from the TikTok dataset's distribution. The TikTok dataset gave the outcome that the biggest hit can be before or after the second-biggest hit video, but one cannot say that it is equally likely to occur before or after.

**Exploration amount during hot streaks** Furthermore, Garimella et al.'s research results show that during hot streak periods, the tweets are longer, contain more media, and are about various topics [8]. Similarly, my thesis results indicate that authors try more diverse hashtags during hot streak periods than before them. In future work, TikTok authors can be examined during hot streak periods, if the length of their videos increases during hot streak periods.

**Hot streak analysis** For the hot streak analysis, Garimella et al. try to find if there is a relationship between the position of the hot streak in a user's lifecycle and the age of the user on Twitter in terms of the time since a user's first tweet. They also examined whether there is a parallelism between the follower amount and the presence of a hot streak in a user's lifecycle [8]. In the future work of this thesis, one can show the distribution of the number of followers vs the length of the hot streak duration. Also, if the relevant data is obtained, one can identify the change in follower amount when the hot streak began and when the hot streak ended.

## 2.1 Hot streak characterization

**Number of hot streaks** Additionally, Garimella et al. determine how many hot streaks each user has in their career lifecycles on Twitter. They found that most of the users experienced less than 4 hot streaks, and a certain number of users possessed more than 10 [8]. My thesis research calculated the start of hot streaks as the earliest timing of the first three biggest hit videos, and the end of the hot streaks as the latest of the first three biggest hit videos. So, there is only one hot streak gauged per author, if there exists one. In the future work of this thesis research topic, one could also identify hot streak periods as described in Garimella et al.'s study.

**Length of hot streaks** Similarly, Garimella et al. also sort the hot streak length of each user and find the longest one. They conclude that most users experience short hot streaks with a length of at most 20 tweets [8].

In my study, we only analyze one hot streak per author, if it exists. Thereby, in future work, the length of the hot streak distributions for a TikTok dataset can also be examined.

**Position of hot streaks** As mentioned earlier, Garimella et al. discovered that hot streak timing changes according to the age of the user on Twitter. If the users are old in terms of the time of their first tweets, then it is more likely that they can experience hot streaks, than the users who are new on the platform because of the follower amount and increased knowledge. It has been observed that hot streaks can occur at any time for new users [8]. My thesis's future work can observe whether the age on the platform is an identifier of the timing of a hot streak. Currently, my research is focused on when the hot streak period starts, ends, and what its length is.

**User properties** What Garimella et al. also analyzed differently than my study is if the properties of users (age, number of followers, tweets, friends) affect the existence of hot streaks. They realized that only the follower amount has a significant effect, and when it is high, these users have hot streaks [8]. Again, the future work of my thesis can make this differentiation between users and offer comparative results to understand how age, number of followers, and the number of videos are correlated with hot streaks.

**Influence of hot streaks on popularity** In order to understand if hot streaks affect the success of social media content creators on Twitter, Garimella et al. calculated the ratio of the number of retweets during the hot streak period to the total number of retweets in the whole career. They identified that the retweet amount is not much affected by hot streak periods [8]. In my study as future work, for instance, the ratio of the number of plays during a hot streak period to the total number of plays in the whole career of a TikTok author can be examined to understand if hot streaks affect popularity to a large extent.

**Changes during the hot streak** In Garimella et al.'s study on Twitter, during the hot streak they observed a huge boost in the retweet amount, and follower amount compared to before and after hot streaks. Also, the retweet and follower amounts were higher after the hot streak than before it. They assert that the result was also like that for follower amount change because an increase in retweets increased follower count [8]. Again, as future work of this thesis topic on the TikTok dataset, one can analyze the differences between the number of plays, likes, shares, comments, and follower count before, during, and after hot streak periods.

**Popularity metric counts per follower** Since the boost in the retweet amount increases the follower count, Garimella et al. also analyze the number of retweets of each follower before, during, and after hot streak periods. The results indicate a substantial increase in the retweet count per follower during hot streaks compared with before and after, but the same ratio doesn't differ indicatively for before and after the hot streak [8]. Similarly, for instance, on the TikTok dataset, the number of plays per follower can be analyzed for each user during, before, and after hot streak periods to provide a comparison.

**Activity** Activity is defined as the number of tweets of a user including their retweets in Garimella et al.'s research. They discovered that the activity is much more during the hot streak period than the one before, or after [8]. In parallel to this analysis, future work of my thesis study can compare the number of videos produced by a TikTok author before, during, and after hot streak periods to inspect whether the activity is generally increased during a hot streak.

**Result of hot streak characterization** As mentioned in the previous paragraphs, Garimella et al.'s study concluded that hot streaks are periods in a Twitter user's lifecycle when they receive many more retweets, experience an increase in follower amount, and raise in activity [8]. In my thesis, this hot streak characterization would be part of future work since another analysis has been conducted to answer the research questions presented in my study.

## 2.2 Properties that influence hot streaks

This section will summarize the results of the analysis of Twitter data about the properties that influence hot streaks such as network (follower, and retweeter activity), content and activity (content and the number of tweets throughout the hot streak) [8] and compare it with my study to suggest any future work if there should be.

**Network** Garimella et al. proved that the number of retweets that come from new retweeters during the hot streak periods is substantially higher than the previous retweeters of that user. However, these new retweeters tend to get detached from retweeting after a short while compared to old retweeters [8]. As future work of my thesis on the TikTok dataset, the number of plays, likes, shares, and comments of followers and non-followers can be compared before, during, and after hot streak periods to identify if there is a similar pattern of interaction amount to videos, and which type of user affects the hot streak mostly, followers or non-followers.

**Content and activity** The content of tweets in terms of their length, media, hashtag, URL usage, diversity in topics, and whether they are retweets and mentions of replies were investigated in Garimella et al.'s research to understand the changes during a hot streak [8]. They discovered that the ratio of retweets to the total number of tweets was more during the hot streak periods. Also, media usage and tweet length are more, whereas the reply, and the "mention tweet" ratio to the total number of tweets is lower [8]. Similar to my findings on the TikTok dataset, they observed that the selection of diverse topics -meaning topic entropy- is high during hot streak periods [8]. In future work, one can analyze if video length changes or if authors use duet, stitch, and livestream features more throughout hot streak periods.

All in all, a high alteration has been observed in the content and activity throughout the hot streak periods, regardless of modifications in the number of followers or retweeters [8].

**Summary** As Garimella et al.'s study is analyzing the complete lifecycles of users in Twitter [8], my research also examines the TikTok dataset, which consists of the whole careers of authors. Their research is a pioneer in discovering hot streak existence in social media [8]. My thesis is among the first works to analyze the timing of the biggest hits and examine hot streaks on TikTok as a social media platform.

What is mainly different from my thesis is that Garimella et al. also focus on the impact of a user's network on the hot streaks [8]. The influence of the network can be investigated in my future work.

Moreover, Garimella et al. have similar findings to my results, such as exploration increases during the hot streaks. This is different from the results obtained from the hot streak analysis on the artists, film directors, and scientists' works by Liu et al. [16]. This outcome might be because producing content for science, and art takes

more time than generating casual videos on TikTok or posting tweets with text on Twitter.

Additionally, the hot streak length on Twitter was found to be mainly about 10–20 tweets [8], whereas, on TikTok, my thesis discovered that it was mostly less than 50 videos. In future work, hot streaks on TikTok can be gauged not as starting and ending from the index of the minimum of the first three biggest hits to the maximum of the first three biggest hits. The exact clustered success periods of an author could also be determined as in the study by Garimella et al. [8].

Furthermore, Garimella et al.'s analysis of "10 tweets before and after the most retweeted tweet" and discovering there is a regular rise and fall before and after for most of the users, provide an estimation pattern to identify the timing of the biggest hit tweet [8]. In future work, on the TikTok dataset, videos before and after the biggest hit video can be examined for each user in terms of popularity to understand if there is a common pattern of regular rise and fall around the biggest hit video.

Finally, in future work, another TikTok dataset including a huge number of authors' complete lifecycles can be utilized to understand whether the findings of this research are not limited to the datasets used, but are intrinsic to TikTok.

## 2.3 Success predictions in show business

Williams et al.'s study harnessed a dataset of actors' and actresses' careers to examine if success periods can be predicted in show business. Besides the onset of hot streaks, they examined the onset of cold streaks. Additionally, they tried estimating the most remarkable years of actors and actresses. Their results also differed in males and females showing that gender can also be an identifier to understand and predict success in show business careers [22]. My study's future work can investigate hot streaks and timing of the biggest hits for female and male authors separately to understand whether there are huge differences in popularity in TikTok in terms of gender. Similarly, cold streak periods can be inspected, if there are any, to analyze which users are in an unpopular phase for the long term.

Another difference from my study is that Williams et al. utilized a statistical learning model to determine which actors and actresses are close to their popular periods or they have already finished these periods. They can estimate the timing of the most remarkable year correctly in 85%. Additionally, they observed that this most remarkable year mainly occurs at the beginning of each career, and it is observed more in early careers for females than for males [22]. Similar to Garimella et al.' finding, they identify a regular rise in job opportunities before the most popular year and a regular fall in them after that year [8, 22]. In my study, the results show that more users experience the biggest hit in early careers, but the timing of the biggest hit is almost evenly distributed for all users, meaning it can occur randomly within a TikTok author's career lifecycle.

However, the main discrepancy between show business careers and careers on TikTok is that there are fewer opportunities for people to work in the film industry than to produce videos on TikTok [22]. Therefore, in the film industry, there are mainly one-year careers. But, on TikTok, an account can be active and it can continue being popular for many years.

## 2.4 Luck factor in creative careers

Janosov et al. measure luck's influence in the creative careers of book authors, scientists, and various roles in the movie industry [12]. For this purpose, they analyze a dataset consisting of 4 million careers between the years 1902 and 2017[12]. They discovered that in the movie industry, art directors' careers are less affected by luck than producers and composers. Another finding is that book authors experience the influence of luck more than scriptwriters. For the music industry, hip-hop and classical music are more resilient to luck influence than electronic music, and rock music. Finally, luck influences success in the scientific works on theoretical computer science and engineering less compared to the works about space science, astronomy, and political science [12].

Moreover, Janosov et al. tried to answer if collective work can help to estimate the high-impact work's timing. For this purpose, they examined whether there is a relationship between the "network position" and the "individual impact" in creative careers. They concluded that the high-impact work's timing doesn't depend on the network position of an individual, so they claim that luck increases the possibility of impact more than one's collaborations with others [12].

In future work, the methods utilized in Janosov et al.'s study to examine the effects of luck in creative careers [12] can also be applied to a TikTok dataset because TikTok authors' works are also in the creative career category. The luck's influence and impact of collaboration with other authors, or firms on the timing of the biggest hit video can be analyzed.

## 2.5 Influencer Marketing

Haenlein et al. state that in 2020 influencer marketing reached a $10 billion value in business and there is a boost in its usage by companies that conduct business-to-consumer model[10]. Moreover, Leung et al. note that a substantial amount of the advertisement industry is planning to utilize their finances on influencer marketing, and at the end of 2022, the total cost is supposed to be $16.4 billion [14]. Most remarkably, as Fowler et al. mention, during the pandemic period due to COVID-19, unlike other business areas, influencers who utilized that duration making posts about health and working with institutions such as "non-profits", "governments" as well as initiating goods, obtained positive results[7]. Besides, as Pradhan et al. mention in their review on "social media influencers (SMIs)" and "consumer engagement", in 2025, the number of users on social media platforms will be increased to about 4.41 billion while it was 3.6 billion in 2020. So, social media utilization for taking the attention of customers is evaluated as necessary. Pradhan et al. add that the marketing industry is being shaped by social media channels and their functionalities and the most used channel by marketers is Instagram, while the other popular ones are YouTube, Facebook, and Twitter[19].

However, Haenlein et al. mention that although many business areas such as travel, and food utilize influencer marketing on social media channels, most marketers don't have as much knowledge they have for other media platforms and struggle to take action in the correct way on social media for their businesses. Therefore, Haenlein et al. provide information about the highly demanding social media channels and give recommendations to companies about how to use influencer marketing. Furthermore, to determine the correct influencers for their

campaigns, they provide particular questions [10].

Similarly, seeing that there is a gap in the literature about managing company expenses on influencer marketing effectively, Leung et al. try to measure the cost efficiency of influencer marketing that is affected by the influencer, the number of followers of the influencer, and the contents of the influencer[14]. Their findings are listed in the recommendations in the related part below.

Moreover, as there is huge market growth in influencer marketing, Fowler et al. examined the literature that consisted of 150 articles in English and mostly from journals in Global North and that were published from 1999 to 2020 [7]. The recent research about influencer marketing and its future are stated throughout this section.

The aforementioned study by Haenlein et al. expresses that Generation Z, which consists of young people around 25 years old or below, has different tendencies regarding media usage from the previous generations. They prefer Netflix to TV and Spotify to the radio. Therefore, these media consumption choices also generated advertising on mobile platforms, especially on Instagram or TikTok. Furthermore, with the prevalence of these new advertising platforms, influencers began to emerge on social media channels. Some famous people have reached a follower amount exceeding 100 million on Instagram, and 40 million on TikTok. The majority of the companies' marketing funds are planned to be dedicated to influencer marketing at a minimum 10% [10].

First, Haenlein et al. talk about the significance of Instagram and TikTok in influencer marketing compared to other social media channels which were launched many years before such as Facebook, Twitter, and YouTube [10]. They give a summary of features of the prominent social media channels Facebook, Twitter, YouTube, Instagram, and TikTok. It is noted that on TikTok the users are relatively younger -around 20 years old- compared to the ones on other social media platforms and these younger users are critical of conventional advertisements. Therefore, influencer marketing takes their attention more [10].

Facebook has the highest number of users and user activities. On the other hand, Twitter, which is mainly based on text posts, has the opposite statistics. Both social media platforms are unfavorable among users compared with other platforms. Also, marketers don't prefer Facebook and Twitter for influencer marketing as much as they prefer other social media channels [10].

Furthermore, Haenlein et al. assert that one of the reasons why other environments such as Instagram and YouTube are more favorable is the difference in the aim of usage of these channels. While YouTube and Instagram, where users may not know the people they follow in person, are mainly to spend time and have fun, Facebook includes a network of people the users know, and Twitter is used to follow the news and network on Twitter is created based on similar interests about topics posted. As the network can be formed without knowing followers in person on Instagram and YouTube, it is simpler to emerge as influencers on these platforms [10].

It is noted that for the previously mentioned reasons, Instagram and TikTok are more popular for influencer marketing. In summary, a younger user profile compared to other platforms, increasing trend in user interaction, having a variety of post types mainly videos and images, and having more entertaining content than other platforms are the features that provide a better channel for influencer marketing [10]. That's why Haenlein et al.

analyze TikTok and Instagram more in their review. They exclude YouTube, which is evaluated more as a channel like Netflix than social media and it is more expensive to conduct influencer marketing with it [10].

### Prominent features of Instagram and TikTok

In terms of functionalities, it is noted that Instagram has a feed feature based on an AI algorithm, and the posts that a user might engage in are displayed on this feed page. That is, users can see posts of other users even though they do not follow them and therefore influencers might experience less interaction on their content from their followers [10].

Besides previous features such as posting pictures, and videos, in 2016 Instagram introduced the "stories" feature, which allows publishing videos of 15 to 60s that has a 24-hour lifetime [10]., but can be saved on the profile as a story history for later views. Since June 2018, Instagram has also been providing an IGTV video feature for videos of length 15 to 60 min [10].

On the other hand, TikTok is based on posting videos of 15 to 60s, but not pictures. The content on TikTok is mainly producing sounds and some particular movements with those sounds which are popular "trends" for a small period of time [10].

The major content display features of TikTok are the "Following" feed and the "For You" feed. The "Following" feed looks like the feature in Instagram where users can see the contents of the accounts they follow. "For You" feed is based on an AI algorithm and displays posts according to the predictions of what users might be interested in. However, TikTok is similar to YouTube in terms of building a network as users interact with posts of others outside their friend circles. Yet, the functionality of the algorithm, which is not transparent, makes it hard to estimate if a user will be impactful or not with a post. Finally, posts of sponsors are also supported to be shown on TikTok like on Instagram, but not as prevalent as on Instagram [10].

### Recommendations for successful influencer marketing

Recommendations to companies while conducting influencer marketing, derived from other's research, are the following:

- Every social media channel has a unique way of communication, and companies should consider these characteristics while producing content on these platforms. Moreover, companies should first determine the information to be presented and identify the user segment to present this information. After this phase, they should decide on an adequate social media channel to post their campaigns. Firms also need to collaborate with influencers to learn the features of the social media channels and the user segments, so that more efficient communication can be established for their campaigns. Moreover, firms should encourage the interaction of the users on their posts to attract their attention to their brands [10].
- Studies in advertising show that consumers should be faced with repeated advertisements in a combination of social media and other forms like printed ones, or e-mail. Moreover, influencers should not advertise products of brands that are rivals with each other in a short period, and this is provided with agreements that last for a long time. Lastly, picking up the adequate influencer is significant because accessing the

correct user segment through the selected influencer and the influencer's usage of the brand should be real, so companies should make the selected influencer know and embrace their brands or products [10].

- Although studies advise reaching the most popular users on social media in terms of network, it is observed that influencers with fewer connections have more interactions compared to their follower amount and have a more balanced follower segment. Therefore, it is recommended to keep in mind the following information while deciding on whether to reach mega, macro, micro, and nano influencers -the concepts that were generated by marketers [10]:
  - If the aim of the advertisement is for a national or international brand, big influencers that have already been known should be chosen [10].
  - However, if the aim of the advertisement is to be impactful in a particular region or to reach a relevant customer segment, smaller influencers should be selected [10].
  - Leading many influencers can be expensive because smaller influencers might not have managers for operating the engagement. Therefore, it should be decided whether to choose influencers whose follower amount exceeds one million and therefore who have managers [10].
  - Moreover, smaller influencers may not be experts in generating unique posts with high standards. So, firms cannot use their posts in other settings, and posts from other small influencers in the same area will look like each other. Since they are posting in the same area, some of their followers might be the same. Therefore, the advertisements through these influencers cannot reach a wide audience [10].
  - While observing the cost efficiency of influencer marketing on Weibo, Leung et al. state that an increase in the number of followers positively affects the cost efficiency [14]. Therefore, in terms of budget planning, working with an influencer with a large number of followers would be profitable.
  - Fowler et al. explored that a study in 2020 states that there is not much impact difference by using micro or meso-influencers where micro-influencers are the ones with less than 10,000 followers and meso-influencers are the ones with a follower amount between 10000 and 1 million[7].

    On the other hand, Pradhan et al. mention a definition in another study stating that in terms of the number of followers, SMIs can be divided into five categories where celebrity or mega-influencer has more than 1 million followers, macro-influencer has between 100,000 and 1 million followers, micro-influencer has between 10,000 and 100,000 followers, and nano-influencers has less than 10,000 followers. Pradhan et al. also state that in some studies, SMIs are called micro-celebrities and actual celebrities are called macro-celebrities[19]. In the TikTok dataset used in this study, the user types and the corresponding number of users to these types are listed in Table 2.1 to understand whether these users can be influencers in terms of their follower amounts and what type of influencers they can be.
  - Fowler et al. further mention another research in 2021 that claims that since users evaluate that small influencers are more trustworthy, these influencers' product advertisements can be beneficial. Similarly, a study in 2020 also asserts that small influencers can create the desired impact [7].

  As a result, in many cases, selecting small and big influencers in different numbers according to the marketing campaign is advised [10].

- Too much direction about how their products are communicated through the influencers on TikTok and Instagram is not advised. A lot of direction prevents unique content from being generated, and followers can feel detached from the influencers if they change their style of posting substantially due to a marketing campaign's strict requirements. Thereby, influencers' posts should be ratified by firms before they are posted. In this way, there can be a control mechanism for spreading information correctly without affecting the uniqueness of the content. But, there should be some set of rules to guide the influencers to direct the attention of users on the brand advertised [10].

- The study by Leung et al. recommends that advertising "new brands" through influencers decreases cost efficiency because followers find trying new products risky and not yet popular enough[14]. Thereby, firms should use different strategies for old and new brands while making their advertisements and take the risk of not achieving much if they introduce new products using influencer marketing.

| **Influencer Type** | **Count** |
|---|---|
| nano | 417 |
| micro | 149 |
| macro | 125 |
| mega | 69 |

Table 2.1: Number of authors in the TikTok dataset that corresponds to the influencer type according to the number of followers

### Recommendations for identifying the influencers

Moreover, recommendations to the companies trying to find the correct influencers are the following:

- Haenlein et al. state that determining the correct influencers plays a significant role in any advertising through influencers. Social media agencies identify influencers according to the follower amount and the interaction percentage it receives. The number of followers that makes a user an influencer also depends on the area the user is posting. Moreover, considering the budget, firms need to find an influencer, which can work professionally and in the long term [10].

- It is noted that influencers who are more popular than others were also the ones whom people relied on in their offline lives for recommendations on specific topics before they become influencers on social media. Thereby, firms need to determine such influencers with potential interests in the area they are operating in [10].

- Influencers who are more original and establish trustworthy communication are more impactful [10]. Leung et al. also have a parallel result in their study stating that the uniqueness of the influencer increases the cost efficiency of influencer marketing. Moreover, they claim that both advantages and disadvantages of a product should be communicated, to be interpreted as more credible by the followers because the followers are aware that influencers receive money for their promotions [14]. Lastly, according to the existing research's recommendations, Fowler et al. mention that generating content in a sincere style can create a good impact on the followers about the influencer and the products they advertise. But, they add

that to be more sincere, if influencers show their family members or close friends in the posts, this style can generate both negative and positive results, so the amount of inclusion of this close network circle in the posts should be handled carefully. The research in this area still continues [7].

- Influencers who are active on various platforms should be selected to be able to make use of different advertising styles of social media channels and to be able to make an impact on a wider audience [10].
- Also, followers find it more sincere if they can meet the influencers in live settings, so firms are also asking influencers to organize such events [10].
- Another point that is supposed to be more common in the future is using a post that allows direct sales on the social media app [10]. This is also a parallel finding of the study by Leung et al., which claims that content including sponsor information in a more outstanding way such as with links, increases customer interaction and therefore the cost efficiency of influencer marketing [14].
- Leung et al. further assert that the productivity of the influencer should be at a middle level, but not in extreme numbers for influencer marketing to be more cost-efficient [14].
- Lastly, Leung et al. recommend that influencers should give a reason to the followers to keep following and they should behave like leaders [14].

### Future of influencer marketing

As far as the future is concerned, TikTok's algorithm, which determines a post's popularity, is not transparent, so influencers may not predict if they will be successful with their content or not [10].

Moreover, TikTok is unique in its usage across countries because it is the only app that is prevalent worldwide and in China -although under another name Douyin with additional national controls [10]. But, political tensions between China and other countries make TikTok prohibited by these opposing countries to China. However, Instagram has been incorporating TikTok's features in its app specifications, so the features would exist if TikTok gets removed as an app completely [10].

With the increase in using online platforms for sales, firms may harness online advertising more as their next plans and that would accelerate the utilization of influencers for marketing products. However, if people quit their jobs and therefore have less budget to spend on brands advertised by influencers, influencer marketing may lose its popularity and impact on the consumers [10].

Leung et al. suggest that if companies select influencers with a large number of followers, who produce unique posts, evaluate both positive and negative aspects of the brands in their posts, prefer posting about old brands, and use interactivity such as links more while providing sponsor information, these firms can obtain about 16.6% more customer interaction for their brands [14]. However, Leung et al. utilized Weibo-a social media channel with features like Twitter- to make the relevant measurements. Also, their dataset obtained from Weibo included content for more than a month [14]. Therefore, in future research, the cost efficiency measurements can be conducted on a TikTok dataset encompassing longer user lifecycles to obtain more robust results and to test whether

the evaluations are also valid for short video-sharing platforms.

Another possibility for future research Fowler et al. have identified is examining the effects of the collaboration between influencers on the followers and on making impactful advertisements of companies' products [7]. Furthermore, they suggest that there could be a study to observe the impact when influencers provide and encourage heterogeneity through their social media profiles by following a variety of users [7].

Additionally, Fowler et al. mention the finding of "The Associative Network Model of Memory", which suggests that users store the association between influencers and the products they advertise in their memory. Therefore, it is noted that if an influencer uses another influencer in their posts about the product advertisement, the users may also associate the guest influencer in the post with the brand too. Therefore, the companies may obtain more impact on users for their brands, which are assumed to be promoted by more than one influencer. However, it is suggested that there is a need to investigate this possibility of impact with further research. Similarly, it is proposed that companies should also be cautious about the guest influencer's activities on social media not to reduce their brand's awareness or popularity. Lastly, it is recommended that there should be more studies on observing whether negative interactions during the influencer collaboration affect the users' interest in the brands and companies' future contracts with these influencers in a less favorable way [7].

Another research area indicated by Fowler et al. is how the marketers or companies take the attention of influencers should be analyzed to understand which style might receive more reactions and what are the effects of these different communication styles of companies on the users' perception of the brands [7].

Regarding the influencers who work with many brands, recent research claims that it is not perceived positively by consumers if influencers support an excessive number of brands and take attention to the brand excessively in their posts. However, there should also be research on understanding if there are negative effects regarding this excessive use on evaluating an influencer's knowledge in his or her specific area and if users lose their credibility [7].

On the other hand, the study by Fowler et al. revealed the results of the *Influencer Intelligence Report* stating that more than three-fourths of the brands utilize more than three influencers, and less than one-fifth of brands utilise more than 100 influencers. There are concerns about losing the followers' interests by using an excessive amount of influencers. Therefore, it is advised that there should be research about the ideal number of influencers a brand should use and what the threshold value of influencer amount should be [7].

Besides, Fowler et al. mention that either users or influencers are analyzed in nine-tenths of the studies they examined. Therefore, they suggest there should be studies on how to establish the correct method while making advertisements through influencers, communicating with them, and finding a way to calculate the benefits after spending a budget for influencer marketing [7]. My study sheds light on this new research gap with a different aim, as it focuses on showing a way to help marketers find the influencers who are in their hot streak periods or who have achieved their biggest hits.

Moreover, as indicated by Fowler et al., while research on the TikTok platform is not prevalent, studies on blogs, Instagram, YouTube, and Twitter are more common. It is noted that although Facebook, YouTube, What-

sApp, and Instagram are the social media channels with a great number of active users each month from high to low, respectively, they are not the most commonly utilized platforms for influencer marketing. Instagram has almost half of the active users compared with Facebook, but marketers use Instagram for more than nine-tenths of their advertisements through influencer marketing. Therefore, it is suggested that social media channels, which are both popular among users and influencers, should be studied. In addition, the usage of many social media channels by one influencer should be examined to understand how different platforms affect the impact of marketing campaigns through influencers [7]. Since there is a research gap for influencer marketing on TikTok, although its monthly active user amount is close to the number of monthly active users of Instagram [7], my thesis aimed to contribute to building a framework for marketers to choose the successful influencers by analyzing their complete lifecycles.

In addition, Pradhan et al. claim that when studies about "SMIs and consumer engagement" between 2012 and 2021 are investigated, the trustworthiness of the influencer, the interaction that the influencer set up with the followers as if the followers know the influencer in person, making sponsor information more outstanding or not in the posts, choosing influencers who display characteristics representing their followers -"identity similarity", influencers who make posts not much different from the brand's vision that they advertise affect the consumers' or users' interests on the influencers posts and their willingness to continue following them. They also assert that due to these factors, numerous firms work with influencers per advertisement, but not long term, so that they can find influencers whose posts accommodate more to the brand's vision and whose characteristics represent their followers. In that respect, they add that micro-influencers are more successful for return on investments because they can have more "identity similarity". However, it is noted that if it is a big brand, macro-influencers with followers exceeding 1 million are proposed to be more profitable according to a study in 2020 [19].

Nevertheless, the review study by Pradhan et al. suggests that other social media channels such as YouTube, Facebook, and Twitter should also be investigated to find what affects users' engagement with influencers and if there are similarities with the recent findings on Instagram, which have been the social media channel researched substantially, but which still is not enough to construct a common framework about SMIs and consumer engagement [19].

In my thesis, I used TikTok -a less studied platform- to observe and analyze the successful periods of social media users and to investigate whether the timing of their biggest hit and hot streak periods can be predicted. However, in future studies, separate datasets can be constructed containing different types of users- nano, micro, macro, and mega influencers, and similar observations, and analyses can be conducted on them to distinguish if there are different success patterns according to the influencer type.

According to the previously mentioned related work, another future work could be analyzing if the timing of a brand's hashtag usage by more than one influencer on TikTok is close to each other, then this can imply that the brand is working with more than one influencer at the same time. So, it can be explored which type of influencers obtain the most interaction with which posts, whether there is an increase in the number of followers they have after these campaigns, and whether the related posts start a hot streak period for the influencers. Moreover, it can be analyzed whether there is more consumer interaction when brands use few or many influencers for their

marketing campaigns through social media.

In the following chapters, first I mention the existing problem of finding successful influencers on social media, as well as the research questions addressed in my thesis in the “Problem Statement” part 3. Second, I describe the methods applied from other studies in the “Technical Solution” chapter 4. Third, I define the input, output, and functionality of the code files that were used to build the plots for the data analysis in the “Implementation” chapter 5. Fourth, the plots and their interpretations are stated in the “Results” chapter 6. Fifth, the findings of the thesis are summarized in the “Conclusion” chapter 7. Finally, I propose some future directions for further study and mention the limitations identified in my research in the “Discussion and Future Work” chapter 8.

# 3 Problem Statement

As it is stated in the related works section, there has not been much research about what affects social media content creators to make the highest impact through their complete lifecycles and if the time of their most impactful works can be predicted beforehand.

In order to reduce the gap in this research area, my study has aimed to investigate the timing of the TikTok influencers' most popular videos according to the popularity metrics such as the "number of likes" a video receives. Additionally, the distance between the second and first biggest hit is identified to understand if average success comes before or after the biggest hit. In this way, the influencers who are on the rise or fall of their career lifecycles could be discovered. Moreover, the possibility of the effect of using different topics in the previous videos that lead to a hit video could have been examined. All the analysis that is dedicated to understanding the success lifecycle of an influencer could be visualized with an interactive visualization tool so that the results could be more tangible, easier to read, and could be further studied in detail.

To study these goals, my research has been divided into sub-research topics, with various questions aiming to add to the literature, both empirically and theoretically.

- First, I tried to determine when an influencer's biggest hit occurs. The related research question was: Do influencers' biggest hits tend to occur early, late, or at random along their lifecycles and can the method utilized by Sinatra et al. for scientists' careers [21] be used to investigate this problem for TikTok influencers?
- Second, this study tried to respond to the question: "Do influencers experience average success before and after their hot streak periods similar to the success pattern as in the careers of scientists, artists, and movie creators analyzed by Liu et al. [15]?"
- Third, I investigated the onset of a hot streak and tried to detect whether a pattern of exploration or exploitation of topics exists before and during the hot streaks. The question here was "Do exploration and exploitation patterns identified during the success lifecycles of scientists, artists, and movie directors analyzed by Liu et al. [16] also exist for the hot streak periods in the influencers' careers?"
- Fourth, I aimed to visualize hot streak durations and big impact work timing of TikTok users with an interactive visualization tool.

These questions are addressed in sections -Timing of an influencer's hit 4.2, Hot streak detection 4.3, Understanding the onset of a hot streak 4.4- of the Technical Solution chapter 4.

# 4 Technical Solution

After an overview of "the dataset", "the data cleaning process" and "the preparation of a specific data subset" has been provided in the Data Characterization section 4.1, in the subsequent sections 4.2, 4.3, 4.4, 4.5 I specify the methods utilized to answer each research question, the relevant data as input for each method, the technologies harnessed to achieve the results, and the approach to construct the interactive visualizations.

## 4.1 Data characterization

In this initial phase of the study, first, an overview of the dataset used for this research is presented in section 4.1.1. Second, in the following subsections 4.1.2 and 4.1.3, the summary and outcomes of the data cleaning process, and an overview of the preparation of the data subset to be used for the methods in the later phases of the research are included, respectively. Further details about the dataset, like user level and video level popularity metrics' distributions are explained in the Results 6 chapter. Lastly, how the duplicates have been cleaned in the dataset, how the specific data subset has been constructed, and how the plots of the user level and video level popularity metrics have been built are described in the Implementation chapter 5.

### 4.1.1 Overview of the dataset

As a source of data of influencers from social media, TikTok data is harnessed to examine the research questions proposed in this thesis. The aforementioned TikTok data has been constructed by a preceding collaborative study of the Marketing Research Group at the UZH. It consists of 760 randomly chosen authors-users and 101 248 videos produced by these users through their complete lifecycles. The focus of the data collection had been gathering the data of content creators who have uploaded videos with dance moves. The dataset encompasses videos from various categories as well. All in all, the videos in the whole dataset-101 248 videos- belong to 760 authors and consists of 15 unique categories.

### 4.1.2 Data cleaning

At the beginning of my study, the data were analyzed with a data wrangling tool called OpenRefine [18], and duplicates have been identified in the dataset with the date column filtered for each user-author. One example of duplicate data and how it is displayed on OpenRefine is shown in Figure 4.1. This data cleaning was implemented using Python [20]. After the data cleaning for each author, the total number of videos have been reduced by 18 669, while the number of authors -760- and the number of unique categories -15- have remained the same.

The results after the data cleaning are shown in Figure 4.2, and the number of videos in each category sorted by the number of videos in descending order is displayed in Figure 4.3. In the latest version of the dataset, there

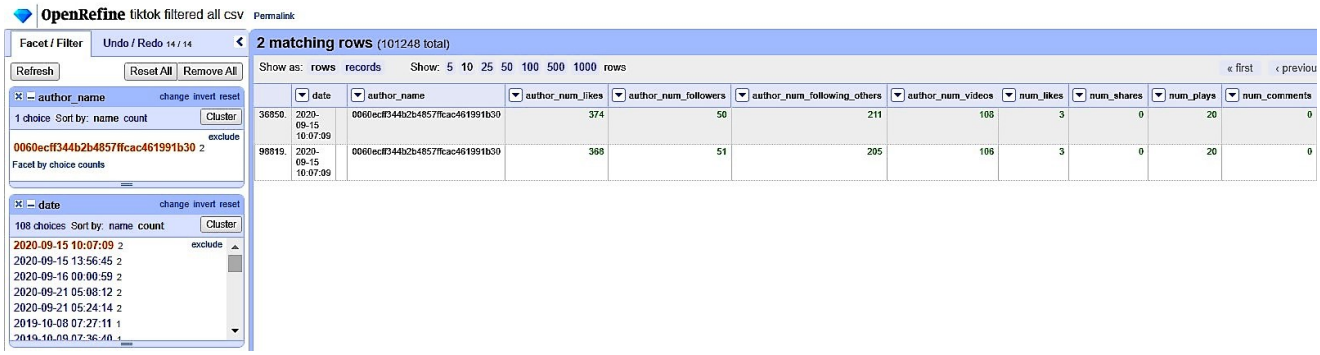



Figure 4.1: Example of duplicate data in the dataset for an author.

| | number of authors | total number of videos | number of unique categories |
|---|---|---|---|
| **Before cleaning the data set** | 760 | 101248 | 15 |
| **After cleaning the data set** | 760 | 82579 | 15 |

Figure 4.2: Number of videos, categories and authors before and after data cleaning.

are 82 579 videos, and the number of videos each author has a range from 5 to 1460. The complete lifecycles of the authors consisted of 3 months to 5 years between the dates 8th of July 2015 and 13th of July 2021, as seen in Figures 4.4 and 4.5. The implementation of data cleaning is described in the Implementation chapter 5.

### 4.1.3 Constructing a subset of the data

After the duplicate removal was completed, a data subset was constructed to use as input for the methods that would be applied in the subsequent parts of the study. As a result, the user level and video-level popularity metrics defined in the Appendix section 9.1 have been calculated on Python [20] for each 760 user's videos. The steps followed to calculate these metrics and how the data subset has been constructed are described in the Implementation chapter 5.

## 4.2 Timing of an influencer's hit

This section explains the method applied to answer the first research question: Do influencers' biggest hits tend to occur early, late, or at random along their lifecycles and can the method utilized by Sinatra et al. [21] for scientists' careers be replicated to investigate this problem for TikTok influencers? Additionally, the data used to process the method are specified.

### 4.2.1 Method

At one phase of the study, Sinatra et al. investigated the impact of the publications of 2887 scientists who have been publishing for a minimum of 20 years, have a minimum of 10 publications and published at least one paper per 5-year period. According to their specific measurement, the number of citations a publication has received 10 years after it was published creates the impact value of this publication [21].

A method Sinatra et al. used to examine the timing of the scientists hit is described below [21].

| Categories | Number of videos |
|---|---|
| misc | 74281 |
| car | 2170 |
| outdoor | 1281 |
| dance | 1012 |
| animal | 724 |
| bike | 676 |
| cooking | 516 |
| beauty | 366 |
| toddler | 297 |
| sport | 286 |
| art | 278 |
| prank | 215 |
| lifehacks | 195 |
| fingerdance | 158 |
| fashion | 124 |

Figure 4.3: Total number of videos for each category sorted by number of videos in descending order.

| | author_name | lifecycle_start | lifecycle_end | lifecycle_duration_months |
|---|---|---|---|---|
| 0 | 97db72ea8bc8b6e94ee7f6aaaa45a4c1 | 2015-07-08 22:57:52 | 2020-10-04 21:55:27 | 63.83 |
| 1 | 005bca08a3d809fe44dd7cc8aae3878a | 2016-06-09 22:47:45 | 2021-06-24 22:16:20 | 61.37 |
| 2 | a296e06324e458b01a6c1dc3f6cf5d97 | 2015-09-24 13:14:25 | 2020-10-07 16:09:44 | 61.34 |
| 3 | 36721f8fda4b44479ad00840d4141bdc | 2016-01-09 00:05:15 | 2020-10-06 18:47:56 | 57.76 |
| 4 | 83719f1bc949a3cd3e7e48b16ecc71a7 | 2016-09-21 00:38:32 | 2021-06-19 13:51:02 | 57.75 |
| ... | ... | ... | ... | ... |
| 755 | 6ffa98bd622fc877c146e7f309b5b5d2 | 2020-07-06 03:30:20 | 2020-10-06 04:42:04 | 3.07 |
| 756 | 1098fec935d8eb40d9de0589ee2887a7 | 2020-07-12 06:08:45 | 2020-10-11 10:51:54 | 3.04 |
| 757 | 0ffb373b885f3b9199e95f481e987e47 | 2020-06-29 10:15:19 | 2020-09-28 01:04:35 | 3.02 |
| 758 | df50b91a57c4c29ff9817799accb53db | 2020-07-15 08:40:40 | 2020-10-13 10:58:05 | 3.00 |
| 759 | b1721c693b472fd24841bbbc0effca21 | 2020-07-11 16:46:41 | 2020-10-09 17:53:31 | 3.00 |

760 rows × 4 columns

Figure 4.4: Lifecycle start, and end time, and the duration (in months) of the lifecycle of each author at the time the TikTok dataset was generated.

| lifecycle_start | lifecycle_end |
|---|---|
| 2015-07-08 22:57:52 | 2021-07-13 11:33:39 |

Figure 4.5: Minimum lifecycle start and maximum lifecycle end time in the TikTok dataset at the time it was generated.

First, Sinatra et al. define the following parameters to use later while exploring the timing of the biggest impact works [21]:

**N:** Total number of works of a scientist

**N*:** The index of the highest impact work of a scientist

**N*/N:** The relative index of the highest impact work of a scientist

Then, Sinatra et al. plot the probability of N*/N and the complementary cumulative distribution (CCD) of N*/N. The plots clearly show that through the career of a scientist, the scientist can have the highest impact work at any time -meaning highest impact work can occur randomly through a scientist's career lifecycle [21].

In the other phase of their study, Sinatra et al. shuffle the impact of work for each scientist in the dataset to create a randomized dataset and then calculate the probability of N*/N, and the complementary cumulative distribution function (CCDF) of N*/N with the shuffled dataset. They find that there is no significant difference between randomized and real careers [21].

Another analysis in Sinatra et al.'s research shows that the increase in the productivity of a scientist lead this individual to come up with the highest impact work. Therefore, they conclude that there is a correlation between productivity in a career and the impact of work [21].

### 4.2.2 Applying the method on TikTok dataset

In this thesis's study, the TikTok dataset offers video level popularity metrics that can show the impact of a video. To replicate the method of Sinatra et al. [21], this thesis uses these metrics defined below to identify the timing of a TikTok influencer's biggest hit. The corresponding probability and CCD plots are shown in the Results chapter 6. In the second part of the analysis, the dataset was shuffled for each of these parameters to create a dataset of randomized careers. Lastly, a statistical test called Mann-Whitney U Test was applied to compare the distribution of N*/N for the original dataset-representing real careers and the distribution of N*/N for the shuffled dataset-representing randomized careers.

#### Video level popularity metrics

**num_likes:** Number of likes each video of an author has received

**num_shares:** Number of shares each video of an author has received

**num_plays:** Number of plays each video of an author has received

**num_comments:** Number of comments each video of an author has received

The Mann-Whitney U Test compares any two given datasets and gives an output called "p_value" which shows the probability whether two datasets belong to the same distribution. If the value is above 0.05, then it can be

claimed that there are no significant differences between the two datasets: they could come from the same distribution.

In my study, all p_values were above 0.05 and mainly close to 1. Thereby, the statistical test results showed that the datasets are almost equal. According to this outcome, it can be claimed that there is not a significant difference between the timing of the highest impact work of real careers of content creators and their randomized careers in the TikTok dataset utilized.

### 4.2.3 Data

The data subset described in Appendix section 9.1.1 of the TikTok dataset was used to process the method defined in this section on the original data and to produce the plots using the bokeh library of Python for visualizations [4].

In this phase of the study, to examine the randomized careers, shuffled datasets were also produced from the original dataset. Two types of randomization have been constructed.

- One dataset for the average of the 100 times of shuffling. While producing the average of 100 times of shuffling for the video-level popularity metrics, separate datasets for each 100 times shuffling were not produced. Only one dataset containing the average of 100 times of shuffling for each video-level popularity metric was generated.
- 25 separate datasets were generated for 25 times of shuffling.

Lastly, as in the "Data characterization" phase, two types of data subsets were prepared from these shuffled datasets.

- One data subset for the average of the 100 times of shuffling.
- 25 data subsets for the 25 separate shuffled data.

These data subsets were produced to plot the distributions of the related metrics utilized in the method defined in this section. The steps for shuffling the data, preparing the datasets, building the plots are specified in the Implementation chapter 5 and the plots that were produced using the bokeh library [4] are displayed in the Results chapter 6.

## 4.3 Hot streak detection

This part of the study introduces the method used to answer the second research question: Do influencers have a similar success pattern as in the careers of scientists, artists, and movie creators analyzed by Liu et al. [15]? Namely, examining if influencers experience average success before and after their hot streak periods. The data used to process the method is also specified.

### 4.3.1 Method

A specific duration that consists of high-impact works that occur more densely than other periods in a career defines a hot streak period. Liu et al. investigated the career works of artists, film directors, and scientists and

found out that high-impact works follow each other and pop up noticeably at a time-resulting in a hot streak [15].

In one part of their study, Liu et al. discovered that the timing of the most popular productions occurs randomly in each career. To discover this outcome, they investigated every career for when first 3 of the most popular productions occurred. They defined "N" as the total number of productions and N* as the most popular work's index, N** as the 2nd most popular work's index, and N*** as the 3rd most popular work's index [15].

In another part of their research, Liu et al. found out that there is an estimated pattern in all careers for the distance between the biggest and the second biggest hit of a career. This finding was partially the result of a method they made use of [15]:

- First, Liu et al. measure $\Delta$N = N*-N**, meaning the distance between the timing of the most and the 2nd most popular work for each career [15].
- Second, Liu et al. define $\Delta$N/N as the relative distance between the timing of the most and the 2nd most popular work [15].

Liu et al. calculate these values stated above also on a randomized dataset to compare the distributions of the values with the distribution of the values from real careers. They make this comparison by observing the distribution of the equation 4.1 [15].

$$R(\frac{\Delta N}{N}) = P(\frac{\Delta N}{N})/P(\frac{\Delta Ns}{N}) \tag{4.1}$$

The distribution of 4.1 has values substantially close to zero and there is almost equal distance from both the positive and negative side of the zero point. Namely, the highest-impact work can occur before and after the second-highest-impact work in almost equal probability. Additionally, when Liu et al. replicate this investigation to check the pattern for N** and N**, and N* and N***, they find similar results. However, they add that although in real careers timing of the first two high impact works is close to each other, this is not observed in randomized careers. Therefore, the distribution of R for real careers has mainly larger values, than the one for randomized careers [15].

### 4.3.2 Applying the method on TikTok dataset

In this thesis's study, to understand if influencers experience average success before and after their hot streak periods, the previously described Liu et al.'s method that was used in career productions of artists, film directors and scientists [15], was applied to the TikTok dataset.

In the "Data characterization" phase of the study, the data subset from the TikTok dataset including the following metrics below was prepared to be used later. All the metrics and their definitions are included in Appendix 9.1.1.

From that data subset, the metrics below were used for the application of the Liu et. al.'s method for hot streak detection in TikTok users' lifecycles [15]. Namely, $\Delta$N/N was utilized for each video level popularity metric-number of likes, number of shares, number of plays, and number of comments- of each user in the dataset.

**ratio_delta_likes:** ΔN/N: Relative distance between the indexes of the videos that received the maximum "number of likes" and that received the 2nd maximum number of likes in an author's career lifecycle

**ratio_delta_shares:** ΔN/N: Relative distance between the indexes of the videos that received the maximum "number of shares" and that received the 2nd maximum number of shares in an author's career lifecycle

**ratio_delta_plays:** ΔN/N: Relative distance between the indexes of the videos that received the maximum "number of plays" and that received the 2nd maximum number of plays in an author's career lifecycle

**ratio_delta_comments:** ΔN/N: Relative distance between the indexes of the videos that received the maximum "number of comments" and that received the 2nd maximum number of comments in an author's career lifecycle

Additionally, in the Timing of an influencer's hit- 4.2.3 phase of the study, data subsets from the shuffled TikTok datasets including the following metrics below were prepared. All the metrics and their definitions were included in Appendix 9.3.

From those data subsets, the metrics below were used for the application of the [15]'s method for hot streak detection in TikTok users' lifecycles. Namely, ΔNs/N was utilized for each video level popularity metric-number of likes, number of shares, number of plays, and number of comments- of each user in the dataset.

**sh_ratio_delta_likes:** ΔN/N: Relative distance between the indexes of the videos that received the maximum "number of likes" and that received the 2nd maximum number of likes in an author's career lifecycle

**sh_ratio_delta_shares:** ΔN/N: Relative distance between the indexes of the videos that received the maximum "number of shares" and that received the 2nd maximum number of shares in an author's career lifecycle

**sh_ratio_delta_plays:** ΔN/N: Relative distance between the indexes of the videos that received the maximum "number of plays" and that received the 2nd maximum number of plays in an author's career lifecycle

**sh_ratio_delta_comments:** ΔN/N: Relative distance between the indexes of the videos that received the maximum "number of comments" and that received the 2nd maximum number of comments in an author's career lifecycle

Lastly, ΔN was calculated using equation 4.1 to compare the distribution between real and randomized careers. This calculation was conducted for both the average of the 100 shuffled dataset and the 25 separate shuffled datasets.

To observe the distribution of ΔN/N and R(Δ) for each video level popularity metric of each 760 users, histogram, probability mass function (PMF), CCDF plots were constructed in Python using the bokeh library [4] for interactive visualization.

Finally, a Mann-Whitney U Test was applied to compare the distribution of ΔN/N for the original dataset, representing real careers and the distribution of ΔNs/N for the shuffled datasets, representing randomized careers.

After the test was applied to the datasets in this study, all p_values were above 0.05 and mainly close to 1. Thereby, the statistical test results showed that the datasets (original and shuffled ones) are almost equal. According to this outcome, it can be claimed that there is not a significant difference between "the distance between the biggest and 2nd biggest hit" of real careers of content creators on social media and their randomized careers in the TikTok dataset utilized in this thesis study.

By applying the method offered in Liu et al.'s study [15] on the TikTok dataset, I tried to investigate if there is a predictable pattern in all users for the distance between the biggest and 2nd biggest hit of an influencer. So, if the distance is high and positive, the second hit comes late after the first hit. But, if the distance is high and negative, the biggest hit occurs at a later time than the second biggest hit.

Another possible outcome can be as in the study results of Liu et al. [15]: The distribution of 4.1 can be substantially close to zero, and there can be almost equal distance on both positive and negative sides of zero. Namely, the highest impact video of an influencer can occur before and after the second highest impact video at almost equal probability. The evaluation of the results of this phase-Hot streak detection- are included in the Results chapter 6.

### 4.3.3 Data

The data subset described in Appendix sections 9.1.1 and 9.3.1 was used to process the method defined in this section on the original data and shuffled data and to produce the plots using the bokeh library of Python for interactive visualizations [4].

How the outputs-plots for this phase of the study were built are specified in the Implementation chapter 5 and the plots are included and evaluated in the Results chapter 6.

## 4.4 Understanding the onset of a hot streak

This section presents the method harnessed to investigate the third research question: Do exploration and exploitation patterns identified during success lifecycles of scientists, artists, and movie directors analyzed in the study of Liu et al. [16] also exist for the hot streak periods in the influencers' careers? Additionally, the data used to process the method is specified.

### 4.4.1 Method

The research by Liu et al. has analyzed career works of artists, film directors, and scientists to observe if there is a pattern they can use to detect onset of a hot streak. In each field they have discovered that there exists a pattern for the start time of a hot streak, which is close to the change time from a period of exploration of topics to a period of exploitation of topics [16]. Exploration of topics means various themes are used in the works while exploitation of topics means similar themes are harnessed in the productions.

Liu et al. measure the entropy of the "styles or topics" in each career to identify the diversity in the productions-meaning exploration amount. The equation 4.2 has been used to calculate the entropy of topics in artists, film directors, and scientists' careers [16].

$$\hat{H} = -\sum_{i=1}^{m} pi * log(pi) \tag{4.2}$$

In this equation, the parameters are defined as the following [16]:

**m:** Number of unique styles or topics within a career

**pi:** Frequency of style or topic "i" within a career

Liu et al. further explain if a career includes only the same topic, then the equation would be $\hat{H} = 0$. But, if all the styles or topics within a career are unique, then the equation would result in $\hat{H} = log(n)$, where n is the total number of productions in a career[16].

To examine the distribution of the topic entropies -$\hat{H}$- in the whole career lifecycles, Liu et el. also normalize it by $\hat{H} = log(n)$, such that $H = \frac{\hat{H}}{log(n)}$. Also, to understand if there is a tendency for exploration or exploitation before and during hot streaks, they calculate the "H" value for the works before and during the hot streak periods[16].

Then, Liu et el. try to investigate the estimated amount of "H"-topic entropy- before and during hot streaks. For that purpose, they compare the average value of "H" calculated for the works of real careers before the hot streak with the "H" value distribution of randomized careers before the hot streak period. Additionally, they conduct the same comparison between the real and randomized careers for the "during hot streak" period. They plot the distribution of H for randomized and real careers to observe the variations [16].

The findings clearly show that the average value of H for the real careers before the hot streak is substantially higher than the estimated amount and the same value is substantially lower than the estimated amount during the

hot streak period. Thereby, Liu et al. conclude that there is a pattern in their dataset consisting of the lifeworks of artists, film directors, and scientists that indicates an exploration of topics leading to hot streaks and narrowing down topics(exploitation) during the hot streak periods [16].

### 4.4.2 Applying the method on TikTok dataset

In this thesis's study, to analyze if there is a pattern of exploitation of topics followed by an exploration of topics such that end of the exploration period coincides with the start of the hot streak period for the productions of social media creators-influencers-, the method of Liu et al.'s study [16] was applied to the TikTok dataset.

To replicate the method in Liu et al's research [16], this study uses the metrics defined in Appendix section 9.4 to investigate if there is more exploration of different topics before hot streaks than during hot streak periods. As a representation of topics in careers stated in [16], from the TikTok dataset, I used hashtags of users and calculated the hashtag entropy for video level popularity metrics- number of likes, number of shares, number of plays and number of comments for each user before and during hot streak time. Additionally, I calculated the difference between the hashtag entropy before and during hot streak periods for each video level popularity metric to observe the amount of deviation from zero to positive or negative. If the tendency of the distribution is more on the positive values, then it would be concluded that there is more exploration before hot streaks than exploitation and that there can be a pattern of exploration leading to hot streaks during which exploitation is experienced for the TikTok influencers' lifecycles.

To illustrate how equation4.2 is applied to the TikTok dataset for entropy calculations before and during hot streaks, I state an example measurement in equations 4.3, 4.4, and 4.5, 4.6. These equations harness the metrics mentioned below 4.4.2 that have been calculated as user-level metrics and video-level popularity metrics for "number of likes". Equation 4.3 and equation 4.5 measure the hashtag entropies for the number of likes metric before and during the hot streak periods, respectively. Equation 4.4 and equation 4.6 measure the normalized hashtag entropies for the number of likes metric before and during the hot streak periods, respectively.

$$\hat{Hb} = -\sum_{i=1}^{Nonset\_likes} hashtag\_freq * log(hashtag\_freq) \tag{4.3}$$

$$Hbefore\_likes = \frac{\hat{Hb}}{log(video\_total)} \tag{4.4}$$

$$\hat{Hd} = -\sum_{i=Nonset\_likes}^{Nend\_likes} hashtag\_freq * log(hashtag\_freq) \tag{4.5}$$

$$Hduring\_likes = \frac{\hat{Hd}}{log(video\_total)} \tag{4.6}$$

#### A. User level metrics:

**unique_hashtag_count:** (m): Number of unique hashtags of a user

**video_total:** (n): Total number of videos of a user

**hashtag_freq:** (pi): Hashtag frequency of a hashtag of a user

**B. Video level metrics:**

**Metrics associated with "number of likes":**

**Nonset_likes:** Minimum of the video index of the first 3 biggest hits. First video index of the hot streak period.

**Nend_likes:** Maximum of the video index of the first 3 biggest hits. Last video index of the hot streak period.

**Hbefore_likes:** (Hb_likes): Normalized hashtag entropy of a user before the hot streak period starts

**Hduring_likes:** (Hd_likes): Normalized hashtag entropy of a user during the hot streak period

These calculations were made both for real and randomized careers of TikTok influencers. Randomized careers consisted of an average of 100 times shuffled versions of the original TikTok dataset. Then, plots were constructed to compare the exploration amount in real careers with the exploration amount in randomized careers for both before the hot streak and during hot streak periods. How the metrics were calculated and how the plots were constructed are included in the Implementation chapter5 and the evaluation of the results of this phase-Understanding the onset of a hot streak- are added to the Results chapter 6.

### 4.4.3 Data

The TikTok dataset, data subset prepared in the "Data Characterization" phase and stated in Appendix section 9.1 were harnessed to produce a data subset including the metrics defined in Appendix section 9.3 for the original data.

Regarding the randomized career evaluation, 100 times shuffled TikTok dataset and its data subset constructed in "Timing of an influencer's hit" phase and stated in Appendix section 9.3 were utilized to produce a data subset including the metrics defined in Appendix section 9.3 for the shuffled data.

These new data subsets constructed both for original and shuffled data have been used to observe the distributions of hashtag entropies in order to check if a pattern of exploitation followed by exploration exists in social media content creators' lifecycles. To examine the distributions and compare them, histogram plots have been constructed using the bokeh library of Python for interactive visualizations [4]. The plots' screenshots are listed in the Results chapter 6.

## 4.5 Visualizing hot streaks of users

In all the phases of my thesis, the results were plots that show the distribution of the metrics, and these plots were built in order to characterize the TikTok users' popularity metrics 4.1, identify the timing of the biggest hits of influencers 4.2, detect the onset of hot streaks of users 4.3, understand the onset of hot streaks 4.4. For all the plots, Python's bokeh library that provides interactive visualizations was utilized [4].

Legends in each plot in my study were made interactive so that for instance, each plot for each of the 25 shuffled data can be hidden or shown to compare every one of them with the original data's plot. Hovering plot elements

to see detailed information about the coordinates and the bars were also added to present a user-friendly interface for reading plots efficiently.

In this last phase of the study, hot streak durations of users were analysed, the contents and interactivity for the figure were designed, and a specific dataset was prepared to show the desired analysis in the plots.

### 4.5.1 Method

First, the general design was arranged to show the following:

- The hot streak durations of users
- The timing of the highest, the 2nd highest, and the 3rd highest impact videos
- Displaying the above information for users who have the biggest hits either in early, middle or late career of their lifecycles

In order to visualize this design, Python [20] and its bokeh library were utilized [4]. Additionally, interactivity was arranged to offer sorting according to the following:

- Productivity of the users
- Length of hot streak duration
- Relative length of hot streak duration
- Timing(video index) of the biggest hit
- Relative timing(video index) of the biggest hit

As a result, influencers could be categorized by their career lengths and achievements in an orderly fashion.

Moreover, on this design with a hovering mechanism, users interacting with plot elements associated with each author(TikTok influencer) can see detailed information about the author. The list of parameters shown in the tooltip are as below:

- User index
- User name
- Total number of videos (N)
- Video index of the streak's start
- Video index of the streak's end
- Streak length (#videos during hot streak)
- Relative streak length
- 1st hit's time in whole career : It can be early, middle, late

- Video index of the biggest hit (N*)
- Relative video index of the biggest hit (N*/N)
- Video index of the 2nd biggest hit (N**)
- Video index of the 3rd biggest hit (N***)

The information shown in the tooltip changes according to the selection of the user on the ”Select a hit order” dropdown. For instance, when “Biggest hit” is selected, the information of “Video index of the 2nd biggest hit (N**)”, and “Video index of the 3rd biggest hit (N***)” is not displayed in the tooltip.

In order to have a better overview and analyze the figures by comparing information obtained from each phase of the study, the results of plots were placed on one page that can be accessed via the bokeh webserver on localhost.

Moreover, a user manual “visualizationManual.pdf” was prepared as a help menu of the offline web page to guide users about the purpose of each phase, which information they can access in each menu, what kind of interactivity can be experienced in figures, and the evaluation of plots as noted in the Results chapter 6 The access to the Google Drive link[1] for the manual and the bokeh code of the offline web page can be provided upon request.

Before constructing the plots, there were also measurements to gauge the duration of hot streaks for each user and display it on the figure. In this study, the hot streak period for $\#likes$, and $\#plays$ is defined as in equations 4.7, and 4.8, respectively.

$$Streak\ durationlikes = (Nend\_likes - Nonset\_likes) + 1 \tag{4.7}$$

$$Streak\ durationplays = (Nend - Nonset\_plays) + 1 \tag{4.8}$$

As formerly mentioned in section 4.4.2, “Nend_likes” is the last video index of the hot streak period, and “Nonset_likes” is the first video index of the hot streak period. In this research, the hot streak duration is measured as the total number of videos from the beginning of the hot streak till the end of it.

Furthermore, to show the timing of the influencers’ biggest hit, whether it is in the early, middle, or late career, an assumption was made to define the early, middle, and late career periods. The assumption is described as below:

- Assuming N denotes the number of videos of an influencer, an influencer’s career lifecycle was divided into these 3 equal phases, representing early, middle and late phases respectively.
  - early phase $\in [1, \frac{N}{3}]$
  - middle phase $\in (\frac{N}{3}, \frac{2*N}{3})$
  - late phase $\in [\frac{2*N}{3}, N]$
- Assuming N* is the video index of the biggest hit of an influencer, if a user’s biggest hit occurs in

[1] https://drive.google.com/drive/u/0/folders/13e5eRkoQL56DcOzqaohsEI8dG2OAylCE

- $[1, \frac{N}{4}]$, then the hit corresponds to the early career phase
- $(\frac{N}{4}, \frac{3*N}{4})$, then the hit corresponds to the middle career phase
- $[\frac{3*N}{4}, N]$, then the hit corresponds to the late career phase

This additional metric was added to the dataset as "streak_time". This metric and the "streak duration" metric defined in equations 4.7, and 4.8 are further defined in the Data subsection 4.5.2. The implementation of the figure and the results of the evaluations of the plots were appended in the Implementation 5 and Results 6 chapters, respectively.

### 4.5.2 Data

The data subset prepared during the "Understanding the onset of a hot streak" 4.4 phase including the metrics defined in Appendix section 9.4 for the original data, were utilized to produce the figure that illustrates the hot streak duration of users interactively.

In addition to that dataset's metrics, the following metrics below were calculated as explained in subsection 4.5.1 to be used while building the plots:

**streak_duration_likes:** Duration of the hot streak gauged according to the popularity metric-"number of likes". It is defined as the total number of videos from the start of the streak till the end of it.

**relative_streak_duration_likes:** Relative duration of the hot streak gauged according to the popularity metric-"number of likes". It is defined as the ratio of the total number of videos from the start of the streak till the end of it to the total number of videos.

**streak_duration_plays:** Duration of the hot streak gauged according to the popularity metric-"number of plays". It is defined as the total number of videos from the start of the streak till the end of it.

**relative_streak_duration_plays:** Relative duration of the hot streak gauged according to the popularity metric-"number of plays". It is defined as the ratio of the total number of videos from the start of the streak till the end of it to the total number of videos.

**streak_time:** The career time when a user experience the biggest hit according to the popularity metric-"number of likes", or -"number of plays". This metric can have either of these values: early, middle, or late, representing early career phase, middle career phase, or late career phase, respectively.

# 5 Implementation

In this chapter, the sections are dedicated to providing the implementation parts of each phase explained in the Technical Solution chapter 4 and they have the following structure:

- An overview of each input data (including the file name that contains it and the Google Drive path of the file) processed through the Python scripts
- Main functionality of the Python scripts, the name of the script's file, and the location path on Google Drive
- An overview of each output data (including the file name that contains it and the Google Drive path of the file) built by the Python scripts
- An overview of the plot outputs constructed by the Python scripts

All the data and code files can be accessed from the provided Google Drive paths upon request.

## 5.1 Data characterization

The implementation part of this phase, first, consisted of data cleaning, and second, data subset preparation to use the methods obtained from the relevant research and applied to the TikTok data in this thesis.

### 5.1.1 Data cleaning

**Input data (tiktok-data.xlsx):** TikTok dataset in folder[1]

**Script (1_Duplicate_Removal.ipynb):** This Jupyter notebook in folder [2] removes the duplicates in "tiktok-data.xlsx" by checking the duplicates according to the date column. After cleaning the data, it creates an index of videos for each author with a new column-video_order.

**Output data (tiktok_cleaned_data.xlsx):** Cleaned TikTok dataset in folder[1]

### 5.1.2 Data subset preparation

**Input data (tiktok_cleaned_data.xlsx):** Cleaned TikTok dataset in folder[1]

**Script (2_prepare dataframe for plots.ipynb):** This Jupyter notebook in folder [2], creates a data subset of 760 rows corresponding to each 760 author. This data subset includes new columns that stores total number of videos of each author, total number of followers, index of the video that has received the maximum number of likes, relative index of the video that has received the maximum number of likes, etc. It also

[1] https://drive.google.com/drive/u/0/folders/12yAb0op2XFm8WCcFAs6-HylptjZxN8d-
[2] https://drive.google.com/drive/u/0/folders/1j9FGbObiw5k5-xtKKKUFkRRxybJo-taP

generates two other files listing unique categories and influencer types of authors according to the number of followers. Moreover, it calculates the start and end date of the lifecycles for each author as well as their lifecycle durations.

**Output data:** Located in folder [1]

- tiktok_categories.xlsx: Unique categories in the TikTok dataset
- influencerTypeList.xlsx: Lists the authors, number of followers, influencer type according to the number of followers in the TikTok dataset
- dataFrame_from_original data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.1

### 5.1.3 Constructing plots

#### A. User level popularity distributions

**Input data (dataFrame_from_original data_for_plots.xlsx):** Data subset in location [1] and whose metrics are defined in Appendix 9.1

**Script(task1_user.py):** This Python file located in folder [3], constructs the user level popularity metrics' CCDF plots.

**Output plot(s):** Can be accessed on the "User level" tab under the "Popularity Distributions" tab by running the "bokeh serve –show home_page.py" command on a Python IDE like Visual Studio Code.

#### B. Video level popularity distributions

**Input data (dataFrame_from_original data_for_plots.xlsx):** Data subset in location [1] and whose metrics are defined in Appendix 9.1

**Script(task1_video.py):** This Python file located in folder [3], constructs the video level popularity metrics' CCDF plots.

**Output plot(s):** Can be accessed on the "Video level" tab under the "Popularity Distributions" tab by running the "bokeh serve –show home_page.py" command on a Python IDE like Visual Studio Code.

## 5.2 Timing of an influencer's hit

The implementation part of this phase, first, consisted of dataset preparation for the 25 separate shuffling and the average of shuffling 100 times. Second, plots were built that compare the distributions between the original data and the data in these two separate types of shuffling.

### 5.2.1 Preparing shuffled datasets

#### A. Preparing average of 100 times shuffling

**Input data(tiktok_cleaned_data.xlsx):** Cleaned TikTok dataset in folder [1]

[3] https://drive.google.com/drive/u/0/folders/1hvibkAcltwzdP8XrAXbaRQ4dhmmlqKUl

**Script(1_prepareData_for_shuffling100.ipynb):** This Jupyter notebook located in folder [4], creates a new dataset which is the average of 100 times shuffling the original dataset. Secondly, it creates a data subset from the "tiktok_data_after_shuffled_100times.xlsx" dataset, that includes new columns such as the total number of videos of each author, the total number of followers, index of the video that has received the maximum number of likes, relative index of the video that has received the maximum number of likes, etc.

**Output:** Located in folder [5]

- tiktok_data_after_shuffled_100times.xlsx: Contains the average of 100 times of shuffling the original dataset.
- dataFrame_from_shuffled100 data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.3

### B. Preparing 25 separate shuffled datasets

### B.1. Creating shuffled datasets from the original dataset

**Input data(tiktok_cleaned_data.xlsx):** Cleaned TikTok dataset in folder [1]

**Script(2_prepareData_for_shuffling25.ipynb):** This Jupyter notebook located in folder [4], creates 25 datasets each of which is a shuffled version of the original dataset.

**Output data:** 25 shuffled datasets located in folder [6]

- tikTok_760_shuffled1.xlsx: Contains a shuffled version of the original dataset.
- ...
- tikTok_760_shuffled25.xlsx: Contains a shuffled version of the original dataset.

### B.2. Creating data subsets from shuffled datasets

**Input data:** Shuffled datasets located in folder [6]

- tikTok_760_shuffled1.xlsx: Contains a shuffled version of the original dataset.
- ...
- tikTok_760_shuffled25.xlsx: Contains a shuffled version of the original dataset.

**Script(3_prepareRatios_for_shuffling25.ipynb):** This Jupyter notebook located in [4] creates data subsets from shuffled datasets, that include new columns such as the total number of videos of each author, the total number of followers, index of the video that has received the maximum number of likes, the relative index of the video that has received the maximum number of likes, etc.

**Output data** : Location in folder [7]

- df_ShuffledForPlots_1.xlsx: Data subset whose metrics are defined in Appendix 9.3
- ...
- df_ShuffledForPlots_25.xlsx: Data subset whose metrics are defined in Appendix 9.3

[4] https://drive.google.com/drive/u/0/folders/1Wj62J5SFl-YsRL7VVJ4CyWhwIy8RJNnf
[5] https://drive.google.com/drive/u/0/folders/1P2Gp4SA7P88Doxhfzhe8JLq6ZiCO3OYX
[6] https://drive.google.com/drive/u/0/folders/1eO1p58R4cpm4_3AO6PSBCHbAVAL7T_qu
[7] https://drive.google.com/drive/u/0/folders/1VSObjoZGOd85GsROblNMsPOUb0gjf8V9

### 5.2.2 Constructing plots

#### A. Constructing plots for original data and average of 100 times shuffled data

**Input data:** Located in [1], and [5] respectively.

- dataFrame_from_original data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.1
- dataFrame_from_shuffled100 data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.3

**Script(task2_orig_sh100.py):** Located in [8], this Jupyter notebook builds the histogram, CCDF, PMF plots of the relative index of the video of each 760 users for each video level popularity metric (#likes, #shares, #plays, #comments) respectively. Also, it applies a Mann Whitney U Test between the PMF values of the original data and the average of the 100 times shuffled data. On each PMF plot's legend, it shows the statistical test result-p_value.

**Output plot(s)** : Can be accessed on the "Plots for original and shuffled100 data" tab under the "Timing of an influencer's hit" tab by running the "bokeh serve –show home_page.py" command on a Python IDE like Visual Studio Code.

#### B. Constructing plots for original data and each of 25 shuffled data

**Input data:** Located in [1], and [7] respectively.

- dataFrame_from_original data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.1
- shuffled data subsets:
  - df_ShuffledForPlots_1.xlsx: Data subset whose metrics are defined in Appendix 9.3
  - ...
  - df_ShuffledForPlots_25.xlsx: Data subset whose metrics are defined in Appendix 9.3

**Script(task2_orig_sh25.py):** Located in [8], this Python file builds the CCDF, plots of the relative index of the video of each 760 users for each video level popularity metric (#likes, #shares, #plays, #comments) respectively. Also, it applies the Mann Whitney U Test between the PMF values of the original data and the 25 separate shuffled data. It constructs the histogram and CCDF distribution of the 25 statistical test results-p_values.

**Output plot(s)** : Can be accessed on the "Plots for original and shuffled25 data" tab under the "Timing of an influencer's hit" tab by running the "bokeh serve –show home_page.py" command on a Python IDE like Visual Studio Code.

[8] https://drive.google.com/drive/u/0/folders/1IAHxhI26myz2viUd6cvBpvjby5GmRYiI

## 5.3 Hot streak detection

In this phase, data subsets created in previous parts of the study were used and there weren't any new data subsets built as the outputs of this phase.

### 5.3.1 Constructing plots for original data and average of 100 times shuffled data

**Input data:** Located in the folders [1], and [5] respectively.

- dataFrame_from_original data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.1
- dataFrame_from_shuffled100 data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.2

**Script(task3_orig_sh100.py):** Located in [9], this Python script builds the histogram, CCDF, PMF plots of "the relative index distance between the 1st and 2nd biggest hit videos" - namely $\frac{N*-N**}{N}$ of each user for each video level popularity metric (#likes, #shares, #plays, #comments) respectively. Also, it applies the Mann Whitney U Test between the PMF values of the original data and the average of the 100 times shuffled data. On each PMF plot's legend, it shows the statistical test result-p_value. Lastly, it shows the ratio of the histogram values of the original and the average of 100 shuffled data for $\frac{N*-N**}{N}$ to clearly observe the histogram distribution deviation above 1 which is the threshold for randomized careers.

**Output plot(s)** : Can be accessed on the "Plots for original and shuffled100 data" tab under the "Hot streak detection" tab by running the "bokeh serve –show home_page.py" command on a Python IDE like Visual Studio Code.

### 5.3.2 Constructing plots for original data and each of 25 shuffled data

**Input data:** Located in the folders [1], and [7] respectively.

- dataFrame_from_original data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.1
- shuffled data subsets:
  - df_ShuffledForPlots_1.xlsx: Data subset whose metrics are defined in Appendix 9.3
  - ...
  - df_ShuffledForPlots_25.xlsx: Data subset whose metrics are defined in Appendix 9.3

**Script(task3_orig_sh25.py):** Located in folder [9], this Jupyter notebook builds the histogram, CCDF plots of "the relative index distance between the 1st and 2nd biggest hit videos" - namely $\frac{N*-N**}{N}$ of each user for each video level popularity metric (#likes, #shares, #plays, #comments) respectively. Also, it shows the plot of the maximum values of the ratio of the histogram values of the original and the 25 separate shuffled data for $\frac{N*-N**}{N}$ to clearly observe the histogram distribution deviation above 1 which is the threshold for randomized careers. Moreover, it applies the Mann Whitney U Test between the histogram values of the original data and the 25 separate shuffled data. Lastly, it constructs the histogram and CCDF distribution of the 25 statistical test results-p_values.

[9] https://drive.google.com/drive/u/0/folders/1Kmws6V_kx8cOITST1B0gr2urtvmuGXpa

**Output plot(s):** Can be accessed on the "Plots for original and shuffled25 data" tab under the "Hot streak detection" tab by running the "bokeh serve –show home_page.py" command on a Python IDE like Visual Studio Code.

## 5.4 Understanding the onset of a hot streak

The implementation part of this phase first consisted of dataset preparation from the original data subset and from the data subset of the average of shuffling 100 times. Second, plots were built that compare the hashtag entropy distributions before and during hot streak periods for the original data and the average of 100 times shuffled data.

### 5.4.1 Preparing data subsets for plots

#### A. Preparing data subset for original data

#### Part 1:

**Input data:** Located in the folder [1]

- dataFrame_from_original data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.1
- tiktok_cleaned_data.xlsx: Cleaned TikTok dataset

**Script(1_prepare_for_onset.ipynb):** Located in [10], this Jupyter notebook creates new columns to apply the method proposed in [16]. New columns are unique_hashtag_count, total_hashtag_count, freq_hashtag, log_freq_hashtag, log_video_total, hashtag_entropy, Nonset_likes, Nend_likes, Nonset_shares, Nend_shares, Nonset_plays, Nend_plays, Nonset_comments, Nend_comments. They are defined in the Appendix section9.3.

**Output data(dataFrame_for_Onset and hot streak end order.xlsx):** Located in [11] is data subset constructed from "dataFrame_from_original data_for_plots.xlsx" and "tiktok_cleaned_data.xlsx" and has new column values defined in the Appendix section9.4.

#### Part 2:

**Input data (dataFrame_for_Onset and hot streak end order.xlsx):** Located in the folder [11] is data subset constructed from "dataFrame_from_original data_for_plots.xlsx" and "tiktok_cleaned_data.xlsx" which have new column values defined in the Appendix section9.4.

**Script(2_prepare_for_Hb_Hd.ipynb):** Located in [10], this Jupyter notebook creates new columns to apply the method proposed in [16]. New columns are h_before_likes, h_during_likes, h_diff_likes, h_before_shares, h_during_shares, h_diff_shares, h_before_plays, h_during_plays, h_diff_plays, h_before_comments, h_during_comments, h_diff_comments. These columns are defined in the Appendix section9.4.

**Output data(dataFrame_for_Hb Hd plots.xlsx):** Located in [11] is a data subset constructed from "dataFrame_for_Onset and hot streak end order.xlsx" and has new column values defined in the Appendix section9.4.

[10] https://drive.google.com/drive/u/0/folders/1WiNyExcVKBrFPkmaEFzhHzUmyq3_RAij
[11] https://drive.google.com/drive/u/0/folders/1tKCsDHHWc-XBdY73XAbvnKWsNZhXarTQ

### B. Preparing data subset for average of 100 times shuffled data

#### Part 1:

**Input data:** Located in folder [5]

- tiktok_data_after_shuffled_100times.xlsx: Contains the average of 100 times of shuffling the original dataset.
- dataFrame_from_shuffled100 data_for_plots.xlsx: Data subset whose metrics are defined in Appendix 9.3

**Script(1_prepare_for_onset.ipynb):** Located in [12], this Jupyter notebook creates new columns to apply the method proposed in [16]. New columns are unique_hashtag_count, total_hashtag_count, freq_hashtag, log_freq_hashtag, log_video_total, hashtag_entropy, Nonset_likes, Nend_likes, Nonset_shares, Nend_shares, Nonset_plays, Nend_plays, Nonset_comments, Nend_comments. They are defined in the Appendix section9.4.

**Output data(dataFrame_shuffled100_for_Onset and hot streak end order.xlsx):** Located in [13] is data subset constructed from "tiktok_data_after_shuffled_100times.xlsx" and "dataFrame_from_shuffled100 data_for_plots.xlsx" and has new column values defined in the Appendix section9.4.

#### Part 2:

**Input data (dataFrame_shuffled100_for_Onset and hot streak end order.xlsx):** Located in the folder [13] is data subset constructed from tiktok_data_after_shuffled_100times.xlsx and dataFrame_from_shuffled100 data_for_plots.xlsx and has new column values defined in the Appendix section9.4.

**Script(2_prepare_for_Hb_Hd.ipynb):** Located in [12], this Jupyter notebook creates new columns to apply the method proposed in [16]. New columns are h_before_likes, h_during_likes, h_diff_likes, h_before_shares, h_during_shares, h_diff_shares, h_before_plays, h_during_plays, h_diff_plays, h_before_comments, h_during_comments, h_diff_comments. These columns are defined in the Appendix section9.4.

**Output data(dataFrame_shuffled100_for_Hb Hd plots.xlsx):** Located in [13] is a data subset constructed from "dataFrame_shuffled100_for_Onset and hot streak end order.xlsx" and has new column values defined in the Appendix section9.4.

[12] https://drive.google.com/drive/u/0/folders/150Lfppf-mDZPbFY6XL12k4UQ-I9vUs8r
[13] https://drive.google.com/drive/u/0/folders/1Es5J7IhCWmOL9CGbOBFqNblMr_ZuA7oW

### 5.4.2 Constructing plots for original data and average of 100 times shuffled data

**Input data:** Located in the folders [11], and [13] respectively.

- dataFrame_for_Hb Hd plots.xlsx: Data subset built in section5.4.1.
- dataFrame_shuffled100_for_Hb Hd plots.xlsx: Data subset built in section5.4.1.

**Script(task4_orig_sh100.py):** Located in [14], this Python file builds the histogram plots of hashtag entropy before the hot streak of each user for each video level popularity metric (#likes, #shares, #plays, #comments) respectively. The histogram plots for hashtag entropy before the hot streak show the data distribution of randomized careers and the average of real careers. Similarly, this script builds the histogram plots of hashtag entropy before the hot streak of each user for each video level popularity metric (#likes, #shares, #plays, #comments) respectively. Lastly, it produces the hashtag entropy difference before and during the hot streak period of the original data to observe the tendency towards exploration or exploitation around the hot streak.

**Output plot(s)** : Can be accessed on the "Understanding the onset of a hot streak" tab by running the "bokeh serve –show home_page.py" command on a Python IDE like Visual Studio Code.

## 5.5 Visualizing hot streaks of users

This phase of the implementation uses the previous dataset file created for the original data analysis in section 5.4.1 with two new columns added as described in the following subsection.

### 5.5.1 Preparing data subset for plots according to number of likes metric

**Input data(dataFrame_for_Hb Hd plots.xlsx):** Data subset located in folder[11] and built in section5.4.1.

**Script(task5_prepare_df.py):** Located in [15] consists of a function called prepare_dataframe() that creates three new columns streak_duration_likes, relative_streak_duration_likes and streak_time stated in section4.5.2 in Technical Solution chapter4 and adds it to the data subset to be used for the plots for hot streak visualizations.

**Output:** is a data frame to be used for the function that constructs the plots.

### 5.5.2 Preparing data subset for plots according to number of plays metric

**Input data(dataFrame_for_Hb Hd plots.xlsx):** Data subset located in folder[11] and built in section5.4.1.

**Script(task5_prepare_df_plays.py):** Located in [15] consists of a function called prepare_dataframe() that creates three new columns streak_duration_plays, relative_streak_duration_plays and streak_time stated in section4.5.2 in Technical Solution chapter4 and adds it to the data subset to be used for the plots for hot streak visualizations.

**Output:** is a data frame to be used for the function that constructs the plots.

[14] https://drive.google.com/drive/u/0/folders/10OOU8594iJedPZEWW4XCZREgdElBb_Wu
[15] https://drive.google.com/drive/u/0/folders/10h3TuHW4lLuaCws_63lPKd7-YhZXV3ic

### 5.5.3 Constructing the plots according to number of likes metric

**Input data:** Data frame built with function prepare_dataframe() in script task5_prepare_df.py located in folder[15].

**Script(task5.py):** Located in [15] consists of a function called task5() that creates the bar plot for the whole lifecycle of each author, bar plot for the hot streak duration of each author, circle plots to show the 1st, 2nd and 3rd biggest hit timing of each author, a dropdown to sort the authors by productivity, length of hot streak duration, the timing of the 1st biggest hit, the relative timing of the first biggest hit, a second dropdown to show authors' either 1st, 2nd, 3rd, or all hits, a third dropdown to list authors who have the biggest hit in their either early, middle or late careers. There is also a HoverTool added to the figure that shows additional information about the authors, and their hot streak periods.

**Output plot(s)** : Can be accessed on the "Visualize hot streak durations" tab by running the "bokeh serve –show home_page.py" command on a Python IDE like Visual Studio Code.

### 5.5.4 Constructing the plots according to number of plays metric

**Input data:** Data frame built with function prepare_dataframe_plays() in script task5_prepare_df_plays.py located in folder [15].

**Script(task5_plays.py):** Located in [15] consists of a function called task5_plays() that creates the bar plot for the whole lifecycle of each author, bar plot for the hot streak duration of each author, circle plots to show the 1st, 2nd and 3rd biggest hit timing of each author, a dropdown to sort the authors by productivity, length of hot streak duration, the timing of the 1st biggest hit, the relative timing of the first biggest hit, a second dropdown to show authors' either 1st, 2nd, 3rd, or all hits, a third dropdown to list authors who have the biggest hit in their either early, middle or late careers. There is also a HoverTool added to the figure that shows additional information about the authors, and their hot streak periods.

**Output plot(s)** : Can be accessed on the "Visualize hot streak durations" tab by running the "bokeh serve –show home_page.py" command on a Python IDE like Visual Studio Code.

## 5.6 Running the web app on bokeh server

This phase of the implementation describes which files should be used to run the complete web app including all phases of my thesis on localhost using bokeh server. The files below located in folder [16] can be used for this purpose:

**How to execute the files to see the results.docx:** This file located in folder explains the steps to run the complete web app on localhost using bokeh server.

**Requirements.txt:** Lists the Python version and Python libraries that should be installed.

**visualizationManual.pdf:** This manual was prepared as a help menu of the offline web page to guide users about the purpose of each phase, which information they can access in each menu, what kind of in- teractivity can be experienced in figures, and the evaluation of plots.

[16] https://drive.google.com/drive/u/0/folders/13e5eRkoQL56DcOzqaohsEI8dG2OAylCE

# 6 Results

Plots obtained via the Implementation steps in sections 5.1.3, 5.2.2, 5.3.1, 5.3.2, 5.4.2, 5.5.3, and 5.5.4 are examined in the following sections.

## 6.1 User and video level popularity distributions

Newman mentions that if a variable has a range that changes in a great amount like the population of cities or web hits of a large number of users, then while plotting its distribution, if log-scale is used on the x-axis or on both axes, then the resulting plot changes substantially, and in that study for the population metric with log-log scale, the plot was a straight line[17]. This outcome is due to the "power law" characteristic of the distribution [17].

My study's dataset consists of 760 users who have a wide range of values for some user-level popularity metrics and some video-level popularity metrics. When these user-level popularity metrics are plotted in linear scale on both axes, as seen in Figure 6.1, it is observed that there is an exponential decline on the y-axis for the CCDF values. Moreover, the y-axis's exponential decline is lower for the "total number of videos per author" metric compared to the other y-axis's decline for other metrics, because the range of values of "total number of videos per author" [5-1460] are much narrower than the range of values of the other metrics, for instance, than the ones of "total number of followers per author" [6-37500000], as shown in Figure 6.1.

Therefore, a log-linear plot as in Figure 6.2 was built to see if the line becomes straight for these metrics with a wide range of values. Indeed, the plots became straighter after the adjustment, and this change proved that the distribution of the user-level popularity metrics has a "power law" characteristic. In Figure 6.2, since the range of values for the "total number of videos per author" is narrow, it has a higher slope, while the other metrics have a smaller slope as they have a wider range.

All in all, to observe the CCDF distributions of the metrics accurately, having noticed the "power law" characteristic of the data, plots built in log-linear scale provided a better display.

The summary of the results for the user-level popularity distributions are listed as below:

- The range of "total number of videos per author" is $\in$ [5, 1460]. In its range, the "total number of videos per author" has more values close to its minimum than its maximum value.
- The range of the "total number of likes per author" is $\in$ [19, 542 800 000]. In its range, the "total number of likes per author" has more values close to its minimum than its maximum value.
- The range of the "total number of followers per author" is $\in$ [6, 37 500 000]. In its range, the "total number of followers per author" has more values close to its minimum than its maximum value.

- The range of the “total number of followings per author” is ∈ [0, 10000]. Since this is close to a straight line, in this metric’s range, it has more values close to the maximum value.

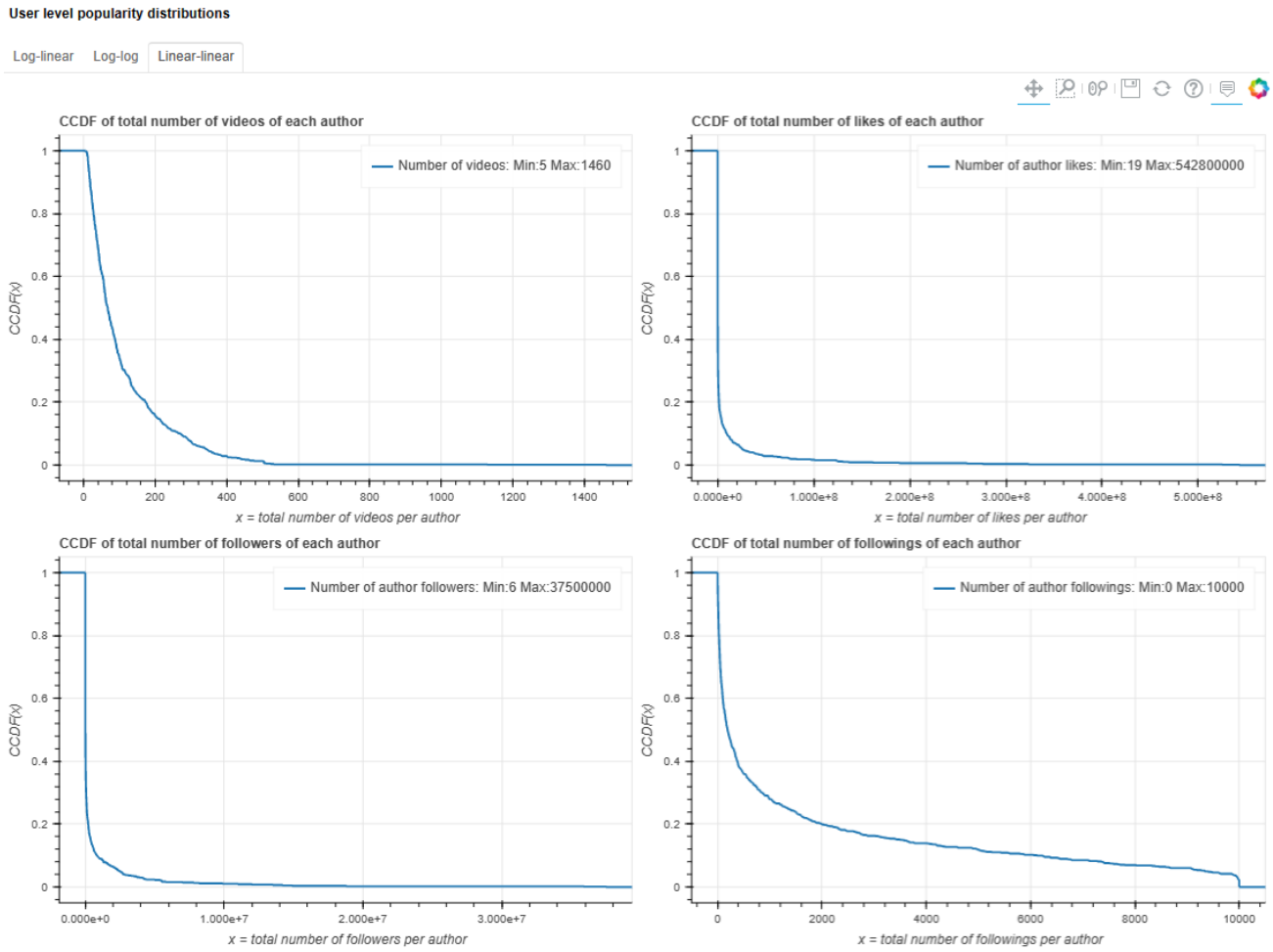



Figure 6.1: CCDF of user-level popularity metrics in linear-linear scale

CCDF plots for user-level popularity metrics in linear scale on both axes. The metrics are the total number of videos per author, the total number of likes per author, the total number of followers per author, and the total number of followings per author, from left to right, respectively.

Similarly, to analyze the video-level popularity metrics’ distributions, their CCDF plots were built and shown in Figure6.3. As there is an exponential decline in y-axis values in each plot and the range of values is substantially high for the x-axis, again log-linear plots were built. The CCDF plots in log-linear scale for each video-level popularity metric are shown in Figure 6.4 and the results of the distributions in the plots are summarized below:

- The range of the “total number of video likes per author” is ∈ [19, 511 378 900]. According to the value of the slope of the log-linear plot, the “total number of video likes per author” has more values in its range close to its minimum than its maximum value.

- The range of the “total number of video shares per author” is ∈ [0, 6 241 222]. According to the value of the slope of the log-linear plot, which is rather straight, the “total number of video shares per author” has more values in its range close to its maximum value.

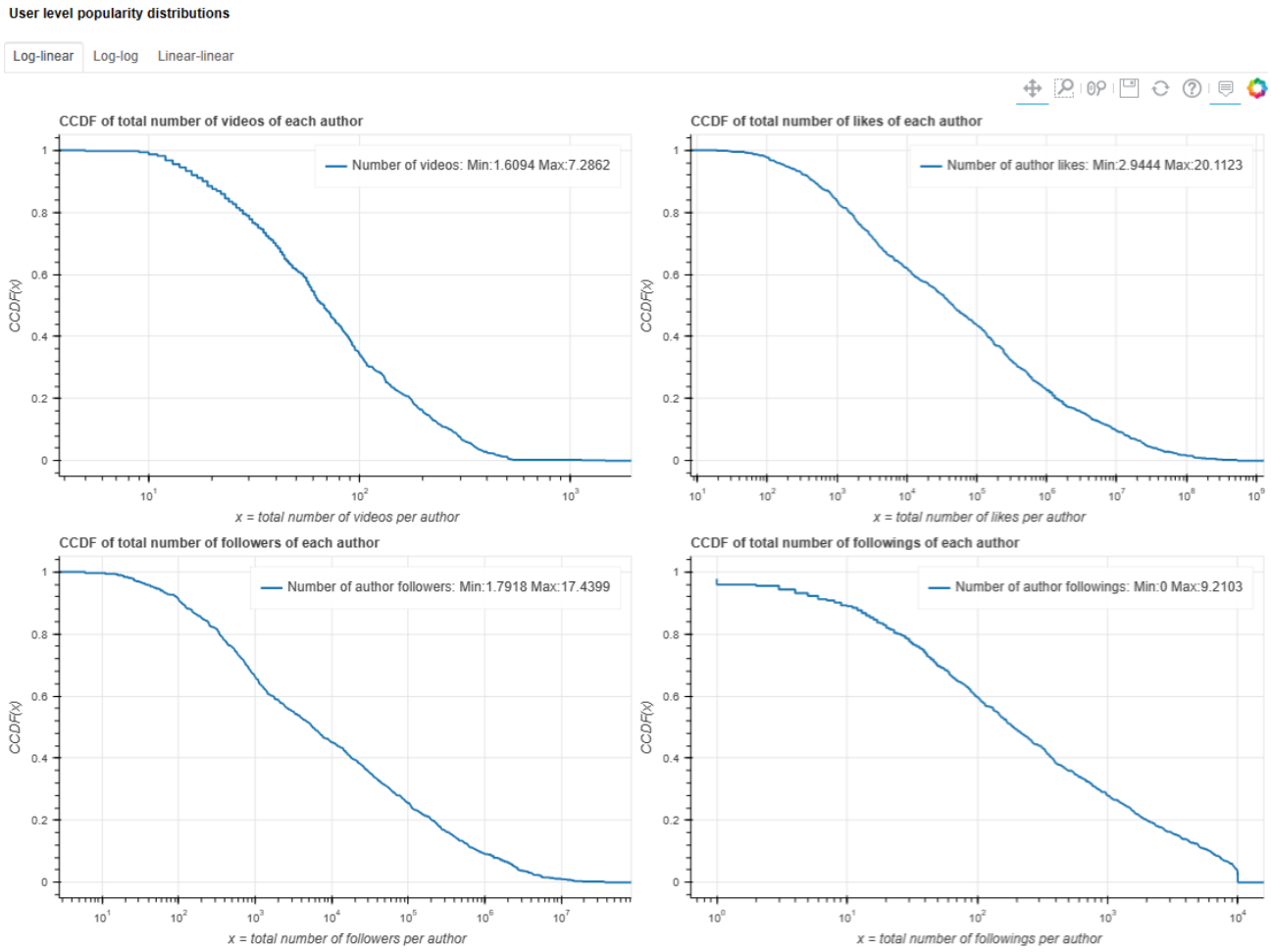



Figure 6.2: CCDF of user-level popularity metrics in log-linear scale

CCDF plots for user-level popularity metrics in log-linear scale. The metrics are the total number of videos per author, the total number of likes per author, the total number of followers per author, and the total number of followings per author, from left to right, respectively.

- The range of the "total number of video plays per author" is $\in$ [77, 4 809 700 000]. According to the value of the slope of the log-linear plot, the "total number of video plays per author" has more values in its range close to its minimum than its maximum value.
- The range of the "total number of video comments per author" is $\in$ [0, 8 420 173]. According to the value of the slope of the log-linear plot, which seems as the highest among the other metrics' plots' slopes, the "total number of video comments per author" has more values in its range close to its minimum than its maximum value.

The plots in log-log axis scale are shown in Appendix 9.2.

As a result, one can say that, since values of the "total number of video shares per author" are closer to the maximum than the minimum value in its range in terms of interactivity, it can be asserted that if a video gets shared, it is highly likely that it can be shared again.

But, compared to the other metrics, the "total number of video shares per author" has the minimum range of values, so the ranges that belong to the total engagement with the videos show that the "total number of video plays per author" has the widest range with [77, 4 809 700 000]. Thereby, in this study, the "number of plays" a video received can be considered a better metric to analyze the timing of the biggest hits, detecting hot streaks, understanding the onset of a hot streak, and examining hot streak durations.

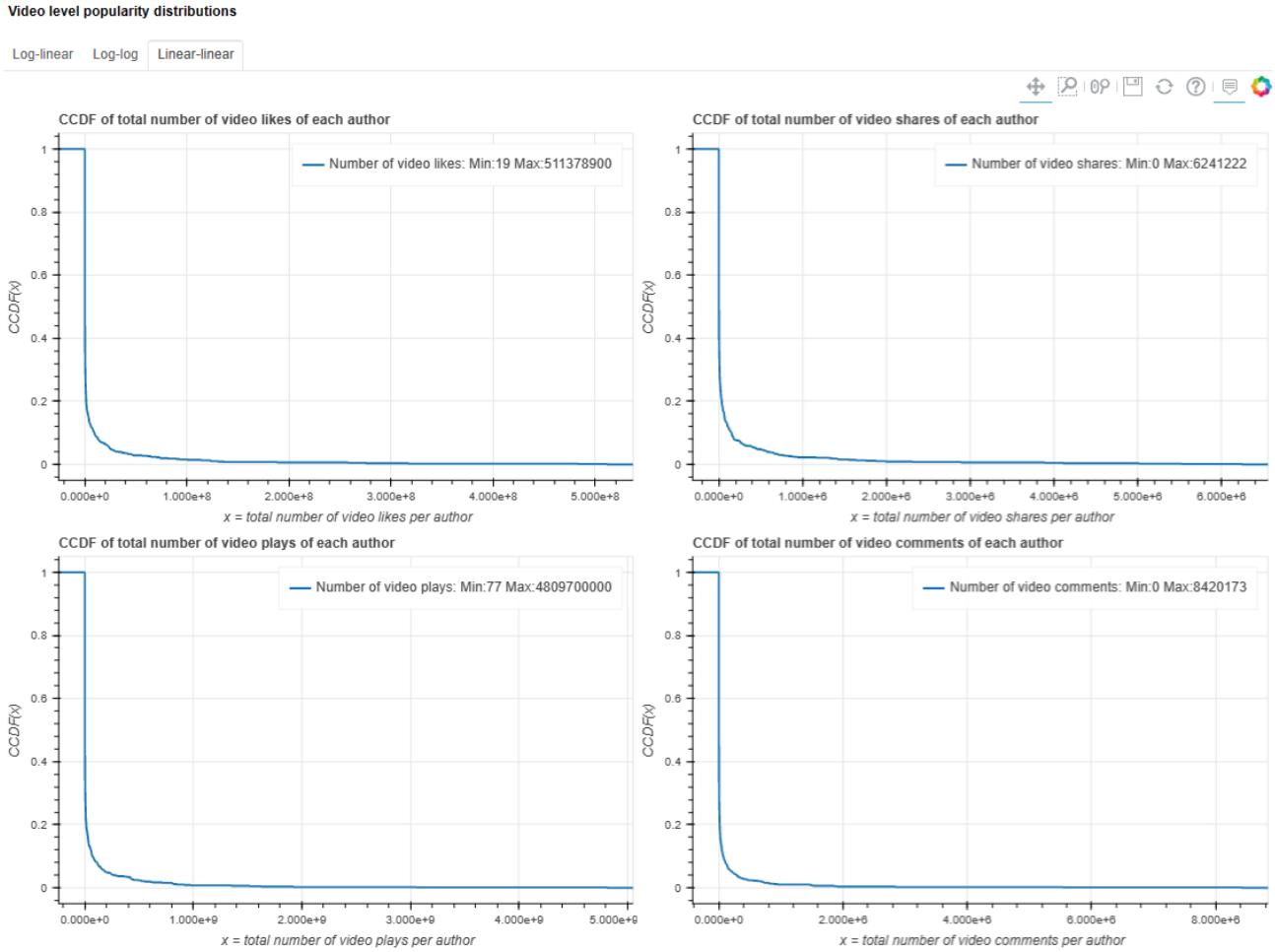



Figure 6.3: CCDF of video-level popularity metrics in linear-linear scale

CCDF plots for video level popularity metrics in linear scale on both axes. The metrics are the total number of video likes per author, the total number of video shares per author, the total number of video plays per author, the total number of video comments per author, from left to right, respectively.

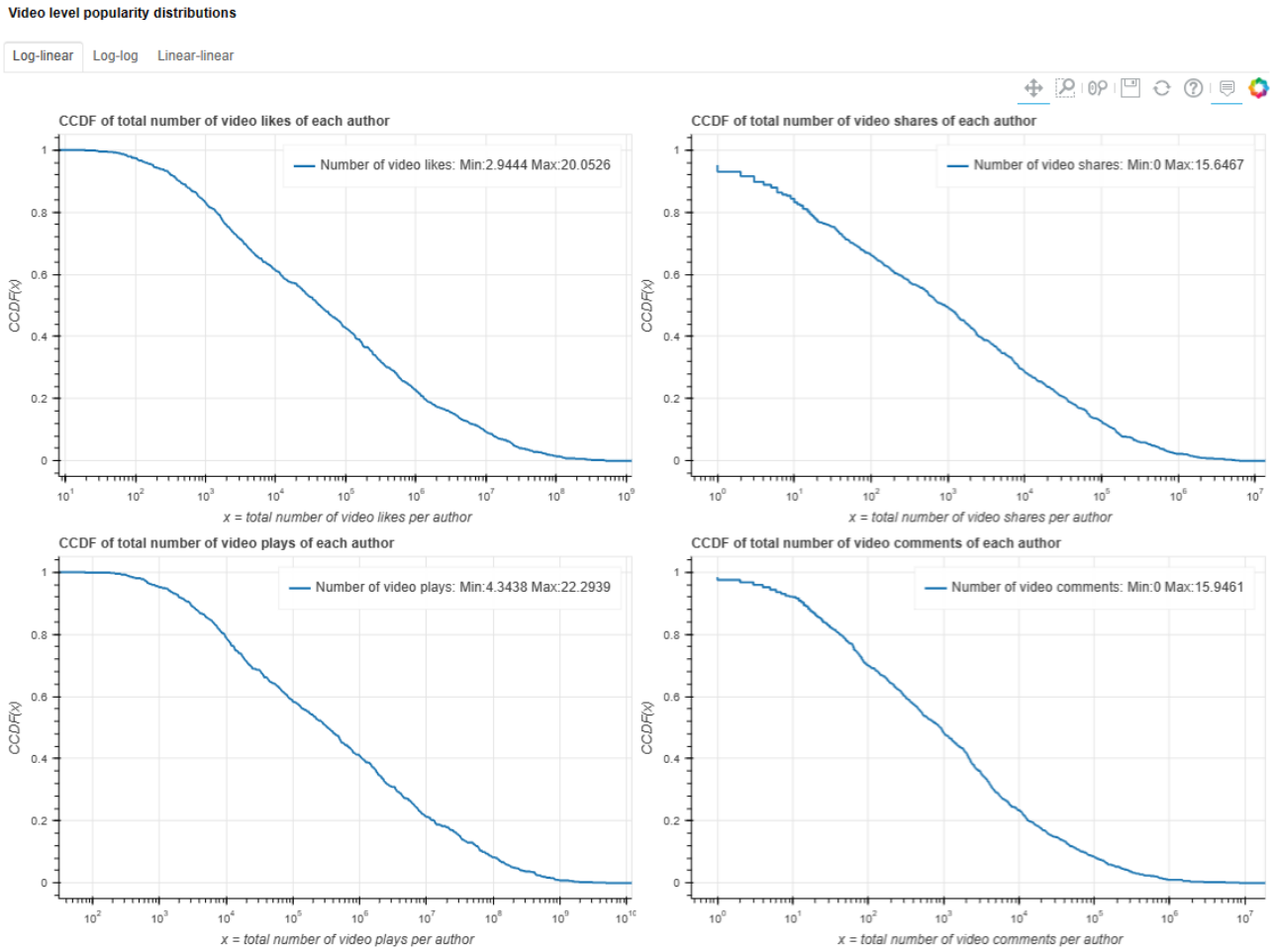



Figure 6.4: CCDF of video-level popularity metrics in log-linear scale

CCDF plots for video level popularity metrics in linear scale on both axes. The metrics are the total number of video likes per author, the total number of video shares per author, the total number of video plays per author, the total number of video comments per author, from left to right, respectively.

## 6.2 Results of the "Timing of an influencer's hit" phase

In this part, I aimed to identify whether TikTok influencers' biggest hits occur in the early, late, or middle phases of their careers.

While building the plots, the parameters below were used to show the distributions of the hits' timings according to the video-level popularity metrics: "number of likes", "number of shares", "number of plays", and "number of comments". The definitions of the parameters below are also stated in Appendix 9.1.1 for each video-level popularity metric.

- $N$ as the "total number of videos"
- $N^*$ as the "index of the highest-impact video"
- $\frac{N^*}{N}$ as the "relative index of the highest-impact video"

Further, analysis of $\frac{N^*}{N}$ would be:

- If $\frac{N^*}{N} = \frac{1}{N}$, then the biggest hit is the first video of an author
- If $\frac{N^*}{N} = \frac{1}{2}$, then the biggest hit video of the author is the one in the middle
- If $\frac{N^*}{N} = \frac{N}{N} = 1$, then the biggest hit is the last video of the author

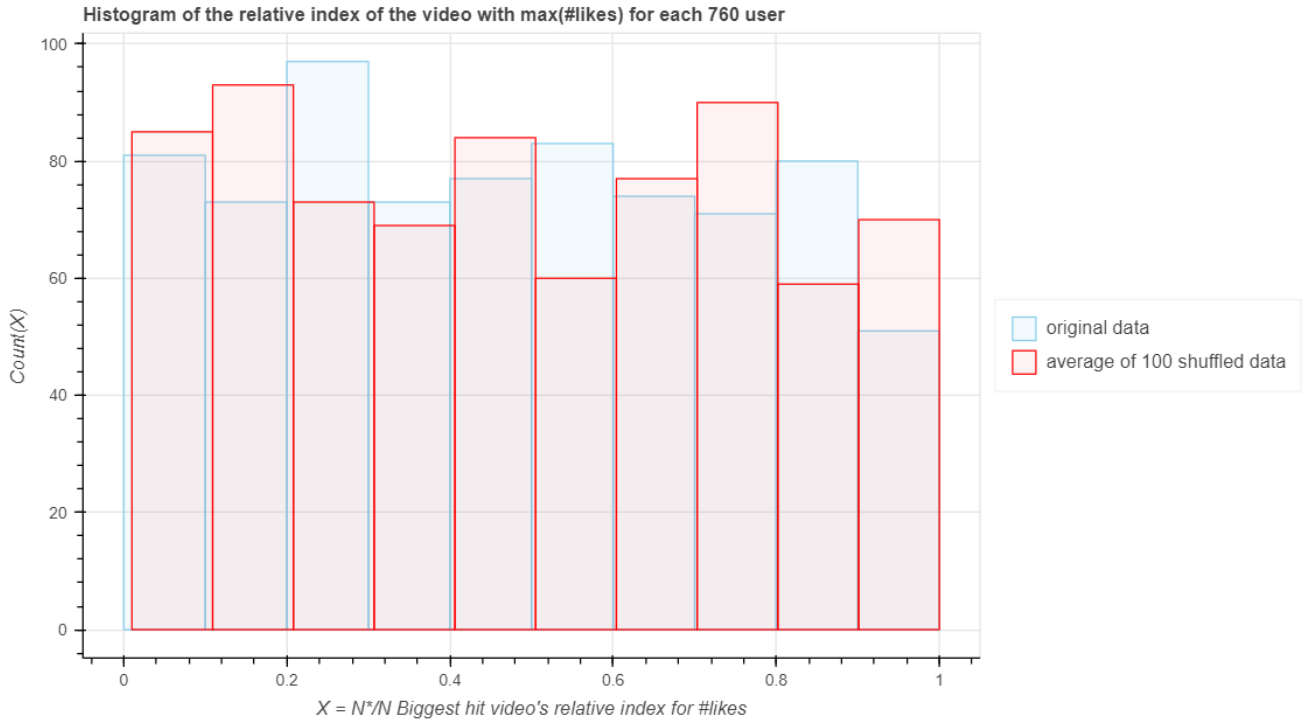
Also, the histogram, and CCDF plots of real, and randomized careers were placed on top of each other to compare if there is a difference between these. In this way, a better estimated inference could be made on the distributions of the dataset, as if it represents a larger number of users.

First, the histogram distributions of the biggest hit's relative video index (N*/N) for the original data according to each video-level popularity metric clearly showed higher values in the early career, as seen in Figures 6.5(a), 6.5(b), 6.6(a), and 6.6(b).

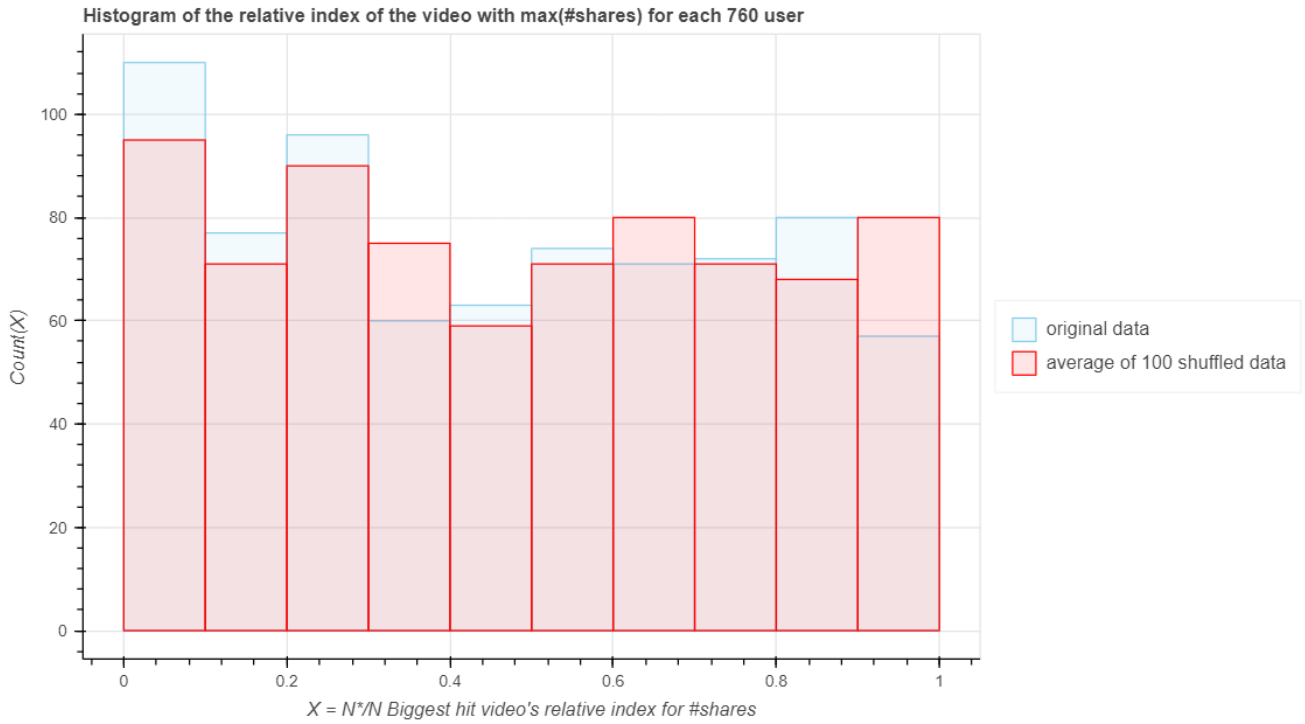
On the contrary, for the average of 100 times shuffled data representing randomized careers, the histogram distributions for the same video-level popularity metrics and parameter (N*/N) follow a random pattern. The close values in the histogram distribution of each Figures 6.5(a), 6.5(b), 6.6(a), and 6.6(b) indicate that the occurrence of hit is random within an influencer's lifecycle.

Moreover, when CCDF plots of the original data and the shuffled data are compared for all video-level popularity metrics as in Figures 6.7(a), 6.7(b), 6.8(a), and 6.8(b), it is seen that the plot of the shuffled data is straighter than the plot of original data for all metrics. This indicates that the distribution of the relative video index of the biggest hit has close values to each other- meaning that the biggest hit can occur at any time in randomized careers.

Lastly, Mann-Whitney U Test results shown in table 6.1 indicate that the original dataset and the average of 100 times shuffled dataset come from the same distribution and there is not a big difference between them.

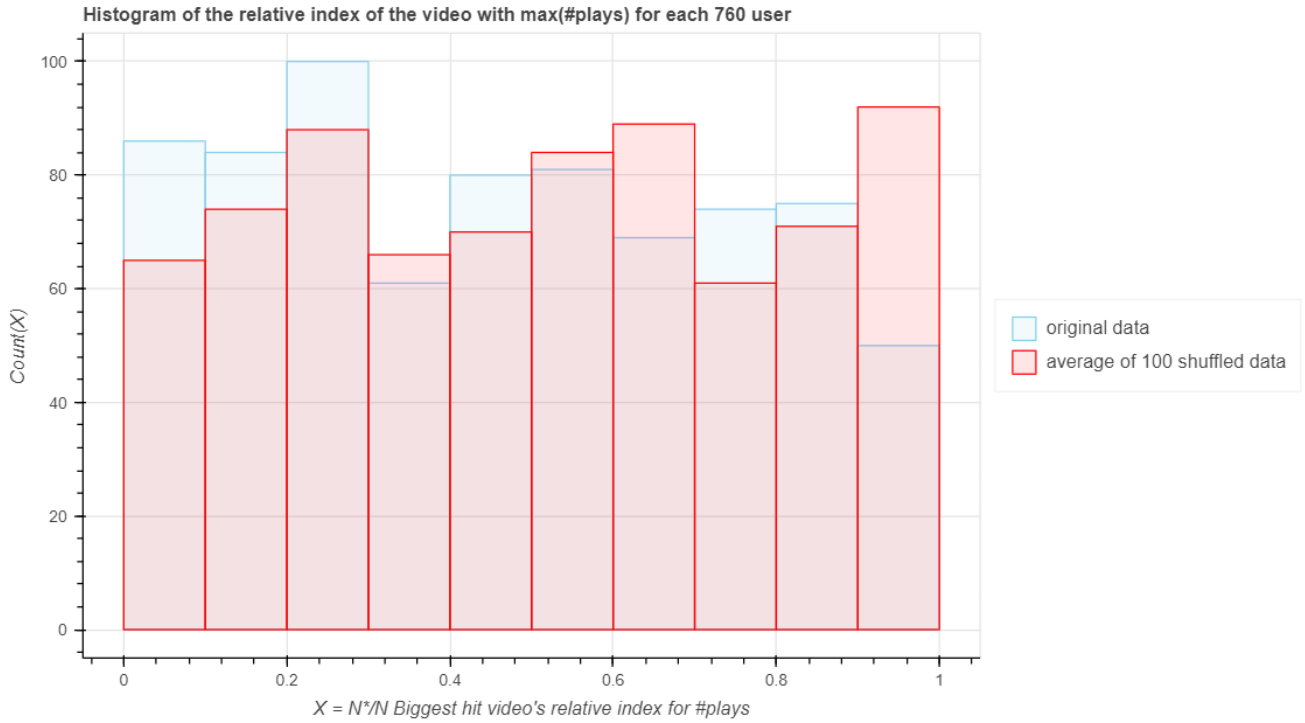



(a) Histogram distribution of the biggest hit's relative video index for for real, and randomized careers according to #likes

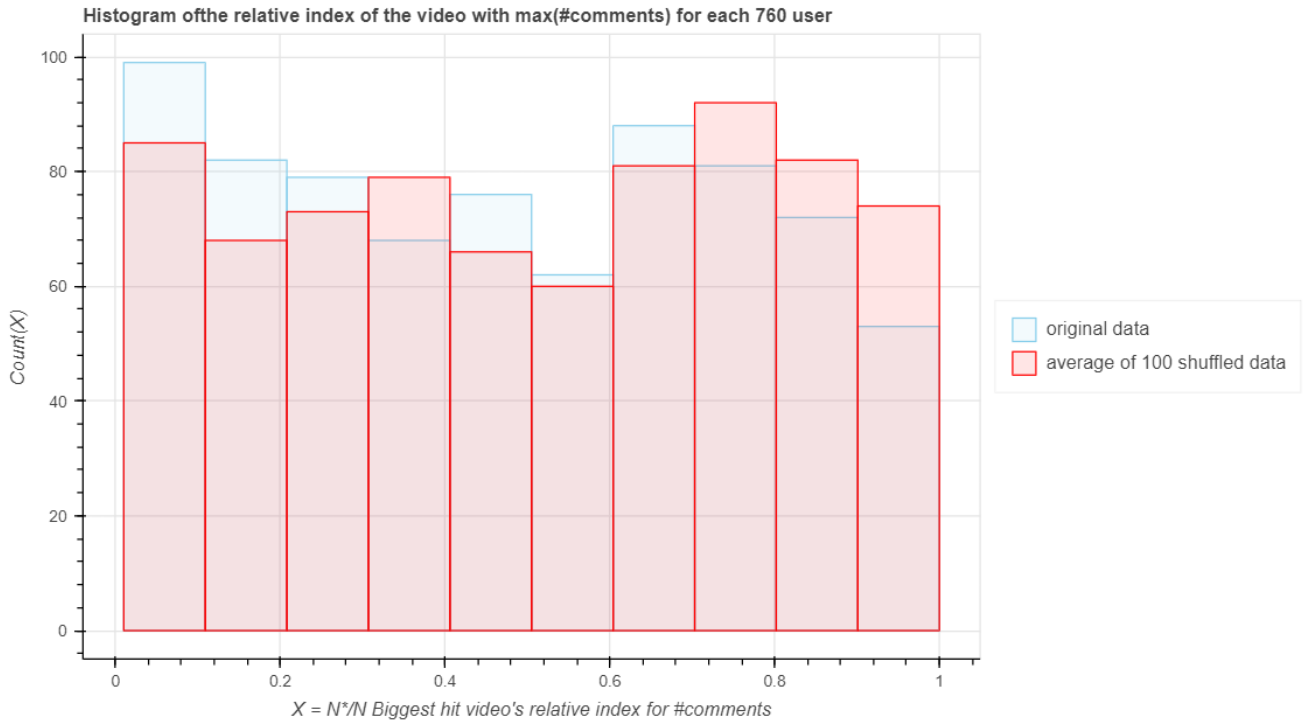



(b) Histogram distribution of the biggest hit's relative video index for for real, and randomized careers according to #shares

Figure 6.5: Histogram distribution of the biggest hit's relative video index for for real, and randomized careers according to (a) max(#likes), and (b) max(#shares)

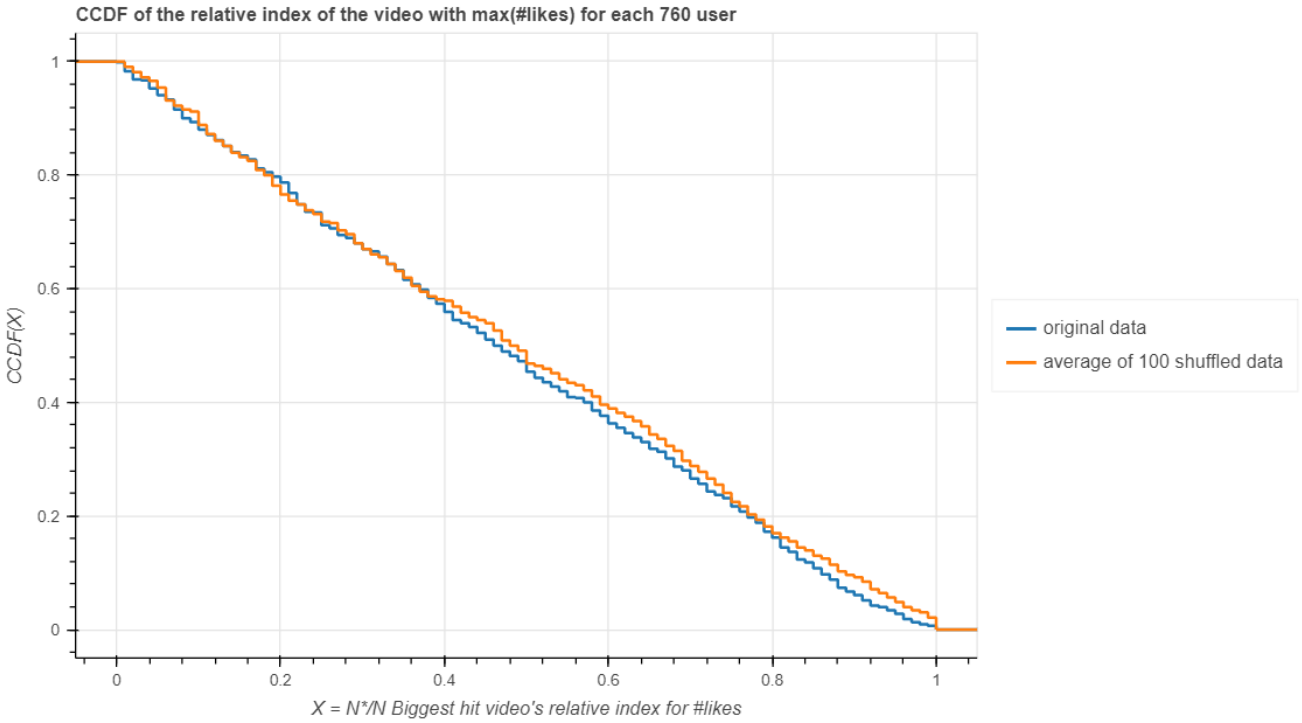



(a) Histogram distribution of the biggest hit's relative video index for real, and randomized careers according to #plays

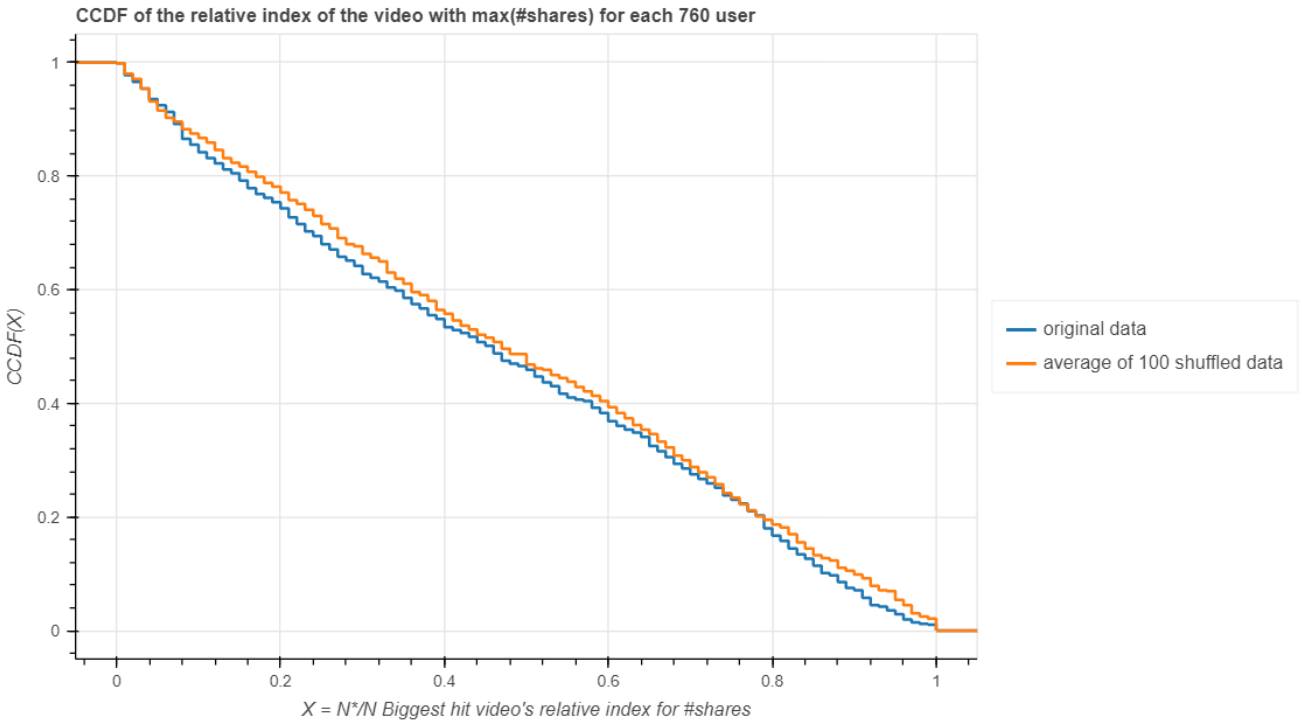



(b) Histogram distribution of the biggest hit's relative video index for real, and randomized careers according to #comments

Figure 6.6: Histogram distribution of the biggest hit's relative video index for real, and randomized careers according to (a) max(#plays), and (b) max(#comments)

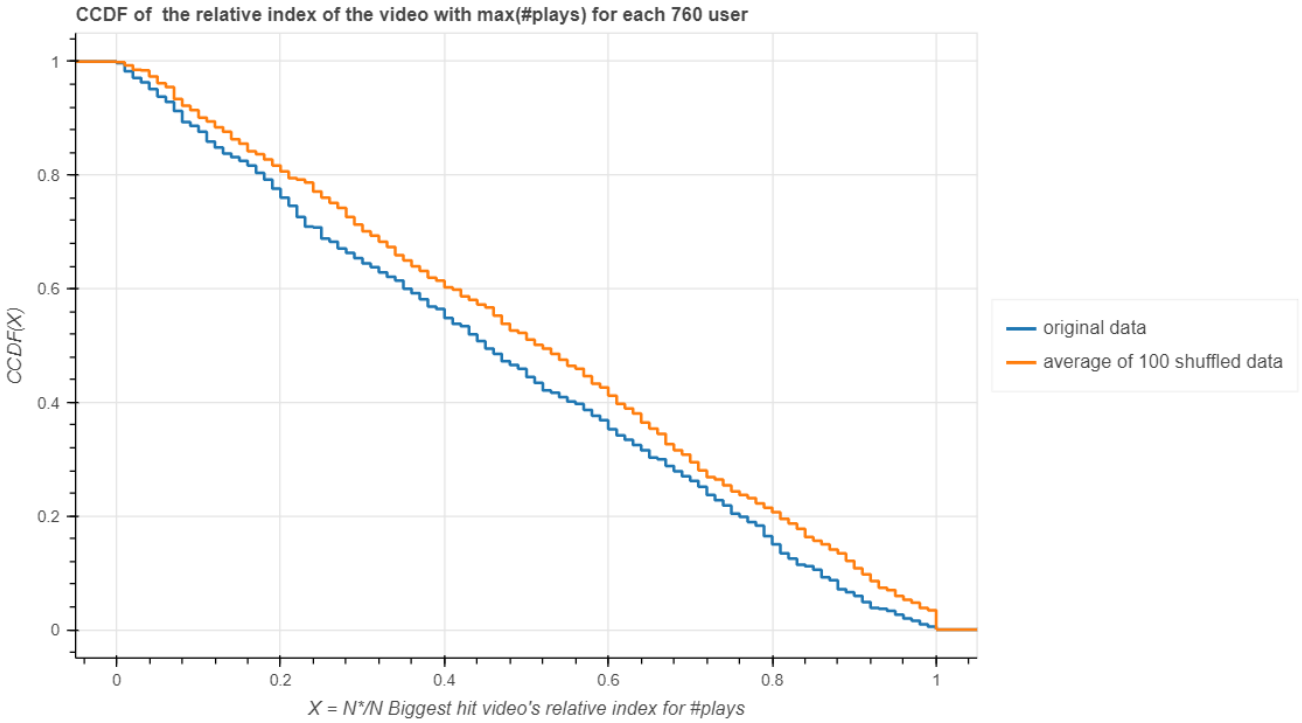



(a) CCDF distribution of the biggest hit’s relative video index for real, and randomized careers according to #likes

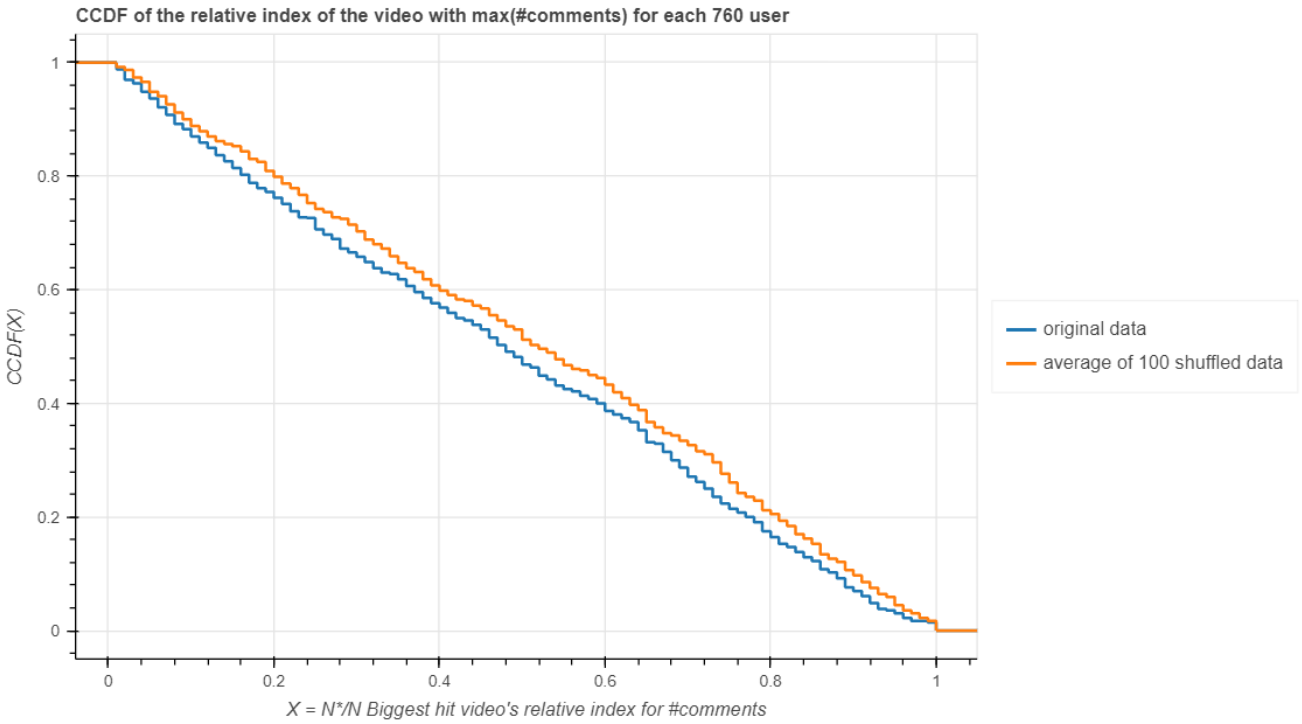



(b) CCDF distribution of the biggest hit’s relative video index for real, and randomized careers according to #shares

Figure 6.7: CCDF distribution of the biggest hit’s relative video index for real, and randomized careers according to (a) max(#likes), and (b) max(#shares)




(a) CCDF distribution of the biggest hit's relative video index for real, and randomized careers according to #plays




(b) CCDF distribution of the biggest hit's relative video index for real, and randomized careers according to #comments

Figure 6.8: CCDF distribution of the biggest hit's relative video index for real, and randomized careers according to (a) max(#plays), and (b) max(#comments)

| Metric | p_value |
|---|---|
| $\#likes$ | 0.8923 |
| $\#shares$ | 0.8489 |
| $\#plays$ | 0.8817 |
| $\#comments$ | 0.8192 |

Table 6.1: p_values of the Mann-Whitney U Test results on original dataset and average of 100 times shuffled dataset according to the metrics $\#likes$, $\#shares$, $\#plays$, and $\#comments$.

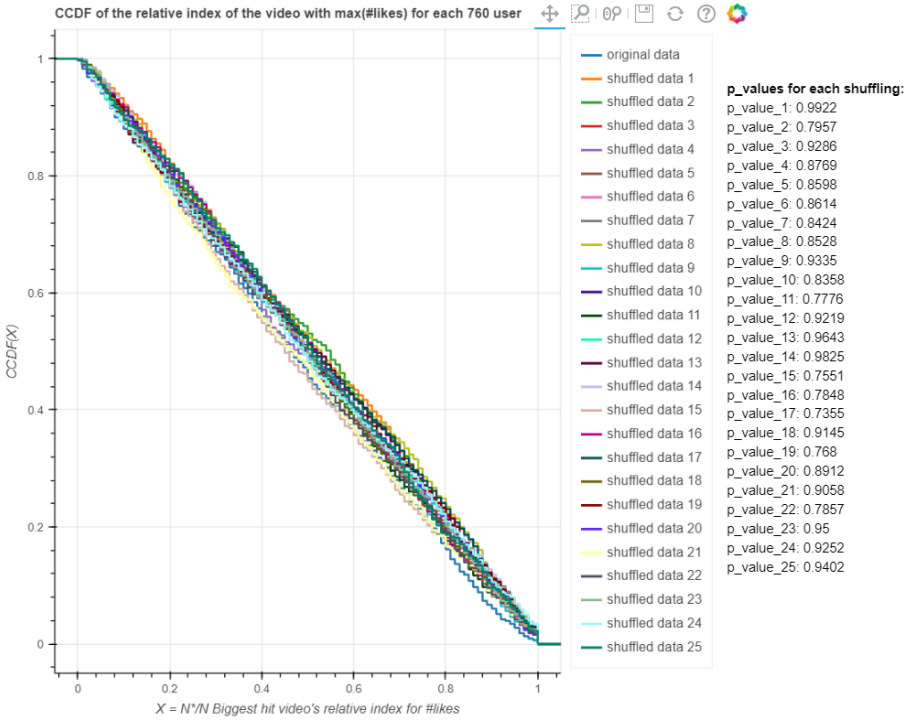
### 6.2.1 Analysis for the 25 shuffled datasets and the original dataset

Regarding the comparative analysis for the other shuffled datasets-25 separate shuffled datasets, to see all the 25 different distributions and compare them with the original dataset's distribution for the relative video index of the biggest hit according to all the video-level popularity metrics, CCDF plots were built.

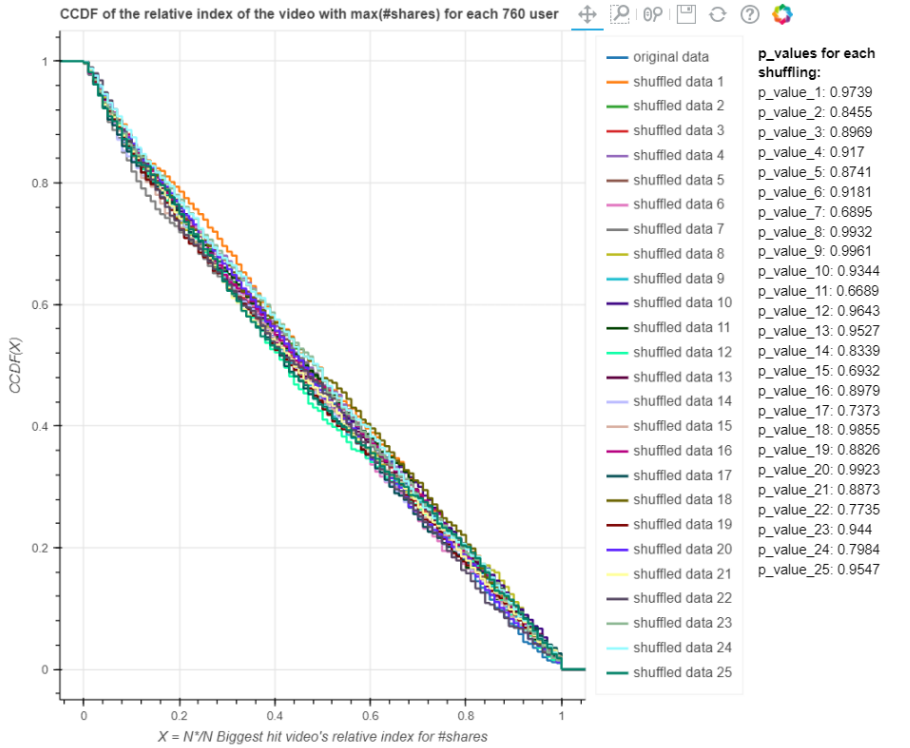
Figures 6.9(a), 6.9(b), 6.10(a), and 6.10(b) show the CCDF plots and the statistical test's results-p_values. In the "Data Characterization phase" 6.1, it was identified that the "total number of plays per author" has the widest range among other video-level popularity metrics, followed by the "total number of likes per author", the "total number of comments per author" and the "total number of shares per author" in descending order in terms of range width. As a parallel result, the CCDF plots for N*/N are straighter for the "number of comments", and the "number of shares" while the CCDF plots of shuffled datasets for the "number of plays" can deviate more from the CCDF plot of the original dataset compared to the other metrics.

Moreover, in general, the plots of the shuffled data are straighter than the plot of the original data for all metrics. This indicates that the distribution of the relative video index of the biggest hit has close values to each other - meaning that the biggest hit can occur at any time in randomized careers.

The statistical test results displayed in Figures 6.11(a), 6.11(b), 6.12(a), and 6.12(b) reveal that p_value distributions range from about 0.44 to approx. 0.99. Since the p_values are substantially larger than the 0.05 threshold value, one can assert the datasets for N*/N of both original and 25 shuffled data come from the same distribution. Namely, there is not a significant difference between the real and randomized careers in terms of the timing of the influencers' biggest hits.

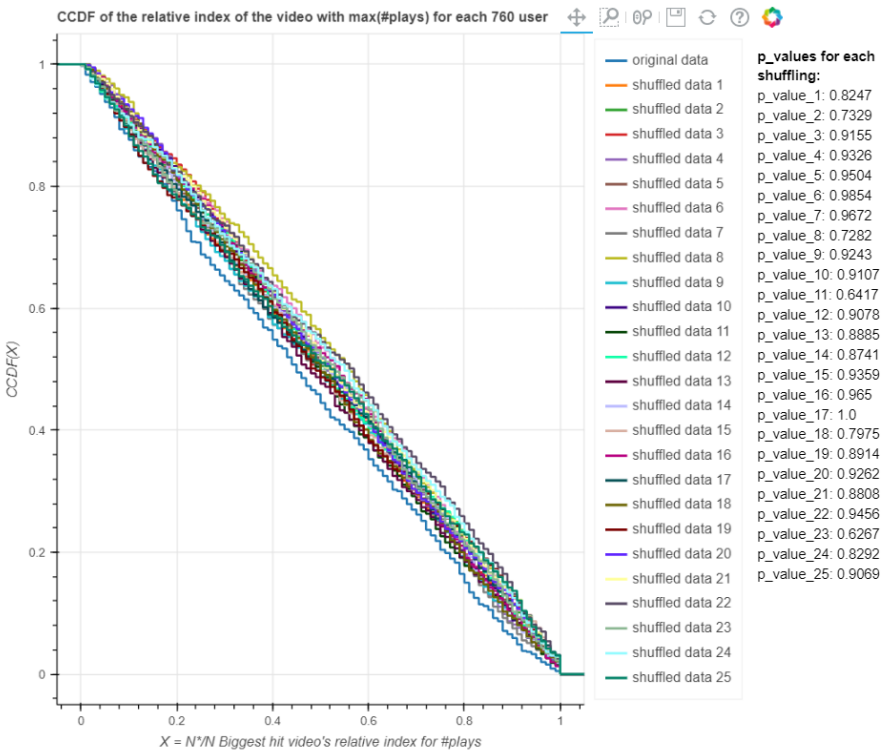



(a) CCDF distribution of the biggest hit's relative video index for real, and randomized careers according to #likes

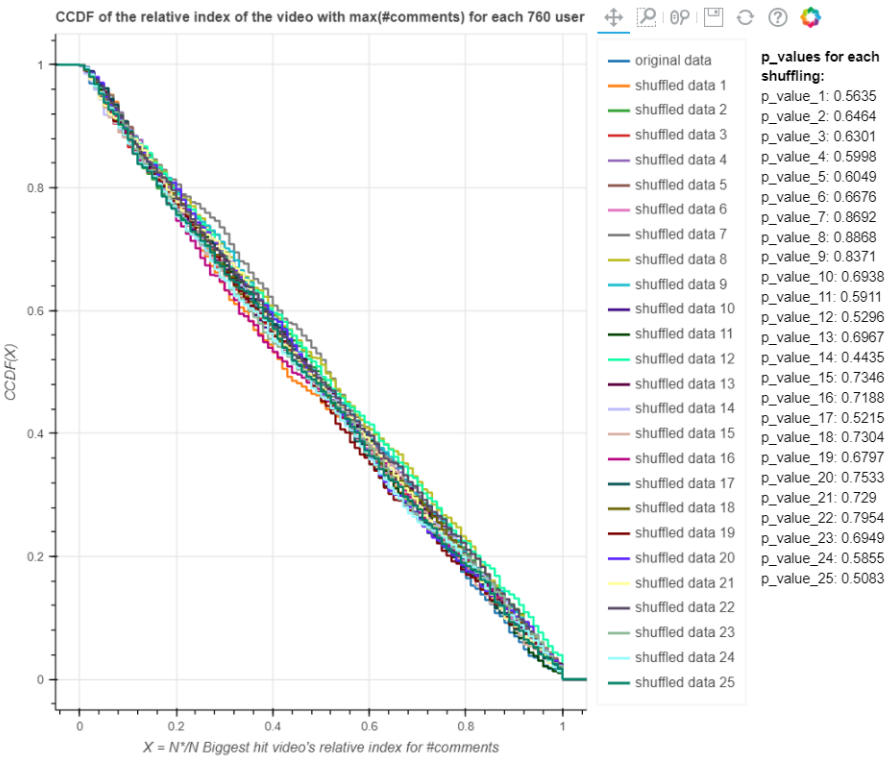



(b) CCDF distribution of the biggest hit's relative video index for real, and randomized careers according to #shares

Figure 6.9: CCDF distribution of the biggest hit's relative video index for real, and randomized careers (25 separate shuffled dataset) according to (a) max(#likes), and (b) max(#shares)

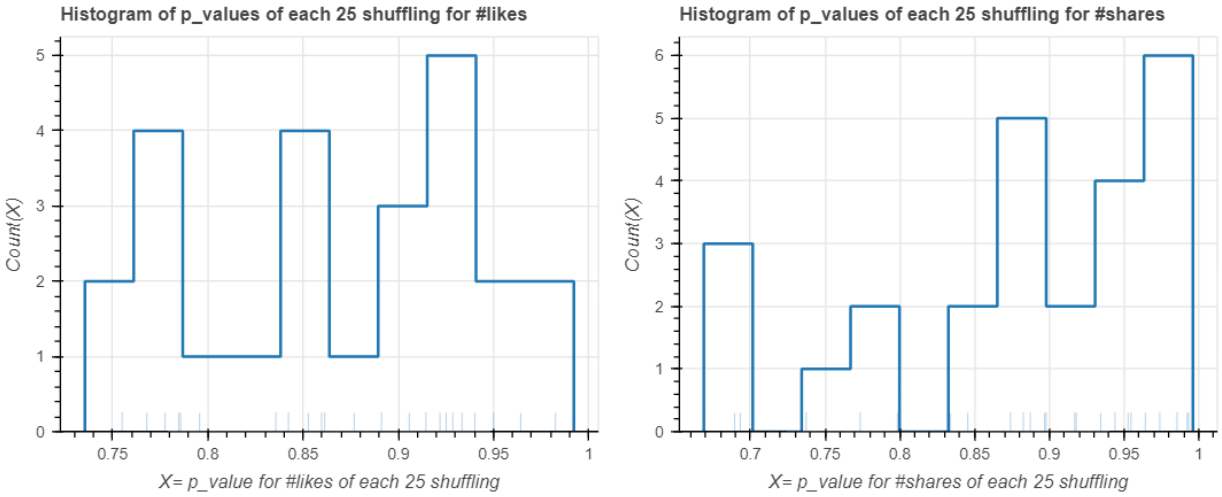



(a) CCDF distribution of the biggest hit’s relative video index for real, and randomized careers according to #plays




(b) CCDF distribution of the biggest hit’s relative video index for real, and randomized careers according to #comments

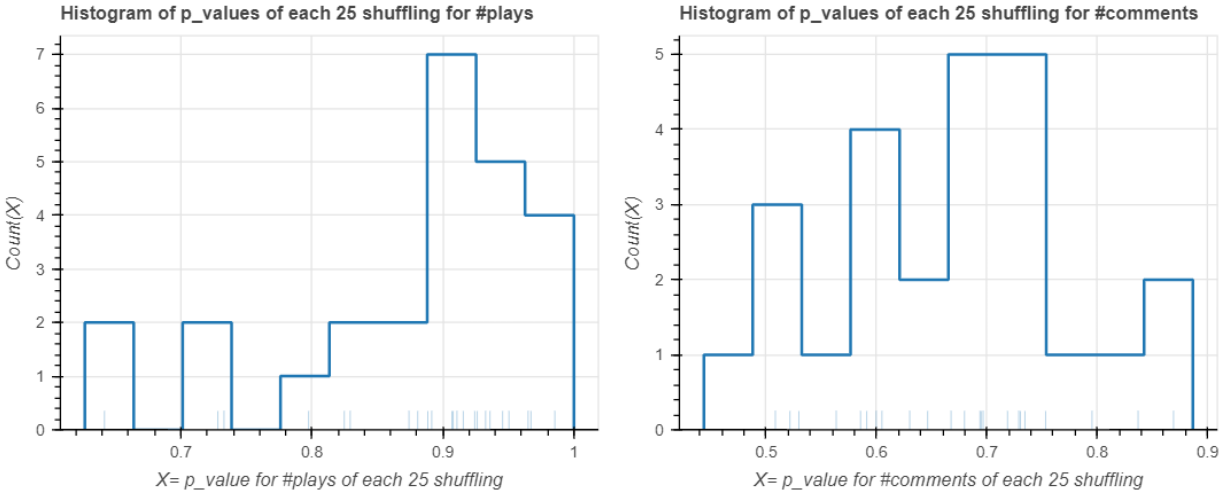
Figure 6.10: CCDF distribution of the biggest hit’s relative video index for real, and randomized careers (25 separate shuffled dataset) according to (a) max(#plays), and (b) max(#comments)




(a) Histogram distribution of the p_values for #likes

(b) Histogram distribution of the p_values for #shares

Figure 6.11: Histogram distribution of the p_values of the Mann Whitney U Test results regarding the biggest hit's relative video index distribution comparison for real, and randomized careers (25 separate shuffled dataset) according to (a) max(#likes), and (b) max(#shares)




(a) Histogram distribution of the p_values for #plays

(b) Histogram distribution of the p_values for #comments

Figure 6.12: Histogram distribution of the p_values of the Mann Whitney U Test results regarding the biggest hit's relative video index distribution comparison for real, and randomized careers (25 separate shuffled dataset) according to (a) max(#plays), and (b) max(#comments)

## 6.3 Results of the "Hot streak detection" phase

After examining the timing of the biggest hit, the next question is whether influencers experience average success before and after their hot streak periods. Namely, the relative distance between the index of the biggest hit and the second biggest hit video was investigated to address this question.

While building the plots, the parameters below were used to show the distributions of the relative distance between the video index of the biggest and the second-biggest hit according to the video-level popularity metrics: "number of likes", "number of shares", "number of plays", and "number of comments". The definitions of the parameters below are also stated in Appendix 9.1.1 for each video-level popularity metric.

- N as the "total number of videos" either in the original or shuffled dataset
- $N^*$ as the "index of the highest-impact video" in the original dataset
- $N^{**}$ as the "index of the 2nd highest-impact video" in the original dataset
- $\frac{N^*-N^{**}}{N}$ as the relative distance between the index of the highest-impact and 2nd highest-impact video in the original dataset
- $Ns^*$ as the "index of the highest-impact video" in the shuffled dataset
- $Ns^{**}$ as the "index of the 2nd highest-impact video" in the shuffled dataset
- $\frac{Ns^*-Ns^{**}}{N}$ as the relative distance between the index of the highest-impact and 2nd highest-impact video in the shuffled dataset

Lastly, as in the method of Liu et al. [16] described in section 4.3.1, the following parameters were also interpreted and used in the plots.

- $\Delta N = N^* - N^{**}$ distance between the index of the highest-impact and 2nd highest-impact video in the original dataset
- $\frac{\Delta N}{N}$ as the relative distance between the index of the highest-impact and 2nd highest-impact video in the original dataset
- $\Delta Ns = Ns^* - Ns^{**}$ as distance between the index of the highest-impact and 2nd highest-impact video in the shuffled dataset
- $\frac{\Delta Ns}{N}$ as the relative distance between the index of the highest-impact and 2nd highest-impact video in the shuffled dataset
- $R(\frac{\Delta N}{N})$ as the fraction of the histogram distribution of $\frac{\Delta N}{N}$ to the histogram distribution of the $\frac{\Delta Ns}{N}$

Histogram, and CCDF plots of $\frac{\Delta N}{N}$ representing real careers, and $\frac{\Delta Ns}{N}$ representing randomized careers are shown in the figures below. These plots of real and randomized careers were placed on top of each other to compare if there is a difference between these. In this way, a better estimated inference could be made on the distributions of the dataset, as if it represents a larger number of users.

First, when the histogram plot in Figure 6.13(a) for the "number of likes" metric is examined, regarding the original dataset, it is observed that the values are substantially close to zero, meaning that the biggest and the 2nd biggest hit timings are close to each other. But, according to the distributions, the timing of the second biggest hit generally is after the timing of the biggest hit. However, the values on the plots for the shuffled data are not mainly close to 0, meaning the second biggest hit timing doesn't herald close biggest hit timing or vice versa. Also, for most influencers, the second hit comes after the first biggest hit, but not before.

Second, when the histogram plot of the original data for the "number of shares" metric is examined in Figure 6.13(b), it shows a similar distribution to the one for the "number of likes". But, the histogram plot of the shuffled data is not like the one for the "number of likes" metric. The distribution of $\frac{\Delta N}{N}$ is similar to the distribution of $\frac{\Delta Ns}{N}$ for the "number of shares" metric.

Third, when the histogram plot of the original data for the "number of plays" metric is examined in Figure 6.14(a), it is observed that the values are substantially close to zero, meaning the biggest and the 2nd biggest hit timings are close to each other. However, according to distributions, in contrast to the pattern on the plot for the "number of likes" metric, the timing of the second hit generally is before the timing of the first hit for the "number of plays" metric. Moreover, the shuffled dataset's distribution also shows values substantially close to zero, but in contrast to the original data, having more negative values indicates that the timing of the biggest hit tends to be before the timing of the 2nd biggest hit.

As seen in Figure 6.14(b), the histogram plot of the original data for the "number of comments" metric shows a pattern similar to the plots of the original data for the "number of likes", but the shuffled data distribution, which shows a tendency towards negative values, is more balanced than the one for the original data.

Furthermore, the results of the Mann Whitney U Test -p-values that have values higher than 0.05 in table 6.2 show that the shuffled dataset and the original dataset come from the same distribution.

| **Metric** | **p_value** |
|---|---|
| $\#likes$ | 0.7259 |
| $\#shares$ | 0.7449 |
| $\#plays$ | 0.8344 |
| $\#comments$ | 0.8082 |

Table 6.2: p_values of the Mann-Whitney U Test results on original dataset and average of 100 times shuffled dataset according to the metrics $\#likes$, $\#shares$, $\#plays$, and $\#comments$.

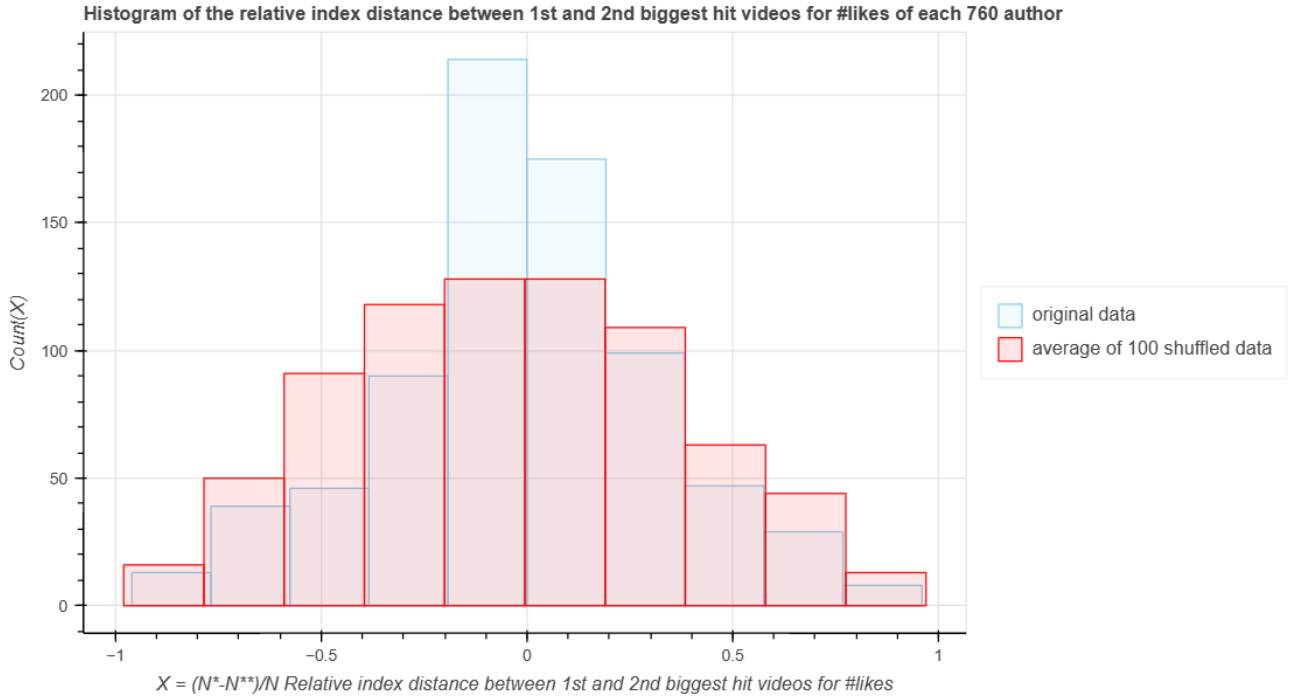
CCDF plots of the real and randomized careers for each metric in Figures 6.15(a), 6.15(b), 6.16(a), and 6.16(b) show that there are fewer variations in the distribution for the original data, but the shuffled data has a more balanced distribution, so the line representing it is straighter on the figures.

Additionally, in Figures 6.17(a), 6.17(b), and 6.18(b), R distribution-the ratio of the histogram plot of the original to the shuffled datasets- for the "number of likes", the "number of shares", and the "number of comments" have a large number of values close to zero. However, there are more negative values than positive ones, meaning

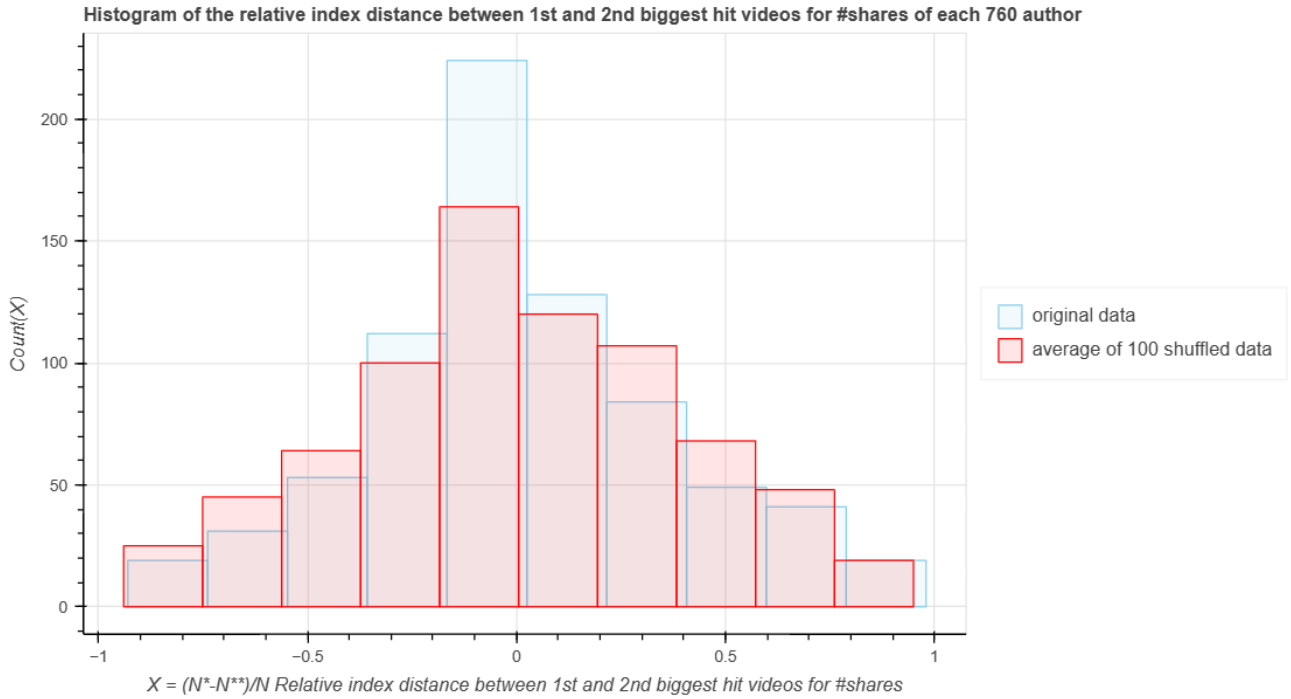
the 2nd biggest hit occurs after the 1st biggest hit in general, but not before. The maximum values for the ratio (Rmax) in each figure are 1.67, 1.37, and 1.56 respectively.

On the contrary, in Figure 6.18(a), R distribution for the “number of plays” has a large number of values close to zero, but it has more positive values than negative ones. Namely, the 2nd biggest hit generally occurs before the 1st biggest hit, but not after. The maximum value for the ratio (Rmax) of the “number of plays” is 1.52.

To sum up, except for the “number of plays” metric, for all the other metrics, the biggest hit’s timing is before the 2nd biggest hit’s timing. But, the magnitude of the “total number of plays per author” for all users in the dataset has the widest range [77- 4 809 700 000]. Also, this range is about 8 times more than the range of the “total number of likes per author”, about 500 times more than the range of the “total number of comments per author”, and about 680 times more than the range of the “total number of shares per author” in the overall distribution. Thereby, one cannot claim when it is more likely for the second hit to occur: before or after the 1st biggest hit. But, from the distribution of data in the figures, it can be asserted that the timing of the biggest and the 2nd biggest hit is close to each other for all video-level popularity metrics both in real and randomized careers.

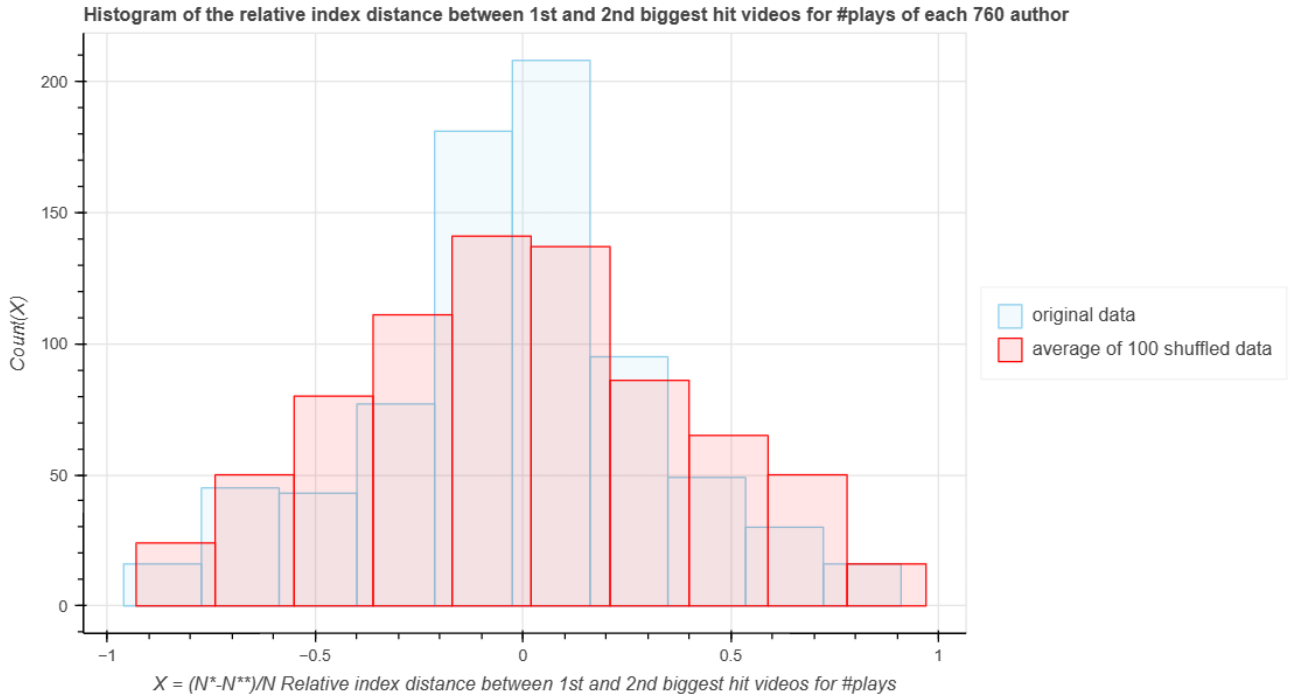



(a) Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #likes

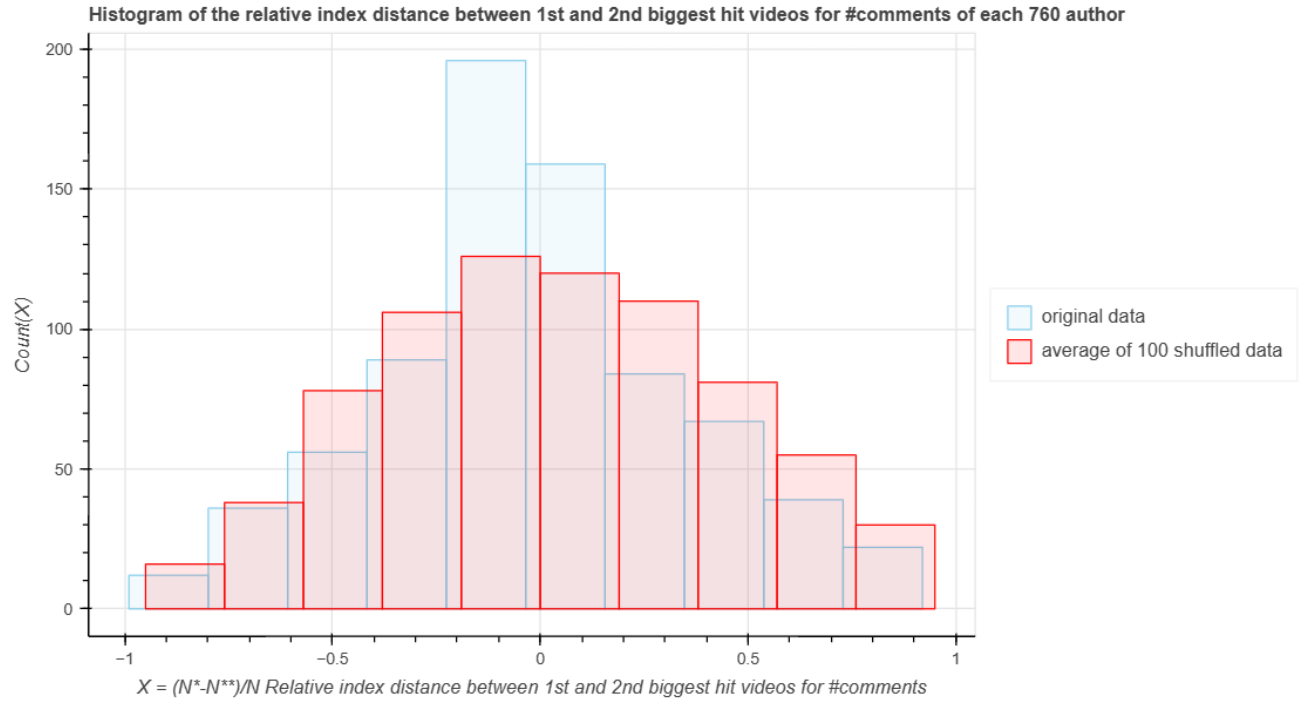



(b) Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #shares

Figure 6.13: Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to (a) #likes, and (b) #shares

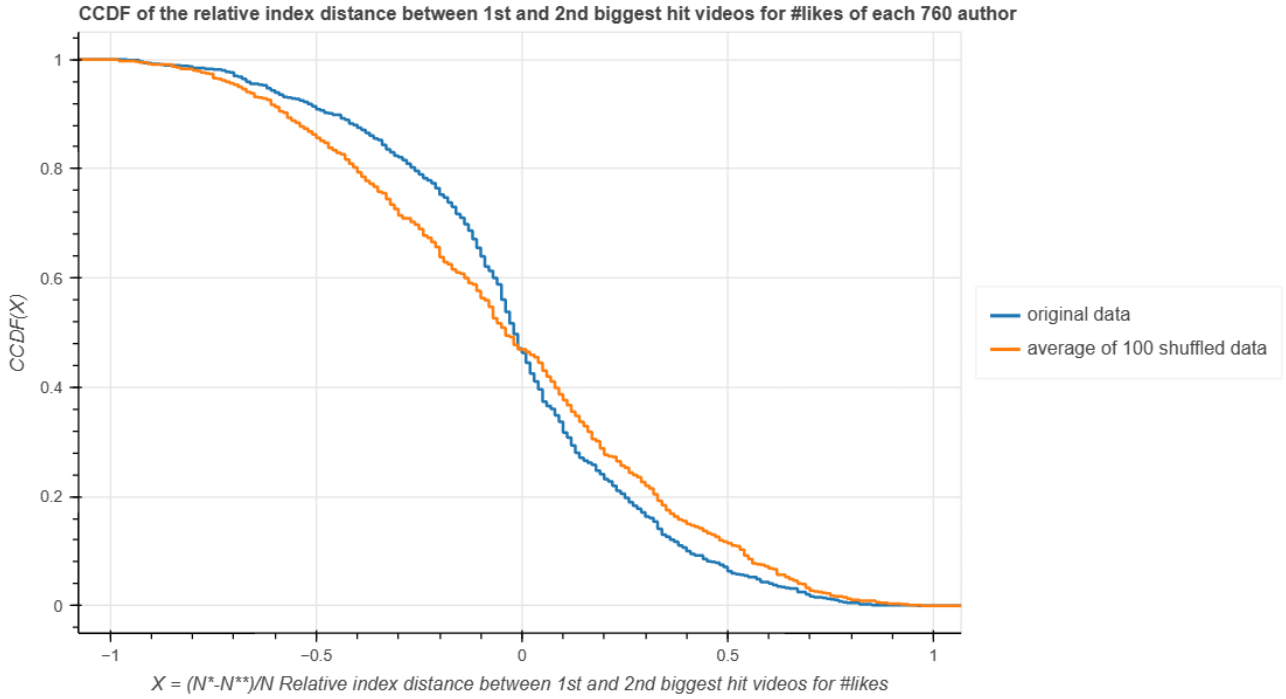



(a) Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #plays

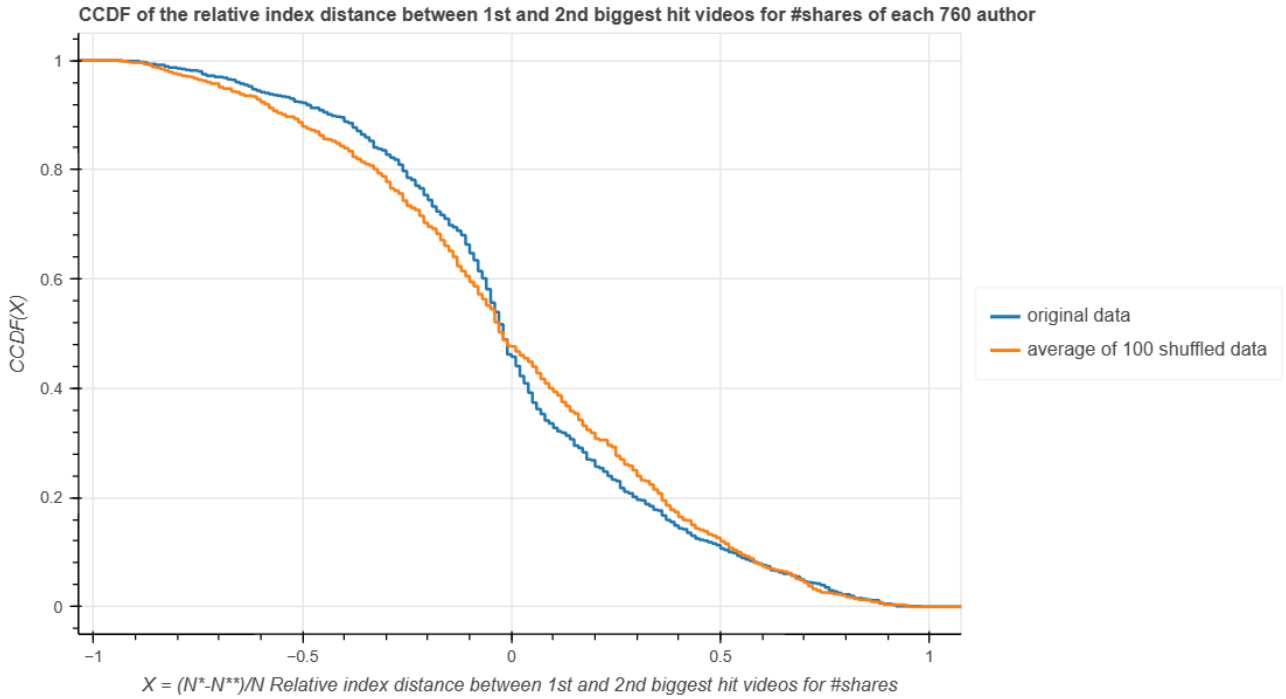



(b) Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #comments

Figure 6.14: Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to (a) #plays, and (b) #comments

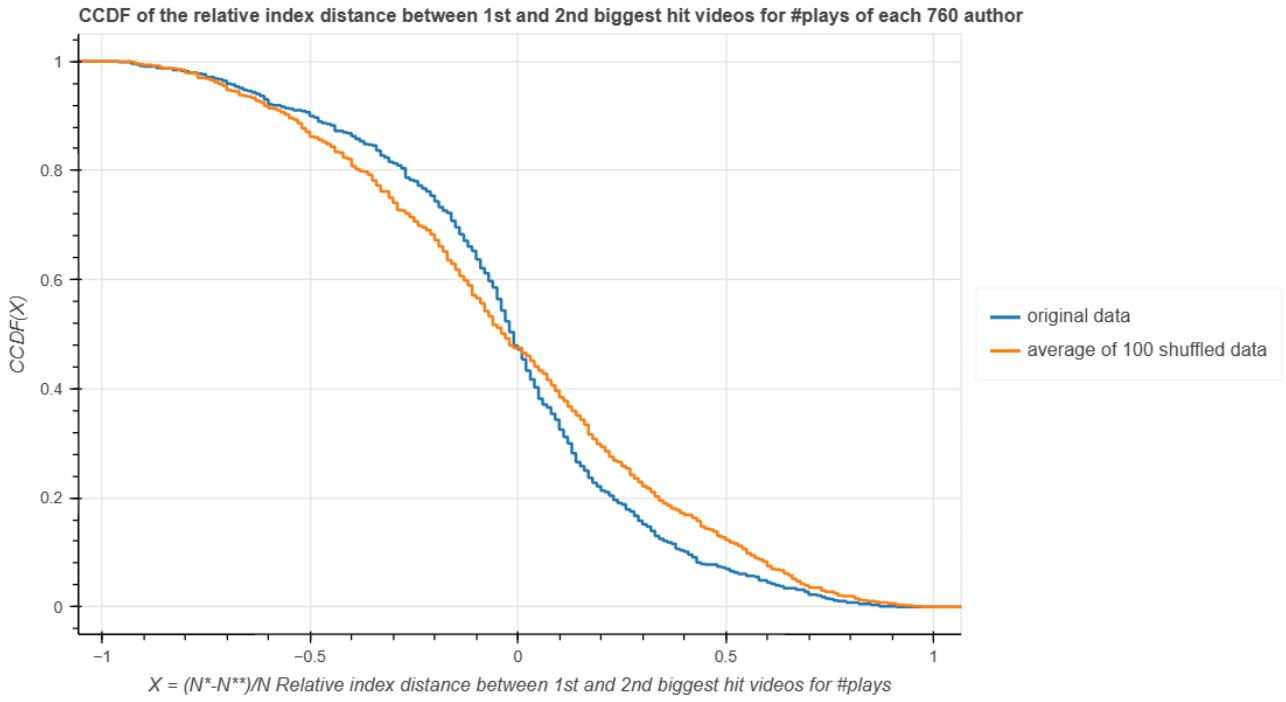



(a) CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #likes

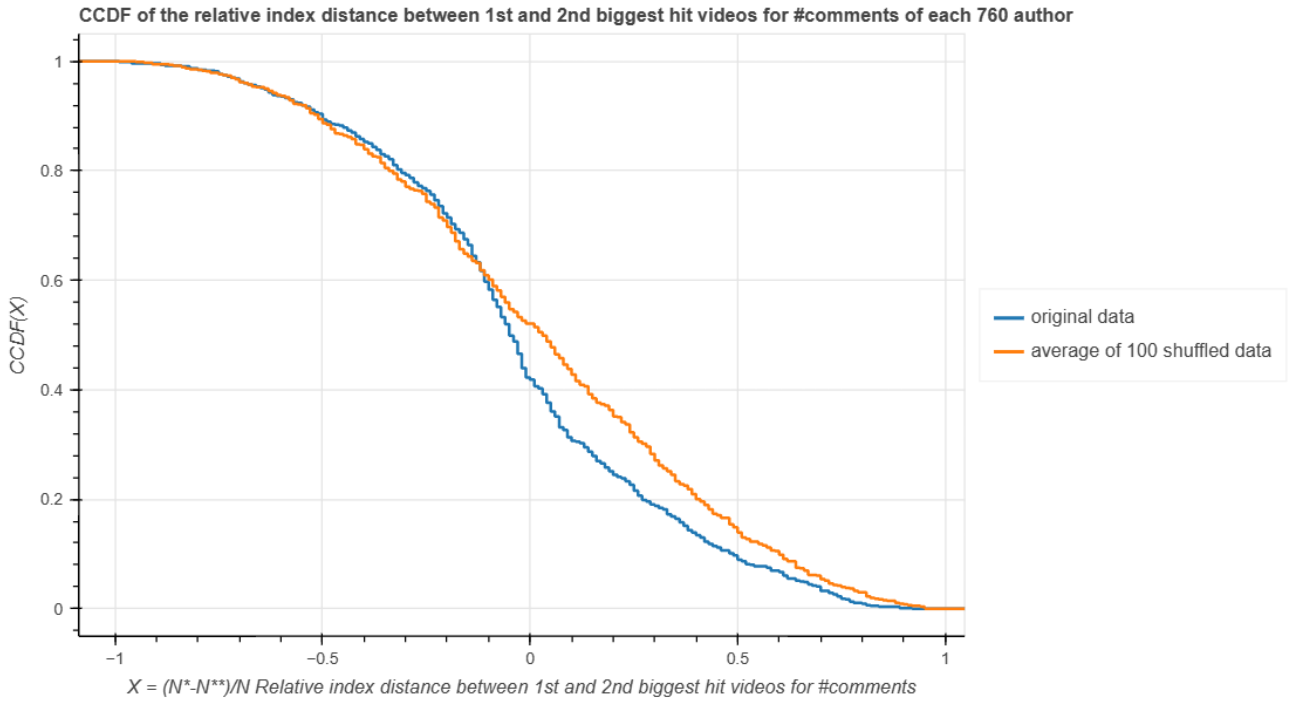



(b) CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #shares

Figure 6.15: CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to (a) #likes, and (b) #shares

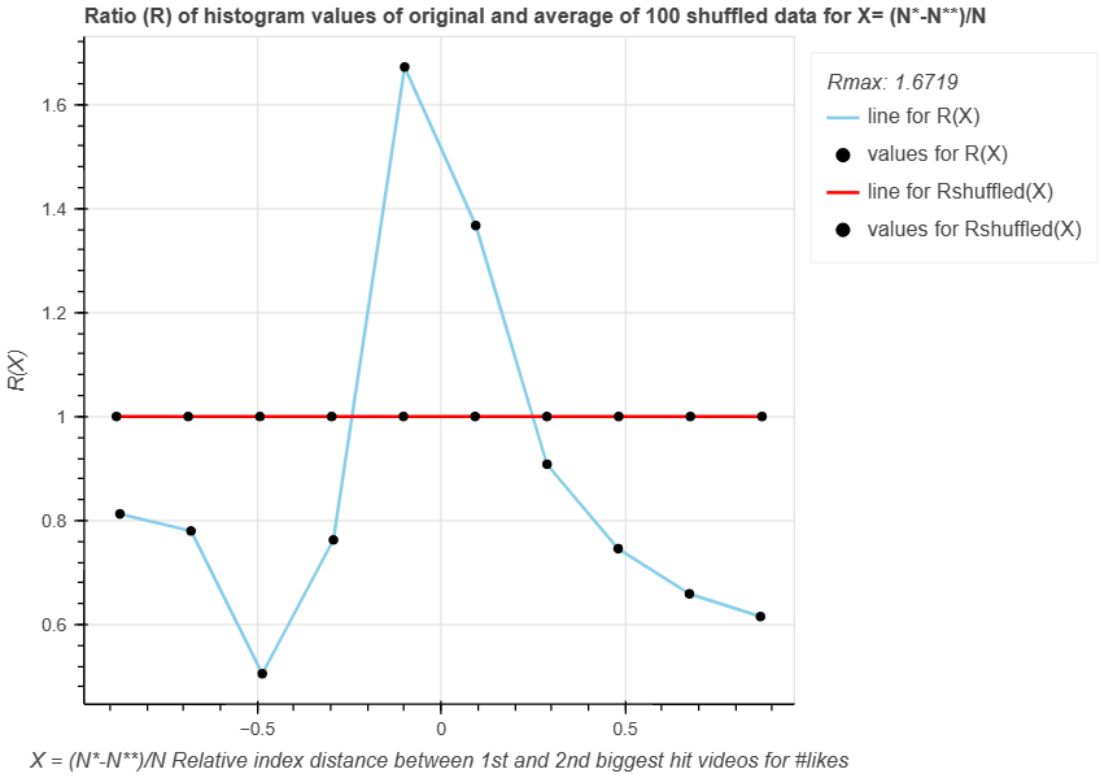



(a) CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #plays

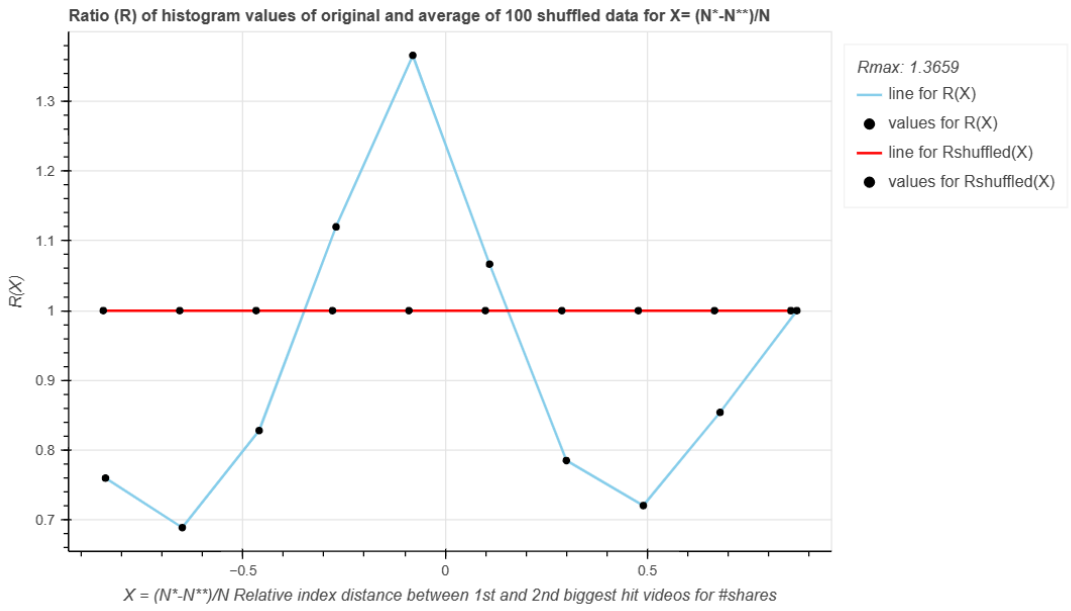



(b) CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #comments

Figure 6.16: CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to (a) #plays, and (b) #comments

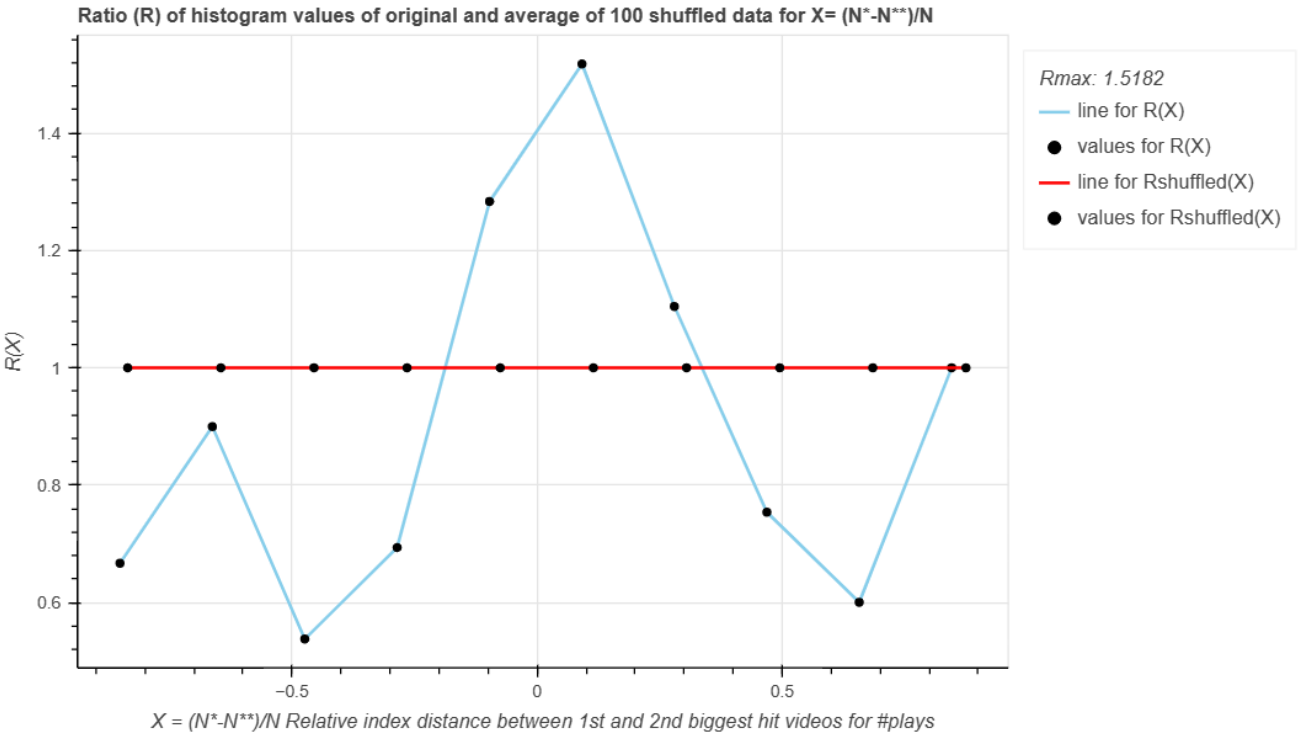



(a) Ratio of histograms distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #likes

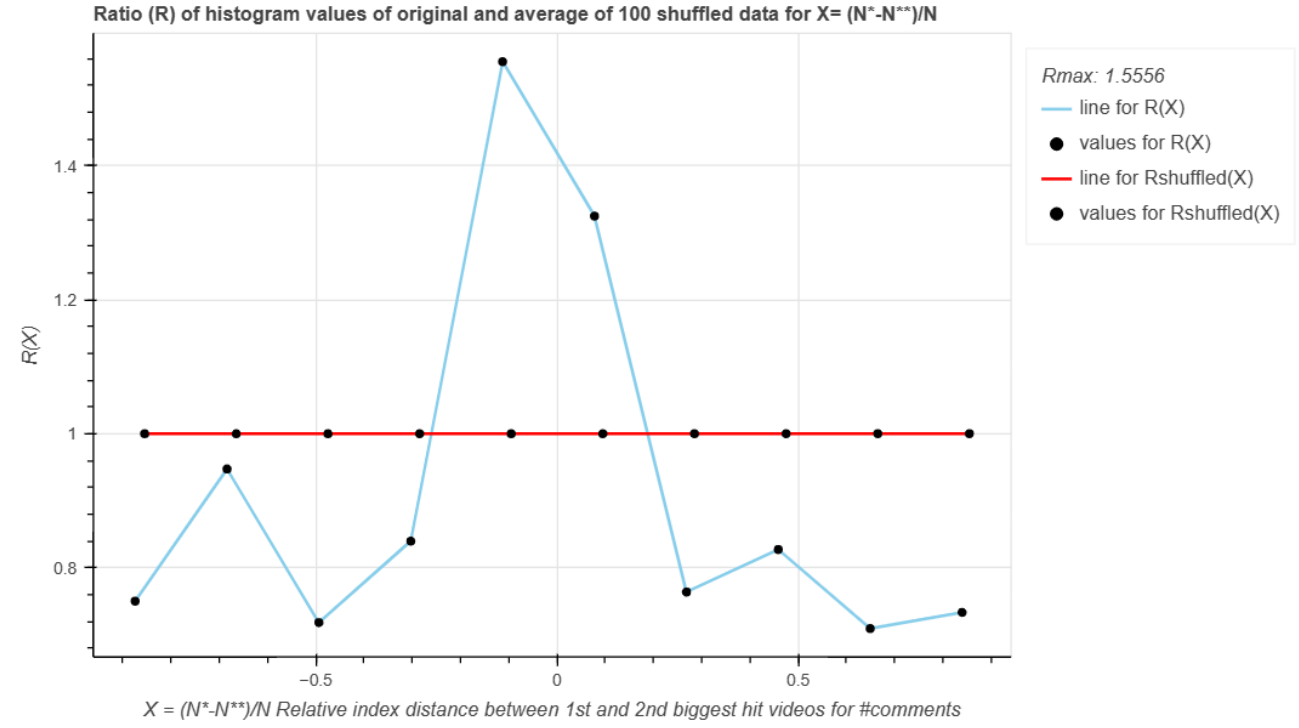



(b) Ratio of histograms distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #shares

Figure 6.17: Ratio of histograms distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to (a) #likes, and (b) #shares




(a) Ratio of histograms distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #plays




(b) Ratio of histograms distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #comments

Figure 6.18: Ratio of histograms distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to (a) #plays, and (b) #comments

### Analysis for the 25 shuffled datasets and the original dataset

To compare the 25 shuffled datasets' distributions with the original dataset for the relative distance between the timing of the highest-impact video and the 2nd highest-impact video according to all video-level popularity metrics, histogram, CCDF, and Rmax plots were built. Additionally, the results of the Mann-Whitney U Test have been shown in figures as the histogram distributions of the p-values for each of the 25 shuffled datasets.

In Figures 6.19(a), 6.19(b), 6.20(a), 6.20(b), 6.21(a), 6.21(b), 6.22(a), and 6.22(b) for the original dataset, it is observed that there are more negative values for $\Delta N/N$ for the "number of likes", the "number of shares" and the "number of comments" metrics. In contrast, distributions have more positive values for the number of plays metric.

However, on the same figures, when the shuffled datasets distributions are examined, for the "number of likes", and the "number of plays" metrics, the distributions' tendencies towards having more negative or positive values are changing. Moreover, the distributions for the "number of plays" are generally more balanced and not close to the zero point much as in the original dataset. Also, there are no high peaks around zero like in the original dataset for the "number of plays" metrics' shuffled data distributions. For the "number of shares" and the "number of comments" metrics, the distributions' tendencies are generally negative.

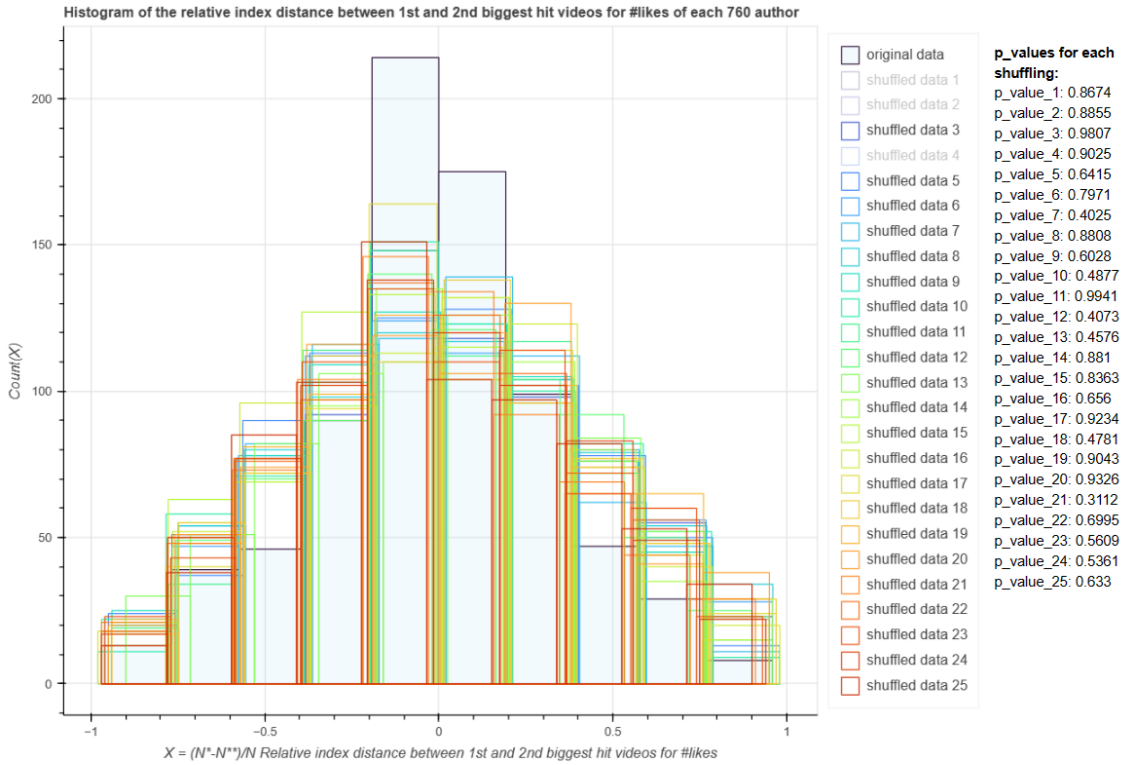
Rmax represents the maximum value of the $R(\Delta N/N)$ for each of the 25 shuffled datasets. If Rmax is equal to 1, then the original dataset's distribution for $\Delta N/N$, is similar to the shuffled dataset's distribution $\Delta Ns/N$, and we cannot say that the TikTok influencers' can experience average success after their biggest hit videos.

The Rmax plots for the metrics: the "number of likes", the "number of comments", the "number of plays" in Figures 6.23(a), 6.24(b), 6.24(a), show that all Rmax values are above 1. However, the Rmax plot for the "number of shares" in Figure 6.23(b) has values mostly above 1. So, the Rmax values, most of which are above 1, can imply a hot streak phenomenon for the TikTok careers.

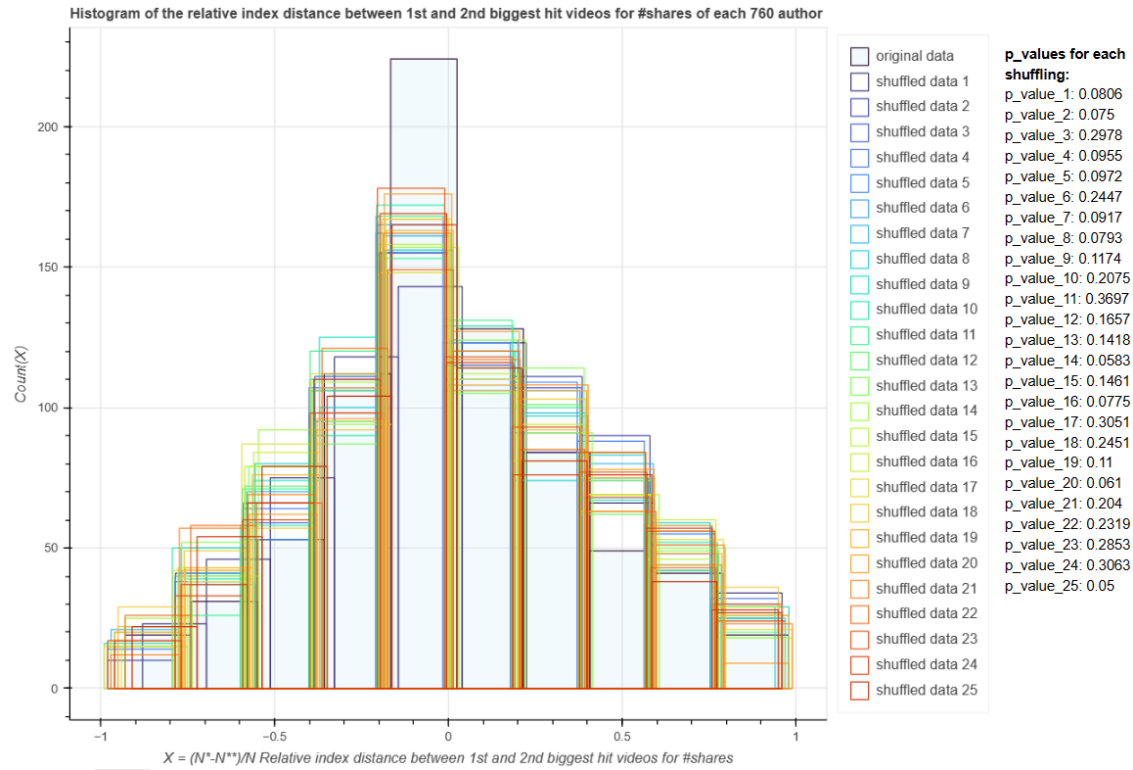
The results of the Mann-Whitney U Tests are shown in Figures 6.25(a),6.25(b),6.26(a), and 6.26(b) as histogram plots. The minimum p-values for each popularity metric are below.

- min[p-value(number of likes)] = approx. 0.3
- min[p-value(number of shares)] = 0.05
- min[p-value(number of comments)] = approx. 0.32
- min[p-value(number of plays)] = approx. 0.25

The p-values that are above 0.05 denote that the original data and the shuffled datasets come from the same distribution and that there is not a significant difference between them.

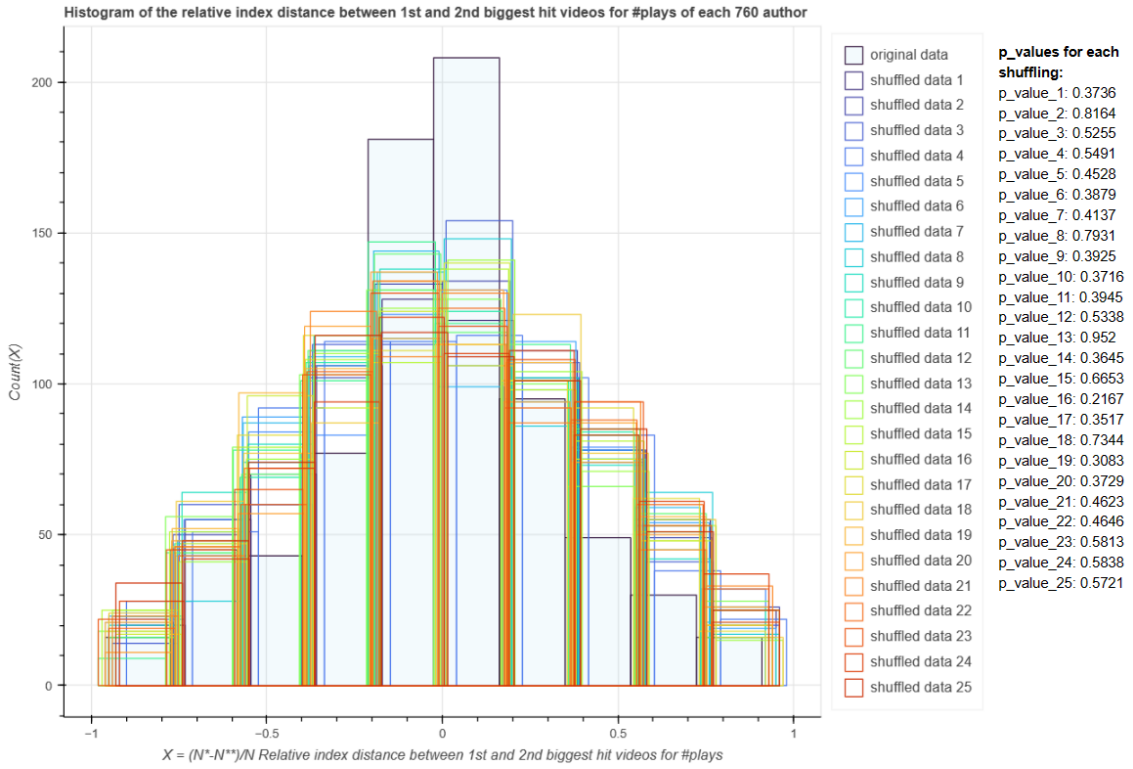



(a) Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #likes

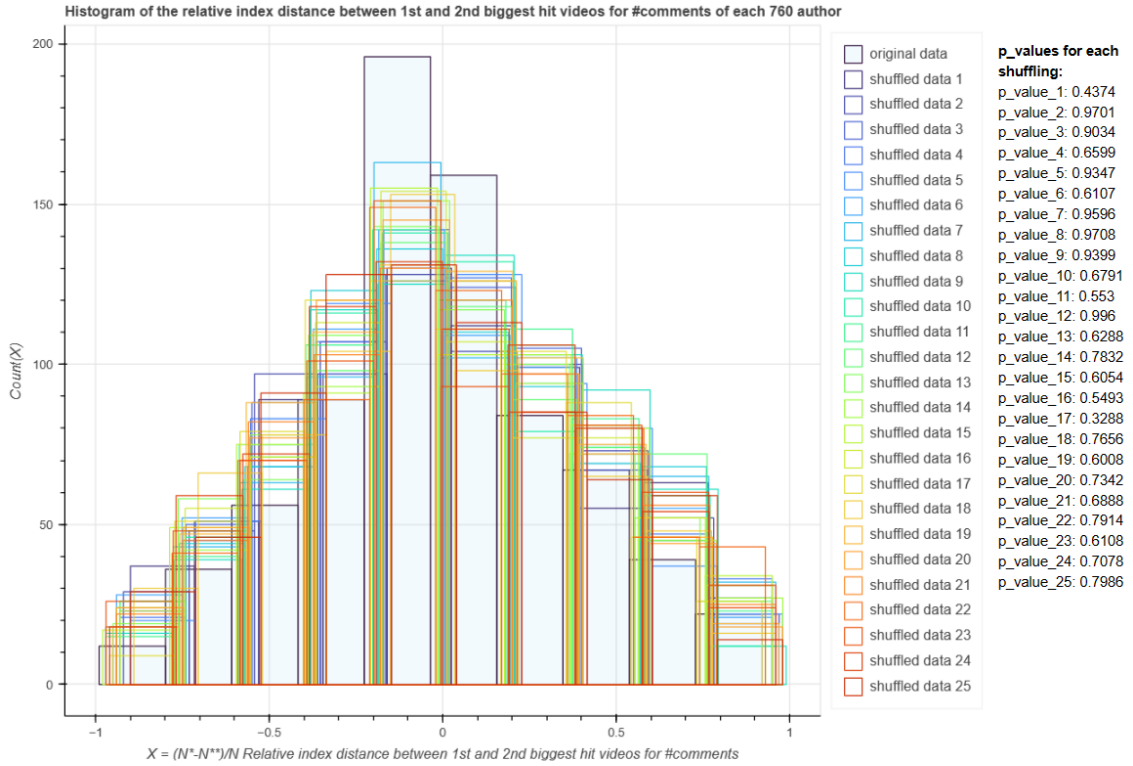



(b) Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #shares

Figure 6.19: Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers (25 shuffled datasets) according to (a) #likes, and (b) #shares

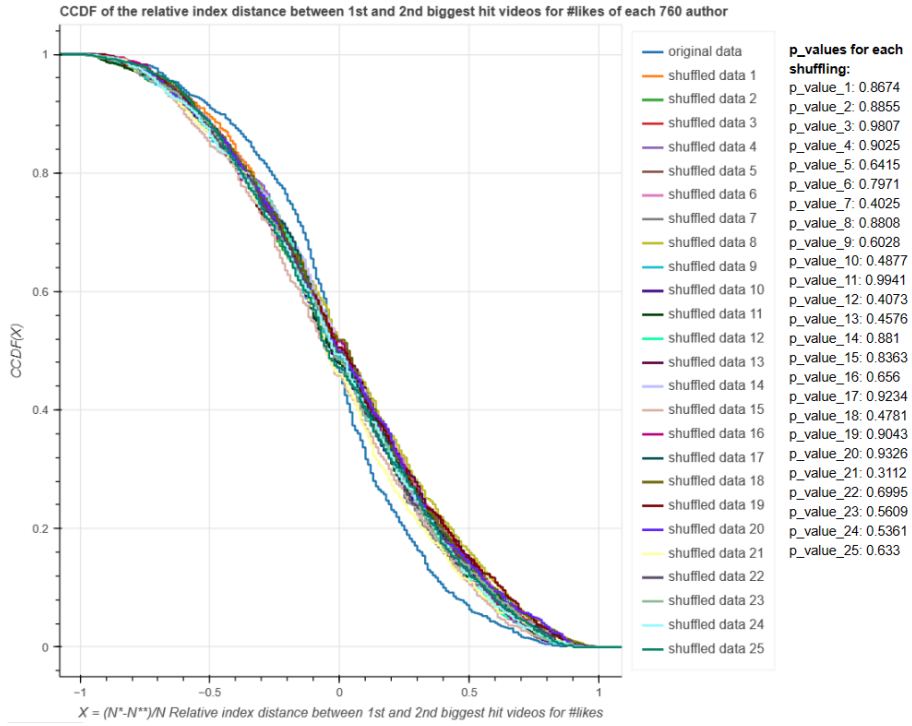



(a) Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #plays

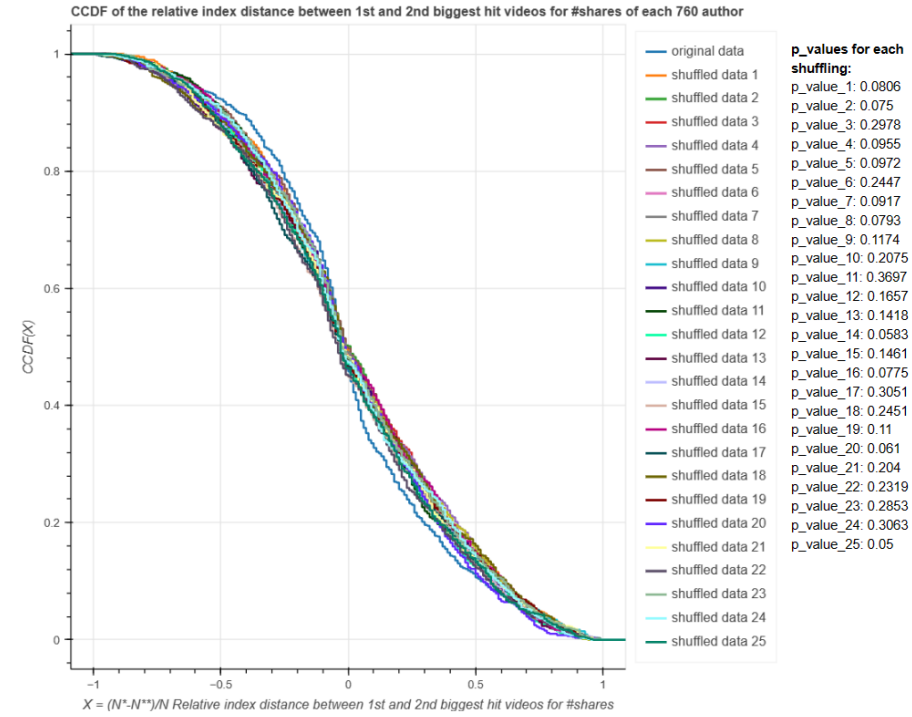



(b) Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #comments

Figure 6.20: Histogram distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers (25 shuffled datasets) according to (a) #plays, and (b) #comments

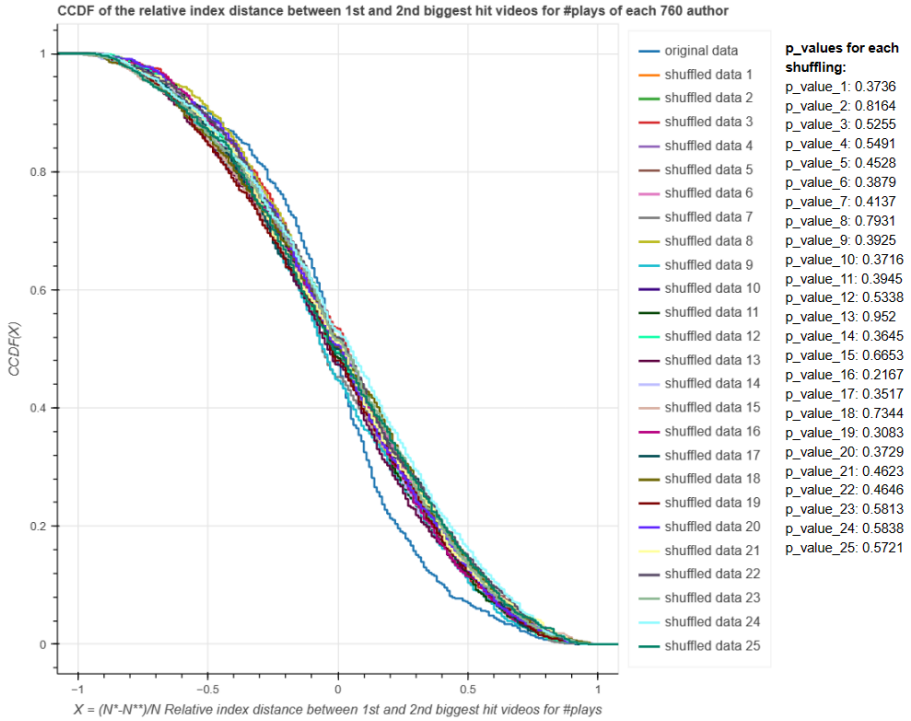



(a) CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #likes

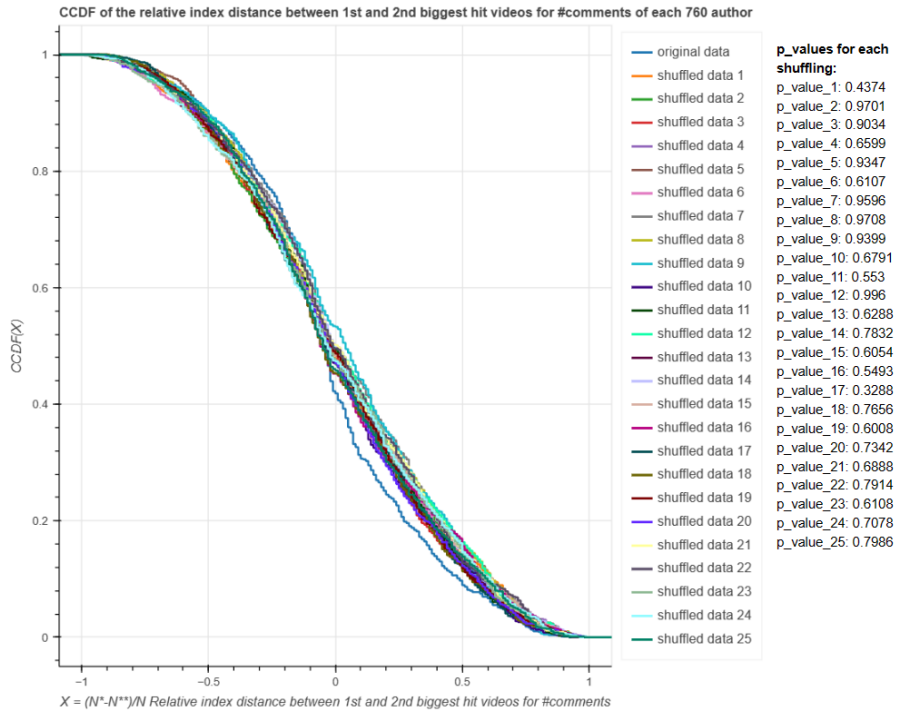



(b) CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #shares

Figure 6.21: CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers (25 shuffled datasets) according to (a) #likes, and (b) #shares

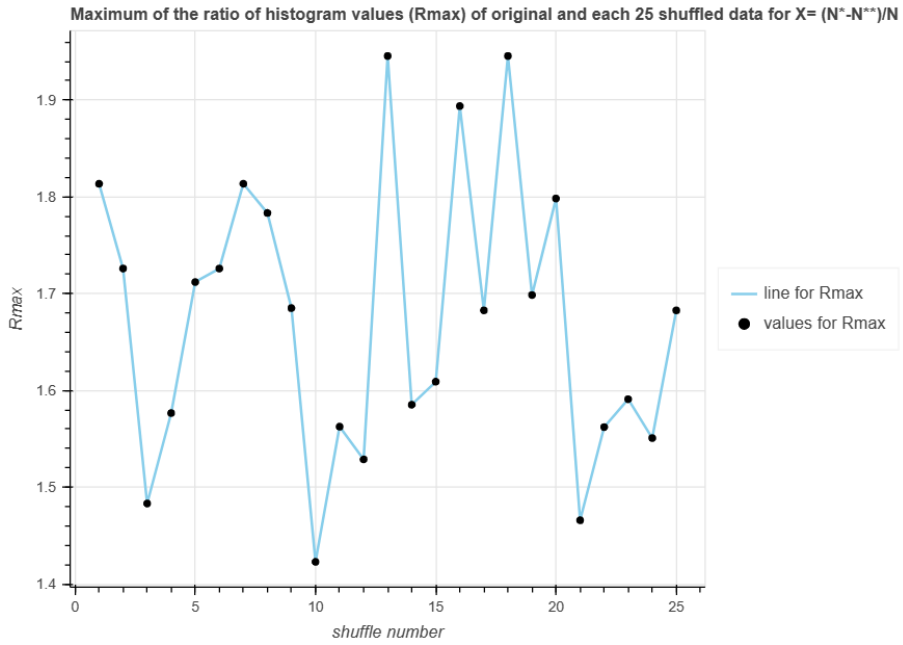



(a) CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #plays

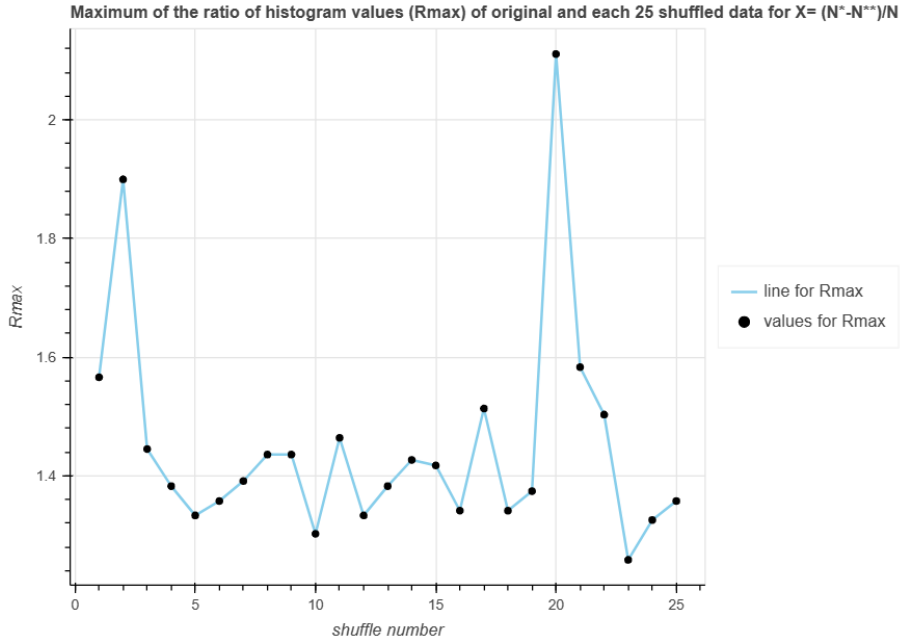



(b) CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #comments

Figure 6.22: CCDF distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers (25 shuffled datasets) according to (a) #plays, and (b) #comments

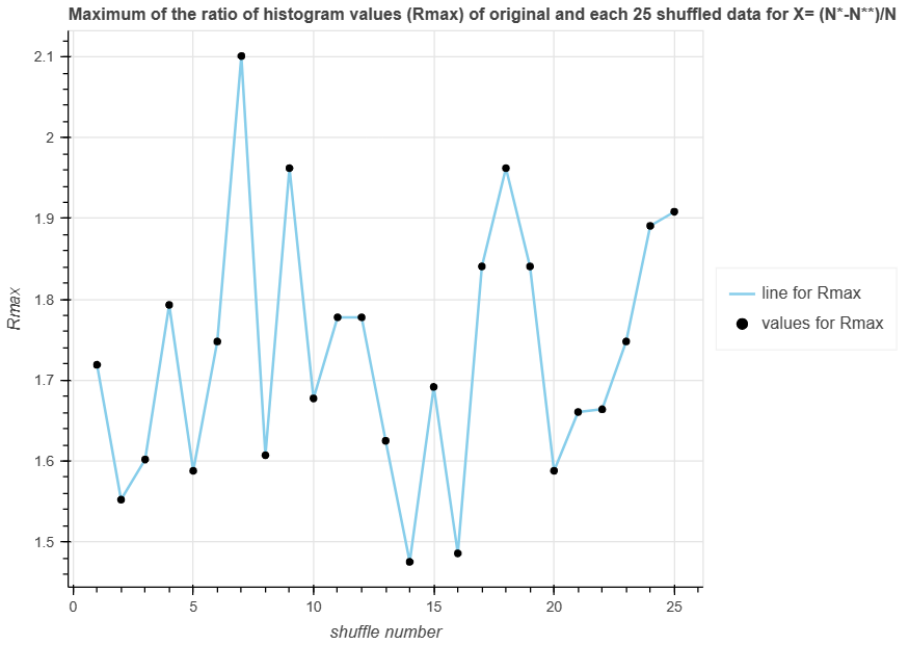



(a) Rmax distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #likes

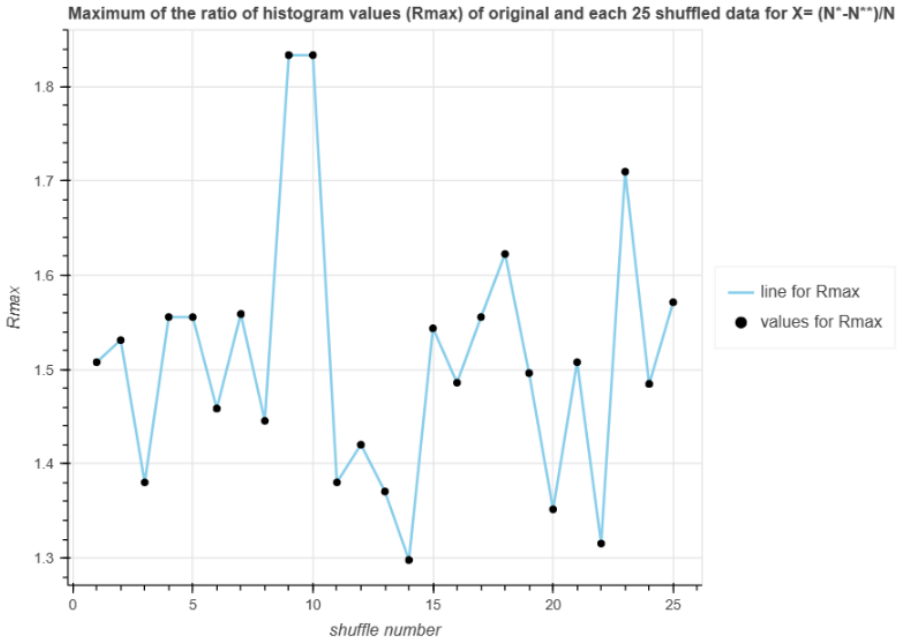



(b) Rmax distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #shares

Figure 6.23: Rmax distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers (25 shuffled datasets) according to (a) #likes, and (b) #shares

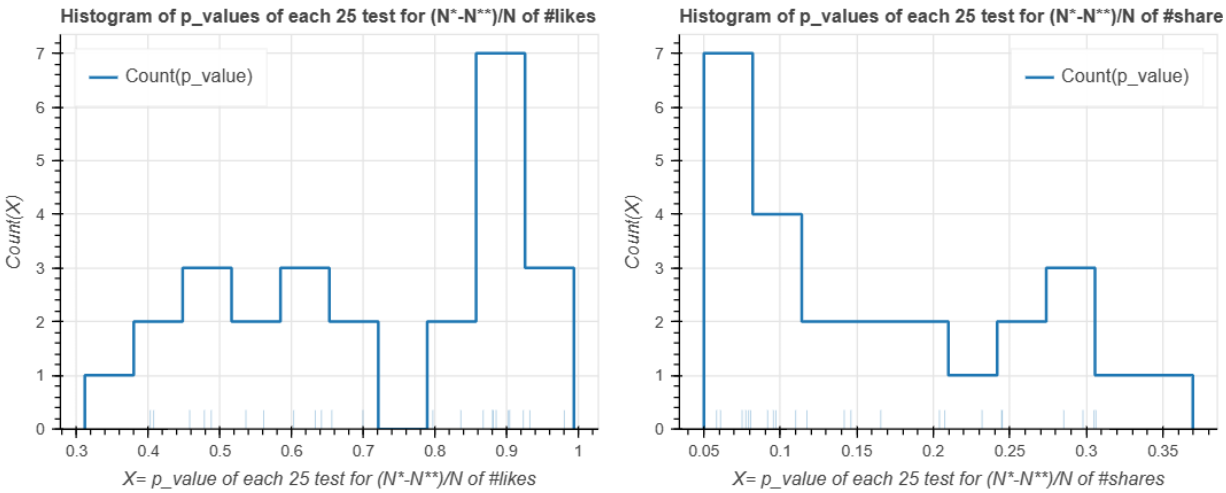



(a) Rmax distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #plays



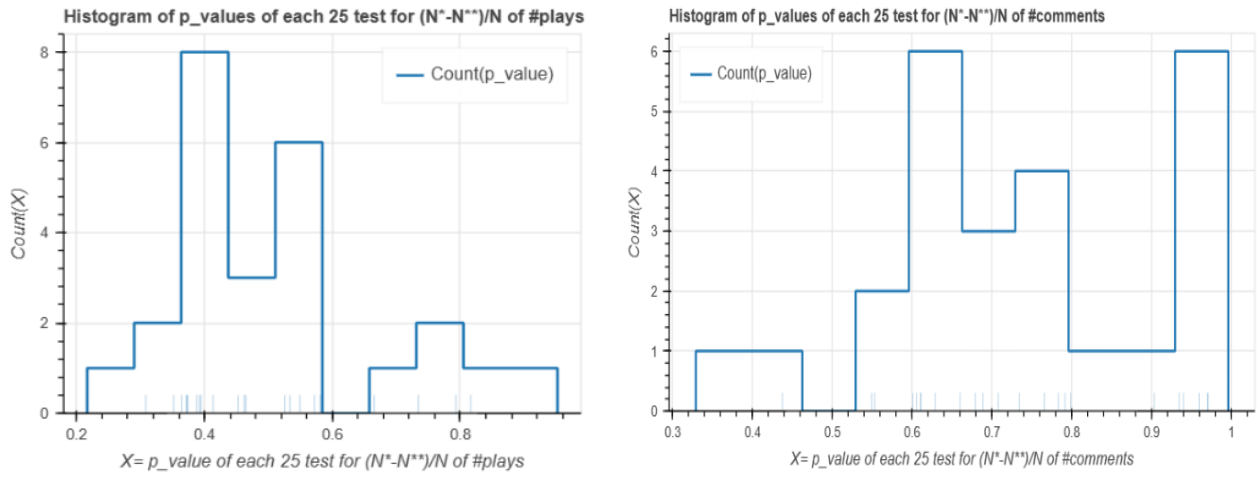

(b) Rmax distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers according to #comments

Figure 6.24: Rmax distribution of the relative index distance between the timing of the highest-impact video and the 2nd highest-impact video for real, and randomized careers (25 shuffled datasets) according to (a) #plays, and (b) #comments




(a) Histogram distribution of the p_values for #likes (b) Histogram distribution of the p_values for #shares

Figure 6.25: Histogram distribution of the p_values of the Mann Whitney U Test results regarding the biggest hit's relative video index distribution comparison for real, and randomized careers (25 separate shuffled dataset) according to (a) max(#likes), and (b) max(#shares)




(a) Histogram distribution of the p_values for #plays (b) Histogram distribution of the p_values for #comments

Figure 6.26: Histogram distribution of the p_values of the Mann Whitney U Test results regarding the biggest hit's relative video index distribution comparison for real, and randomized careers (25 separate shuffled dataset) according to (a) max(#plays), and (b) max(#comments)

## 6.4 Results of the “Understanding the onset of a hot streak” phase

The third research question was whether a pattern signals the start of a hot streak phase, such as the exploitation of topics during a hot streak followed by an exploration of topics before the hot streak.

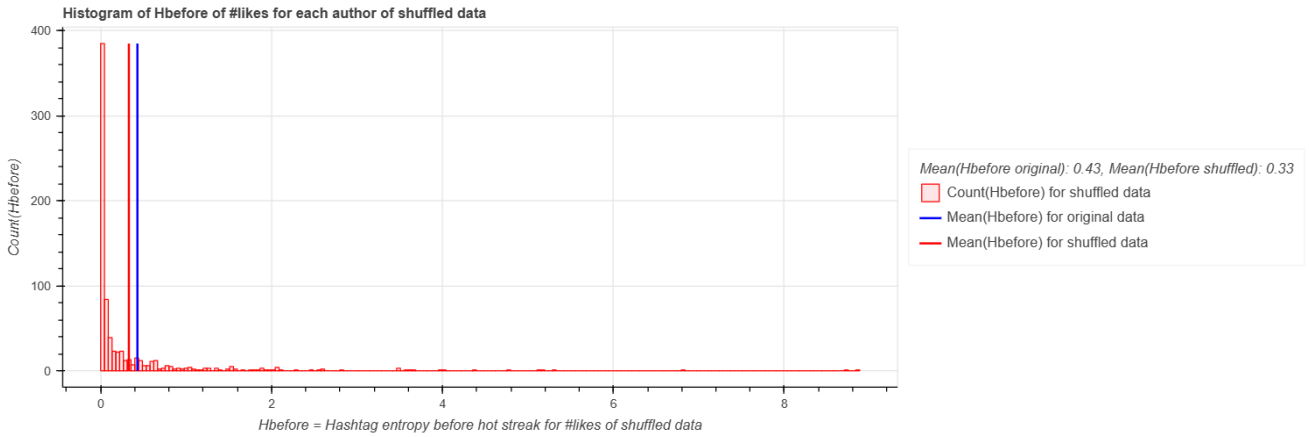
Accordingly, hashtag entropies before and during hot streak periods for the video-level popularity metrics “number of likes”, “number of shares”, “number of plays”, and “number of comments” were calculated with a method described in section 4.4.2.

For the aforementioned video-level popularity metrics, as seen in Figures 6.28(a), 6.28(b), 6.29(a), 6.29(b), 6.30(a), 6.30(b), 6.31(a), and 6.31(b), the mean of the hashtag entropy of the original data during the hot streak period is larger than before it, except for the “number of likes” metric for which there is almost no difference. In randomized careers, the mean of the hashtag entropy of the “average of 100 times shuffled data” during the hot streak period is larger than before it for all metrics. All the comparisons are listed in Figure 6.27

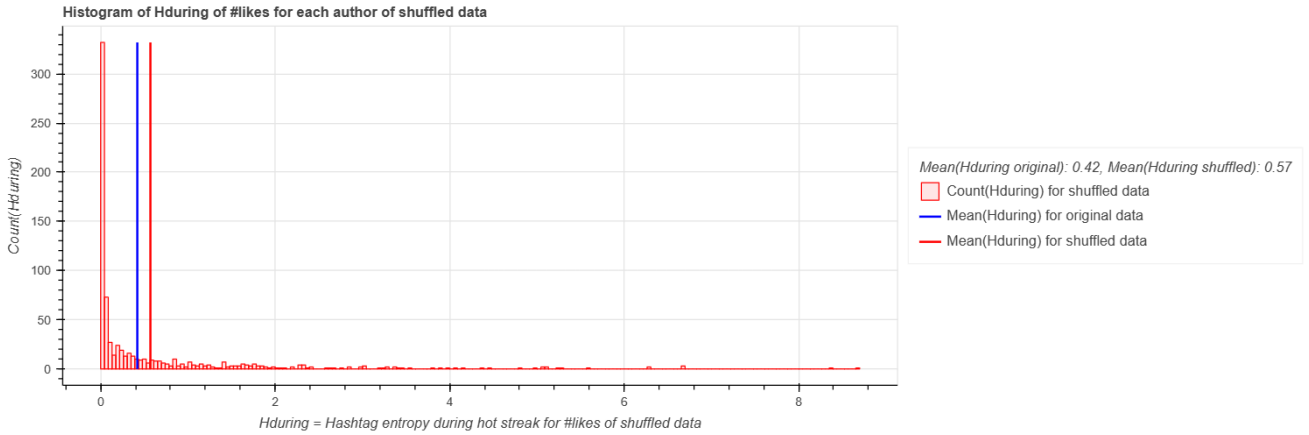
Therefore, it can be asserted that there is more exploration of topics during the hot streak period than before it for the careers of social media content creators on TikTok. That is, TikTok authors try different hashtags more on their content during their hot streak periods than before them. This outcome is similar to the finding of the research conducted by Garimella et al. on Twitter data [8]. As a result, the pattern is different from the one found for the careers of artists, film directors, and scientists, which states that there is more exploration of topics before the hot streak period, while during the hot streak period, the exploitation of topics has been observed [16].

| | #likes | #shares | #plays | #comments |
|---|---|---|---|---|
| Mean(Hbefore original) | **0.43** | 0.39 | 0.38 | 0.38 |
| Mean(Hduring original) | 0.42 | **0.43** | **0.48** | **0.45** |
| Mean(Hbefore shuffled) | 0.33 | 0.34 | 0.33 | 0.33 |
| Mean(Hduring shuffled) | **0.57** | **0.51** | **0.52** | **0.54** |

Figure 6.27: Mean of hashtag entropies before and during the hot streak period for original and shuffled data, separately, according to the video-level popularity metrics

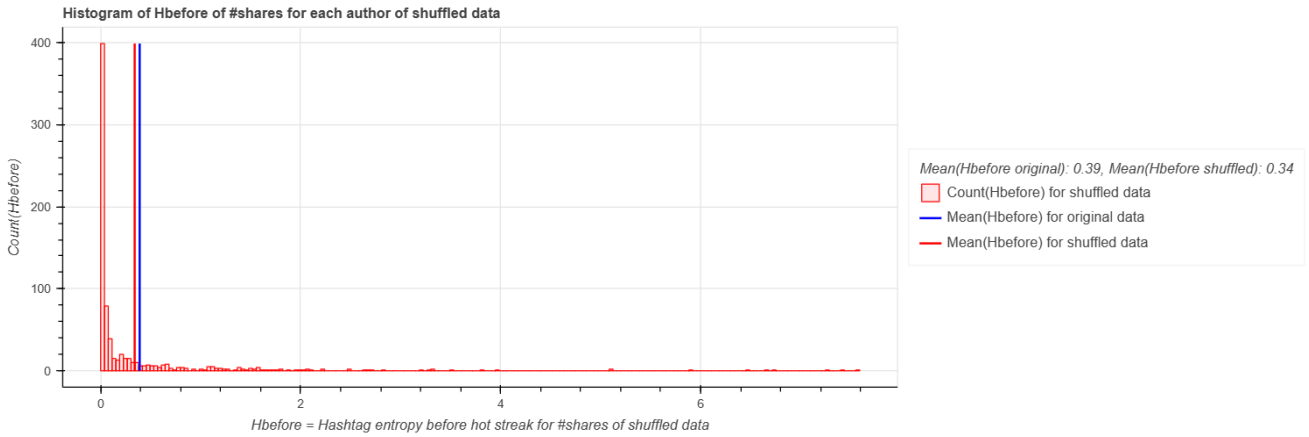



(a) Histogram distribution of the hashtag entropy before hot streak for #likes

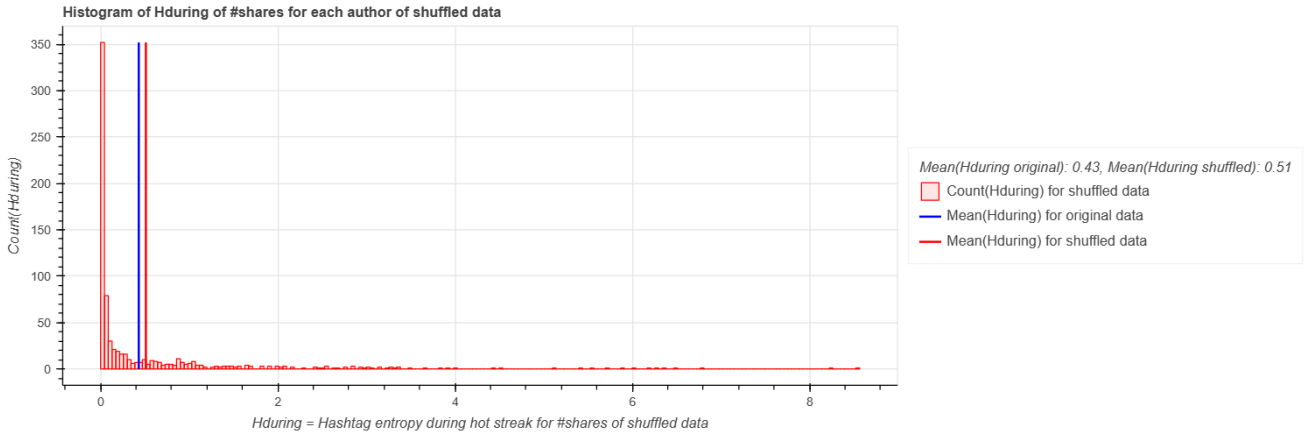



(b) Histogram distribution of the hashtag entropy during hot streak for #likes

Figure 6.28: Histogram distribution of the hashtag entropy before (a) and during (b) hot streaks for average of real, and randomized careers (100 times shuffled dataset) according to #likes

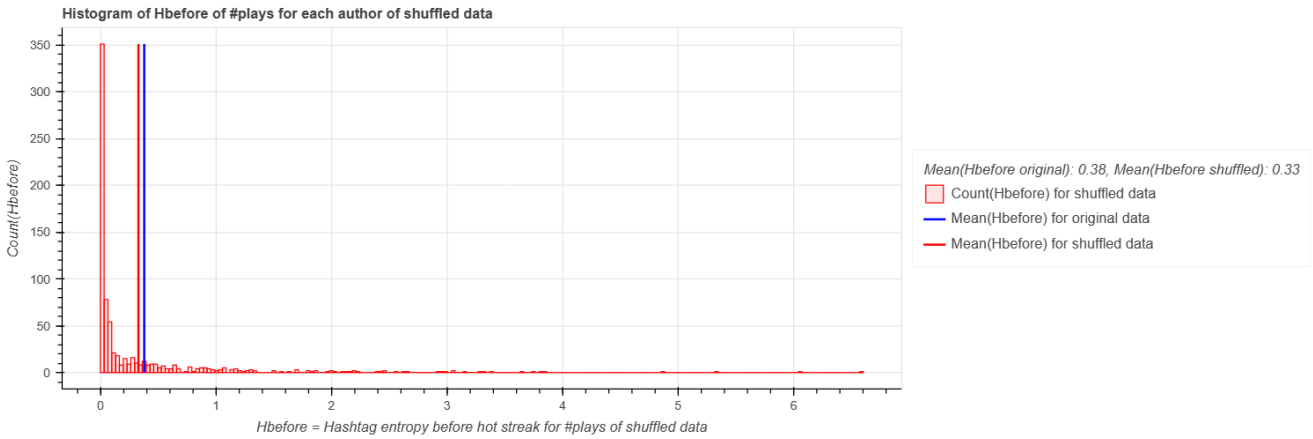



(a) Histogram distribution of the hashtag entropy before hot streak for #shares

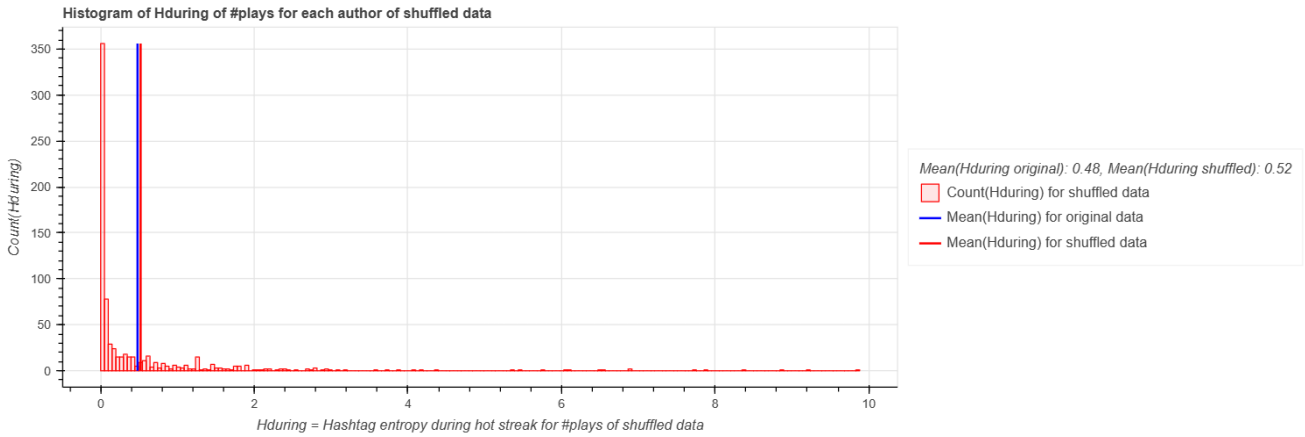



(b) Histogram distribution of the hashtag entropy during hot streak for #shares

Figure 6.29: Histogram distribution of the hashtag entropy before (a) and during (b) hot streaks for average of real, and randomized careers (100 times shuffled dataset) according to #shares

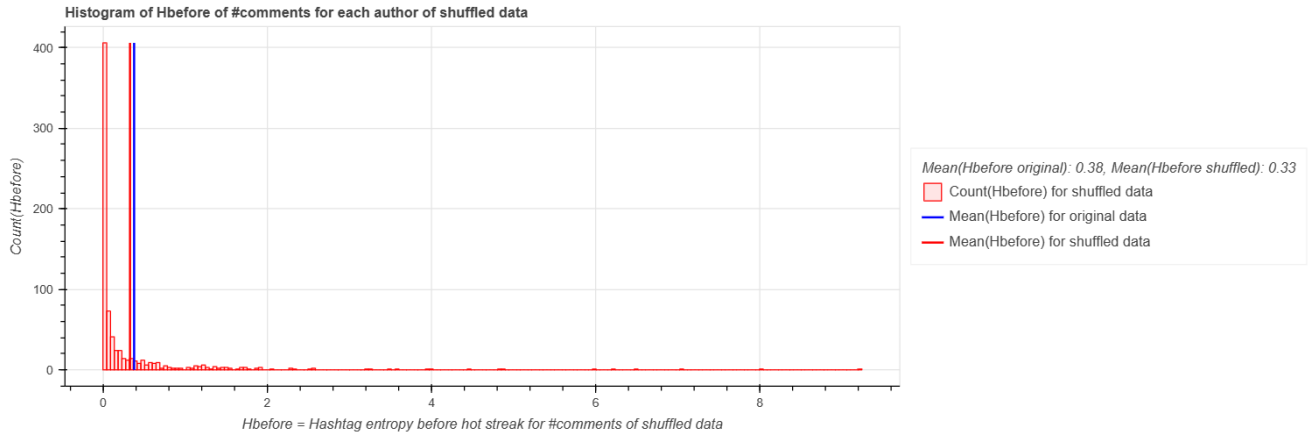



(a) Histogram distribution of the hashtag entropy before hot streak for #plays

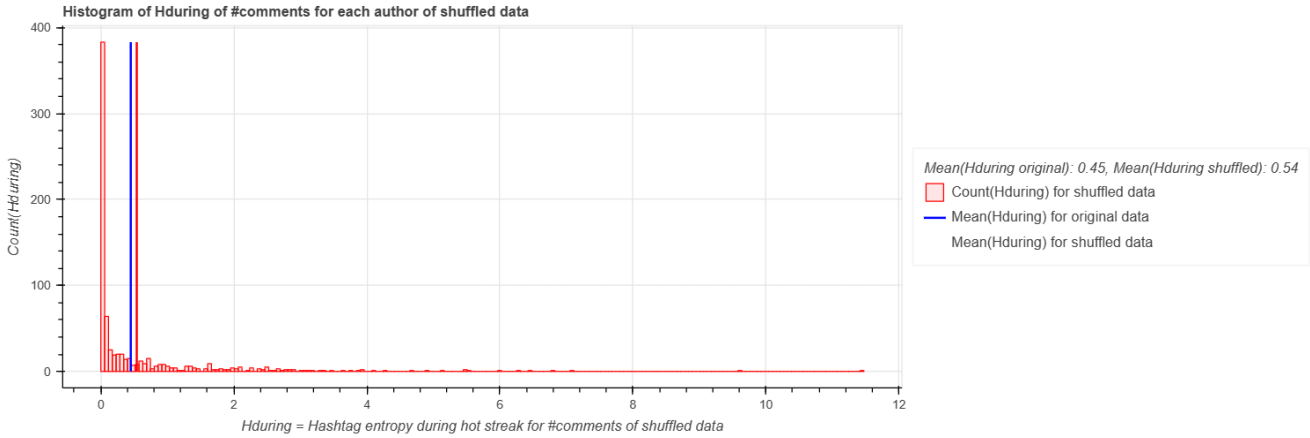



(b) Histogram distribution of the hashtag entropy during hot streak for #plays

Figure 6.30: Histogram distribution of the hashtag entropy before (a) and during (b) hot streaks for average of real, and randomized careers (100 times shuffled dataset) according to #plays




(a) Histogram distribution of the hashtag entropy before hot streak for #comments




(b) Histogram distribution of the hashtag entropy during hot streak for #comments

Figure 6.31: Histogram distribution of the hashtag entropy before (a) and during (b) hot streaks for average of real, and randomized careers (100 times shuffled dataset) according to #comments

## 6.5 Results of the “Visualizing hot streaks of users” phase

In this part of the thesis, I aimed to visualize the hot streak durations of users, and the timing of the the highest, the 2nd highest, and the 3rd highest impact videos of the users during their hot streak periods. Additionally, I wanted to identify how many users experience the biggest hit in early, middle, or late careers, to support the results of the “Timing of an influencer’s hit” phase. I determined the x-axis as “index of the videos”, and y-axis as “index of the users”. Both axes are denoted in Figures 6.33(a), and 6.33(b).

Among the video-level popularity distributions, since the “total number of plays per author” and the “total number of likes per author” have the widest ranges, the “number of plays”, and the “number of likes” metrics were chosen to examine the hot streaks durations of users. The tabs related to these metrics are shown in Figures 6.48, and 6.49, for #likes and #plays respectively.

When a metric is chosen to sort the results in the visualization, it also displays the histogram distribution of the metric associated with the sort option, as seen in Figures 6.34, 6.35, 6.36, 6.37, 6.38. These histogram distributions are evaluated as below:

First, as shown in Figure 6.32, most of the users have less than 150 videos.

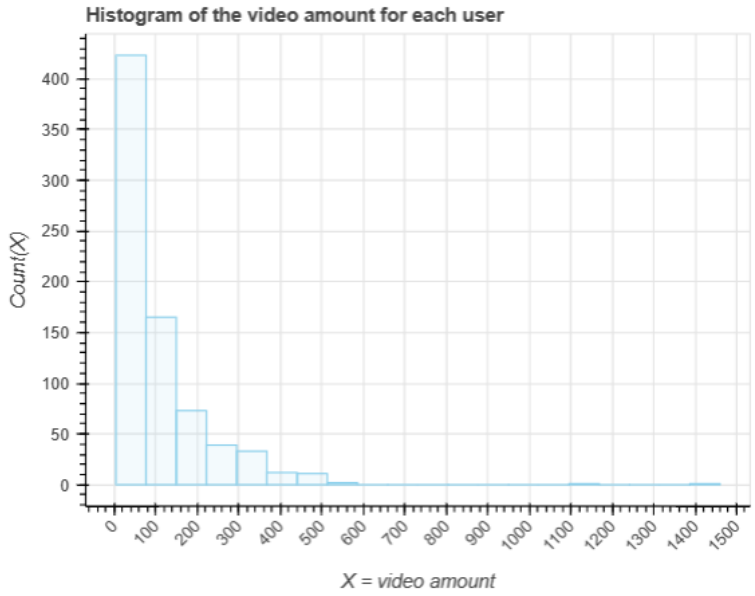



Figure 6.32: Histogram distribution of the video amount of users

Second, according to the “number of likes” metric, the results are below.

- As seen in Figures 6.39(a), and 6.39(b), most users have a length of hot streak duration of less than 50. The duration of the hot streak is generally between the $12\%$ and $50\%$ percent of their career.
- As indicated in Figure 6.41(a), most users have the biggest hit for the video whose index is less than 50.
- As denoted in Figure 6.41(b), there are more users who have their biggest hit in their early career periods.

Third, according to the "number of plays" metric, the results are as below.

- As seen in Figures 6.40(a), and 6.40(b), most users have a length of hot streak duration of less than 40. The duration of the hot streak is generally between the $10\%$ and $50\%$ percent of their career.
- As indicated in Figure 6.42(a), most users have the biggest hit for the video, whose index is less than 50.
- As denoted in Figure 6.42(b), there are more users who have their biggest hit in their early career periods.

Moreover, when a career level is chosen, the title of the plot shows the number of users who experienced the biggest hit in their early, middle, or late careers.

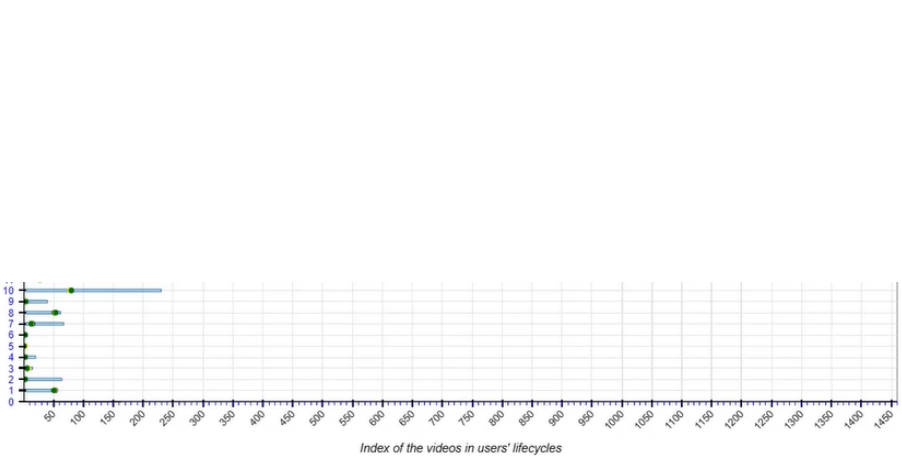
The number of users who have the biggest hit in their early careers (where early career corresponds to the [1, $\frac{N}{3}$] range of the videos of a user, as N is the total number of videos) is higher than the number of users that have the biggest hit in middle, or late career. However, as noted in Table 6.3, the numbers are not so different from each other, so it can be asserted that the biggest hit's timing is random through a TikTok influencer's lifecycle. It can appear either in early, middle, or late career.

| **Career level** | **#Users for** $\#likes$ | **#Users for** $\#plays$ |
|---|---|---|
| early | 271 | 288 |
| middle | 251 | 244 |
| late | 238 | 228 |

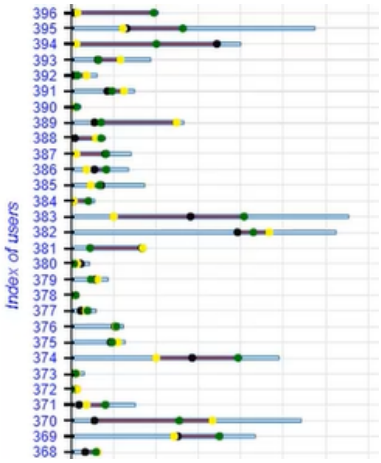
Table 6.3: Number of users who have the biggest hit in each career level for the metrics $\#likes$, and $\#plays$

Finally, when the user moves the mouse over the bars representing the author's lifecycle, or hot streak duration, or to the circles representing the timing of the hits, there is additional information shown as a tooltip. When the user selects a "hit order" from the associated dropdown, the tooltip information changes as the following:

- If "Show no hit" is chosen, the tooltip is displayed as in Figure 6.43.
- If "Biggest hit" is chosen, the tooltip shows "Video index of the biggest hit (N*)", and "Relative video index of the biggest hit (N*/N)" in addition to the information shown when "Show no hit" is chosen. It is denoted in Figure 6.44.
- If "Second biggest hit" is chosen, the tooltip shows "Video index of the 2nd biggest hit (N**)" in addition to the information shown when "Show no hit" is chosen. It is denoted in Figure 6.45.
- If "Third biggest hit" is chosen, the tooltip shows "Video index of the 3rd biggest hit (N***)" in addition to the information shown when "Show no hit" is chosen. It is denoted in Figure 6.46.
- If "Show all hits" is chosen, the tooltip shows "Video index of the biggest hit (N*)", "Relative video index of the biggest hit (N*/N)", "Video index of the 2nd biggest hit (N**)", and "Video index of the 3rd biggest hit (N***)" in addition to the information shown when "Show no hit" is chosen. It is denoted in Figure 6.47.




(a) x-axis represents the index of the videos




(b) y-axis represents the index of the users

Figure 6.33: (a) x-axis and (b) y-axis of plot for hot streak durations

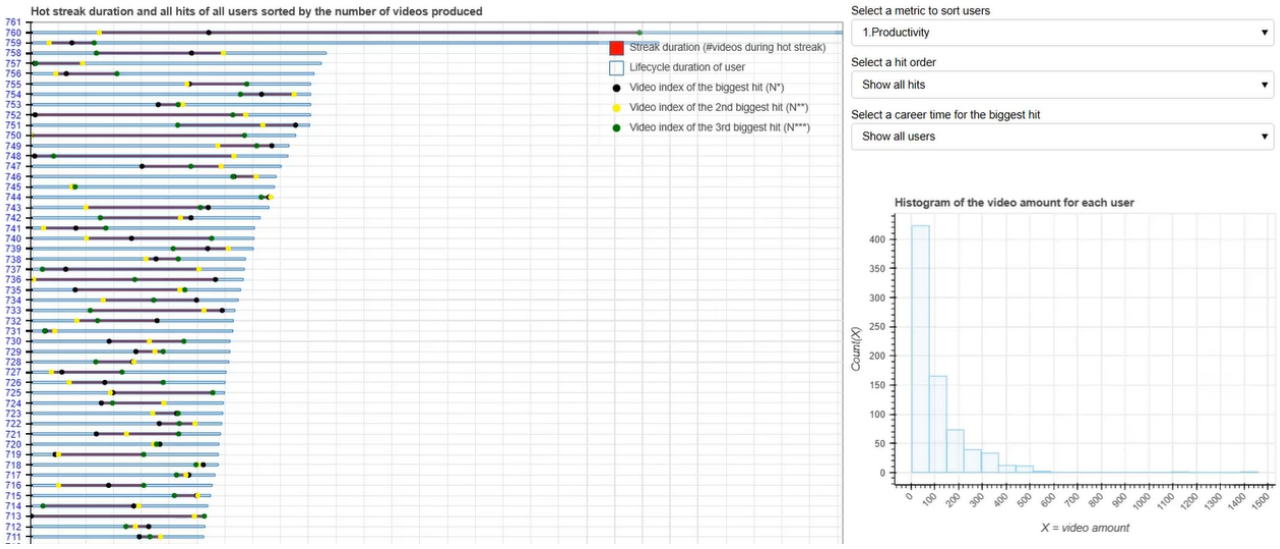



Figure 6.34: Sorting by the number of videos produced and histogram of the video amount of each user

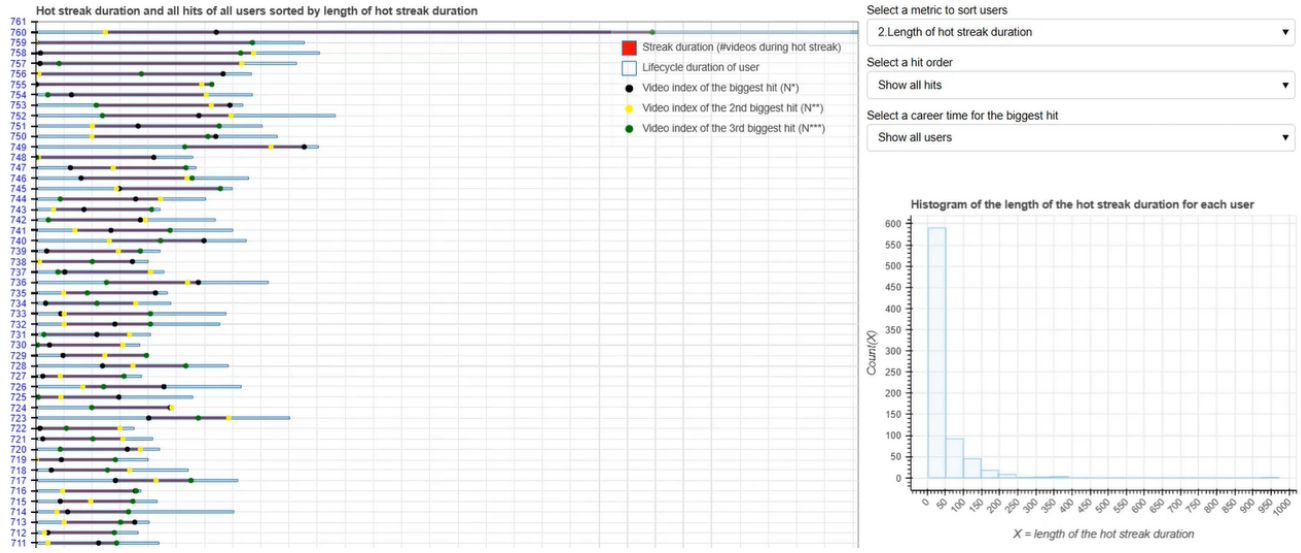



Figure 6.35: Sorting by the length of the hot streak duration and histogram of the length of the hot streak duration

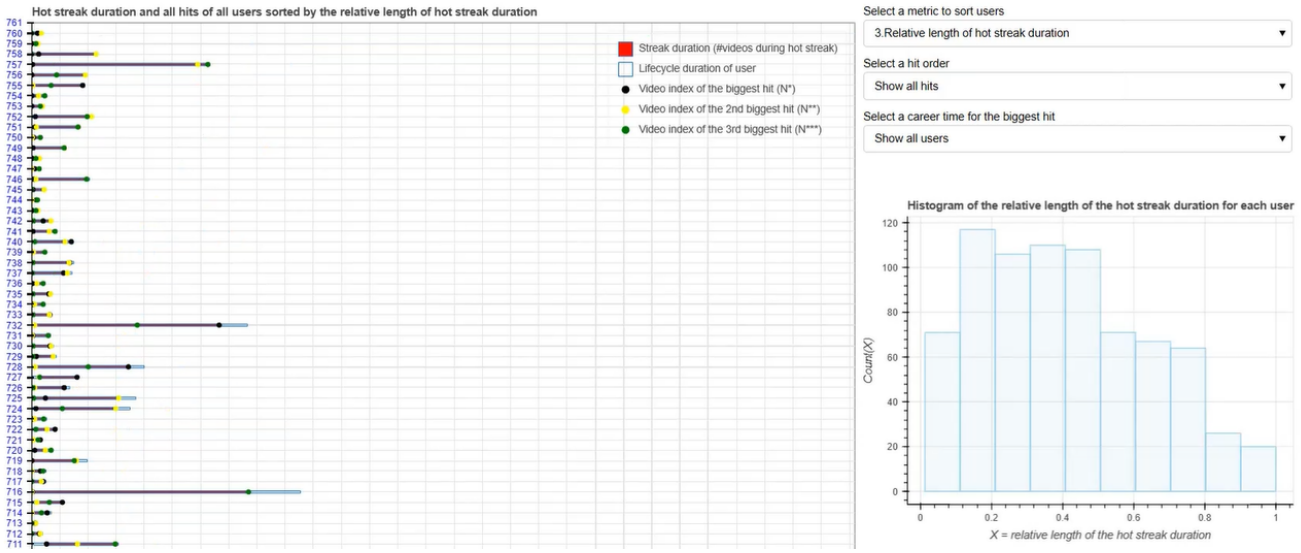



Figure 6.36: Sorting by the relative length of the hot streak duration and histogram of the relative length of the hot streak duration

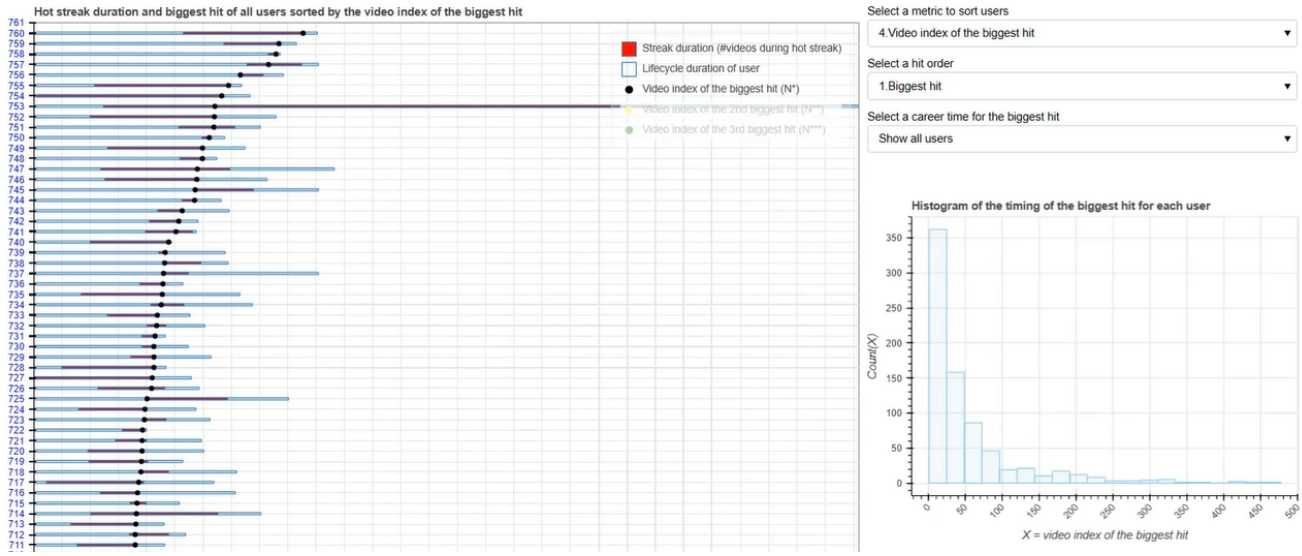



Figure 6.37: Sorting by the video index of the biggest hit and histogram of the video index of the biggest hit

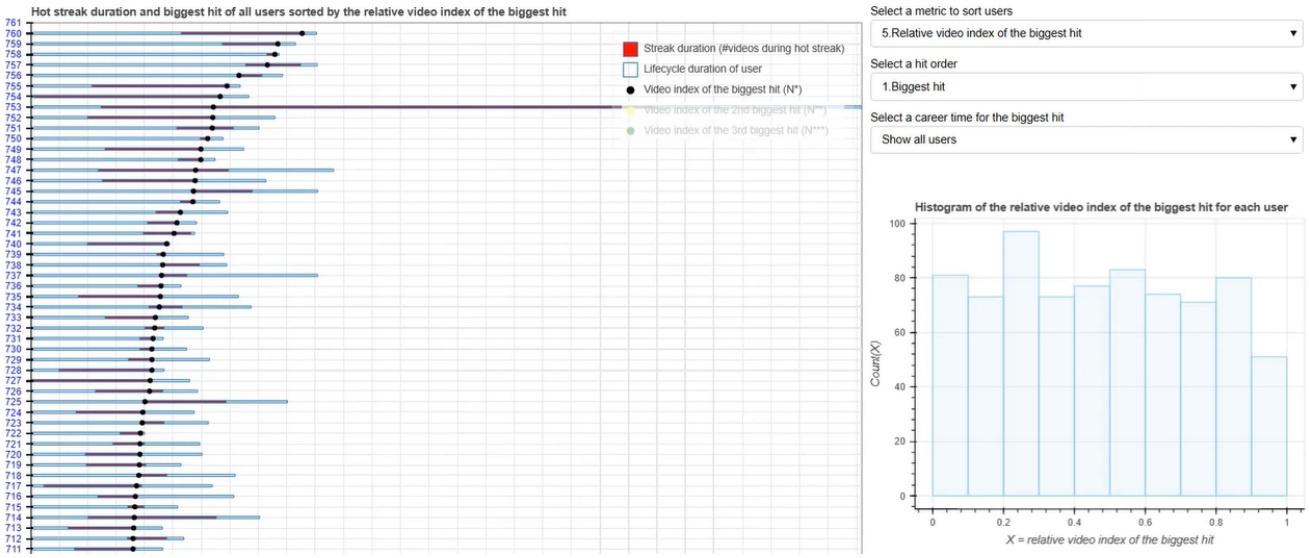
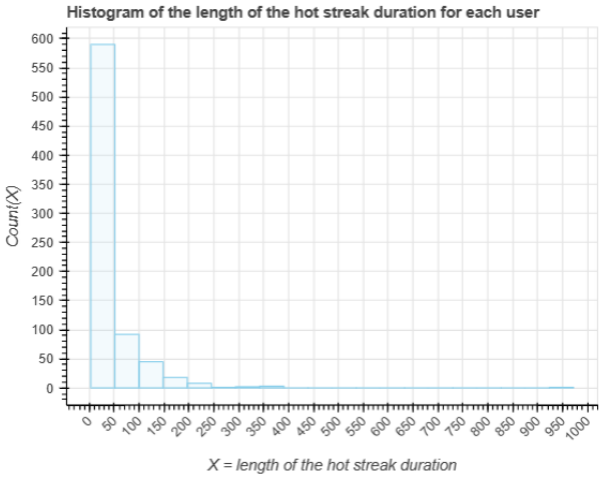
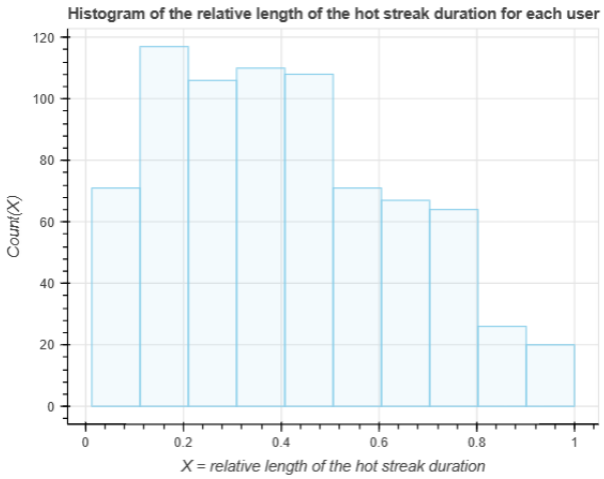



Figure 6.38: Sorting by the relative video index of the biggest hit and histogram of the relative video index of the biggest hit

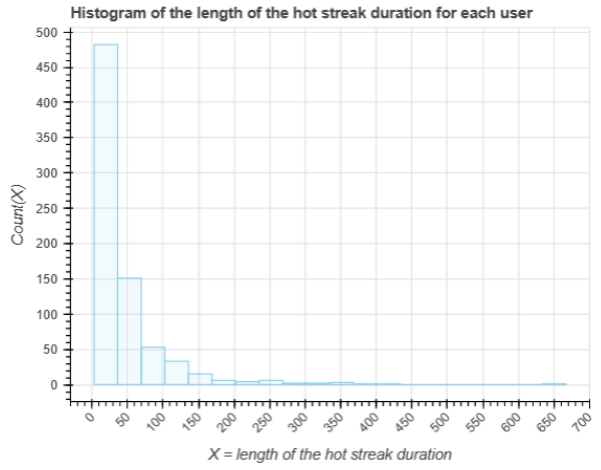



(a) Histogram distribution of the length of hot streak duration #likes

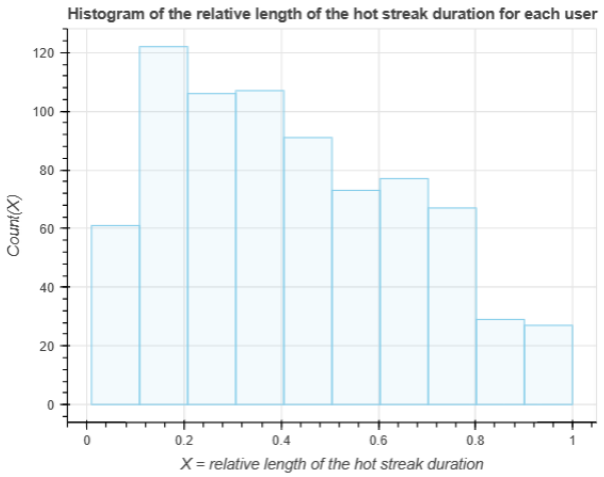



(b) Histogram distribution of the relative length of hot streak duration #likes

Figure 6.39: Histogram distributions of the (a) length of hot streak duration (b) relative length of hot streak duration of original data according to #likes

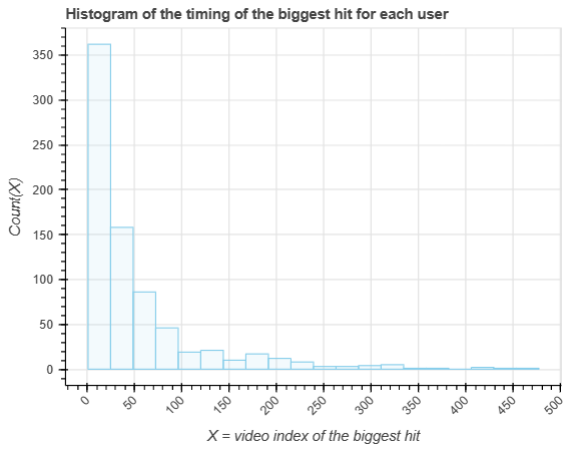



(a) Histogram distribution of the length of hot streak duration #plays

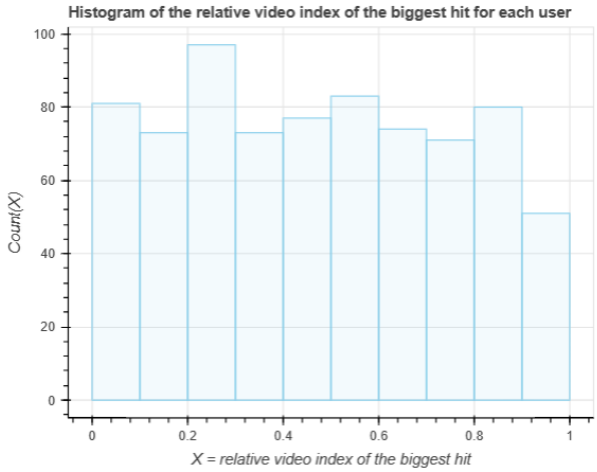



(b) Histogram distribution of the relative length of hot streak duration #plays

Figure 6.40: Histogram distributions of the (a) length of hot streak duration (b) relative length of hot streak duration of original data according to #plays

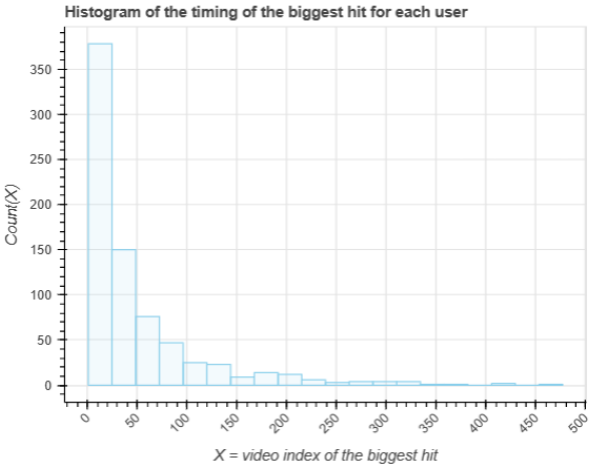



(a) Histogram distribution of the timing of the biggest hit #likes

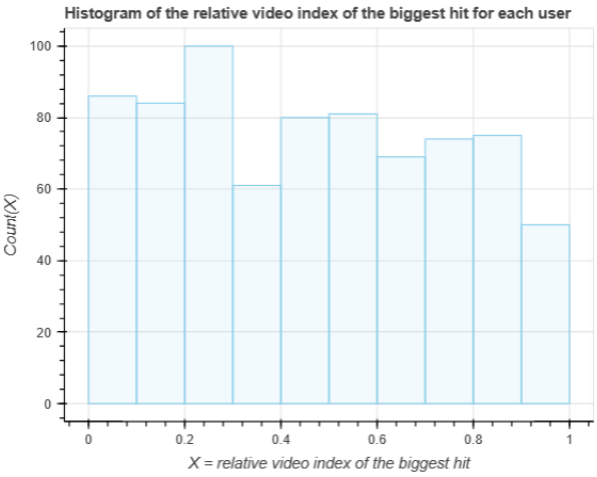



(b) Histogram distribution of the relative timing of the biggest hit #likes

Figure 6.41: Histogram distributions of the (a) timing of the biggest hit (b) relative timing of the biggest hit of original data according to #likes




(a) Histogram distribution of the timing of the biggest hit #plays




(b) Histogram distribution of the relative timing of the biggest hit #plays

Figure 6.42: Histogram distributions of the (a) timing of the biggest hit (b) relative timing of the biggest hit of original data according to #plays

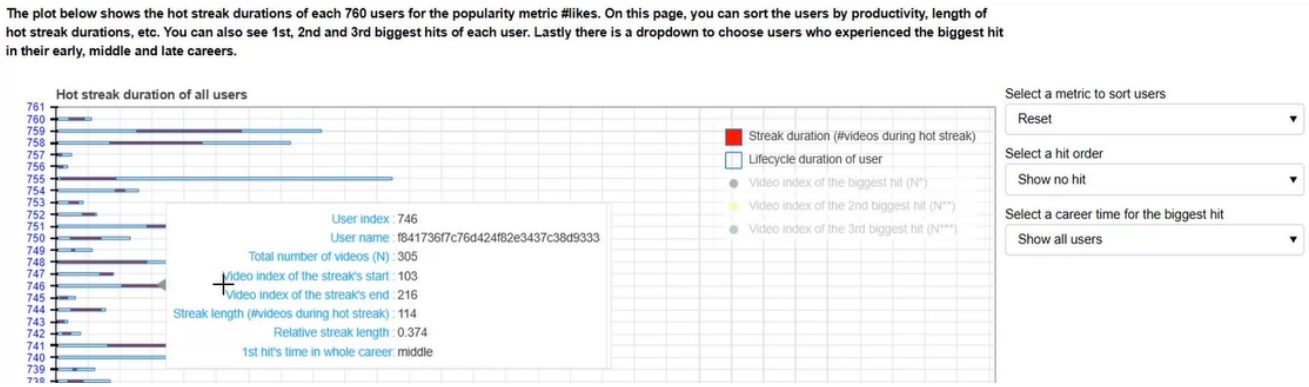



Figure 6.43: Tooltip shown when "Show no hit" is selected among the "hit order" options.

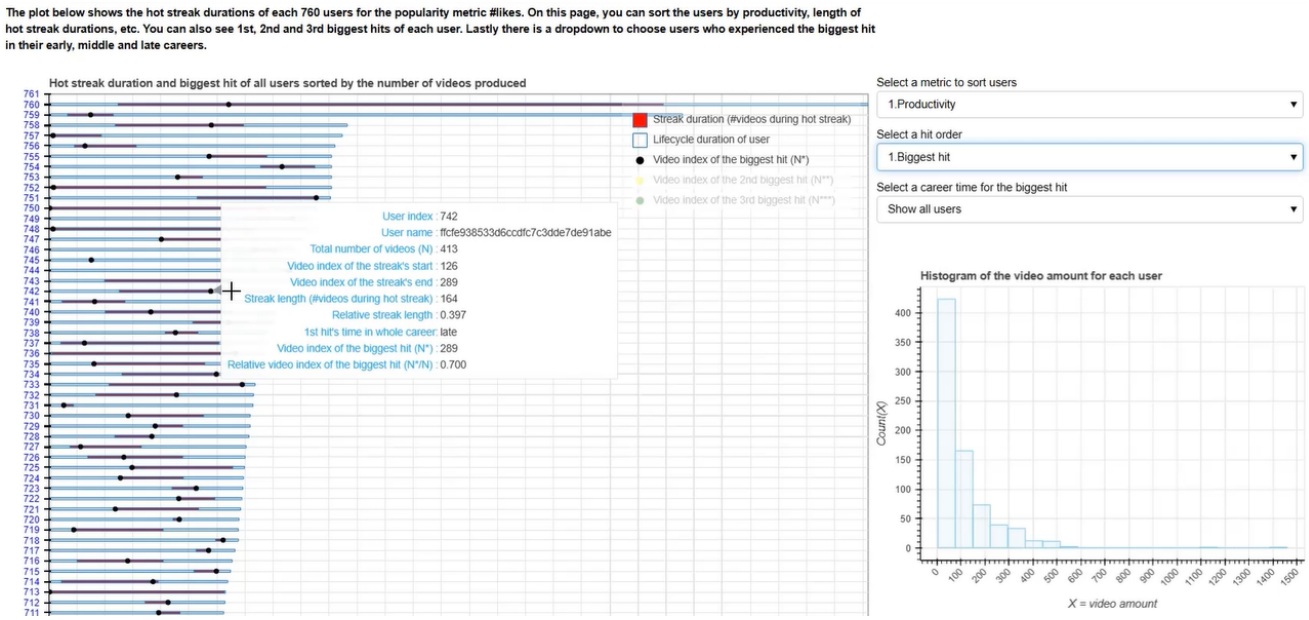



Figure 6.44: Tooltip shown when “Biggest hit” is selected among the “hit order” options. It includes the video index of the biggest hit, and relative video index of the biggest hit information in addition to the information shown on the tooltip when “Show no hit” is selected.

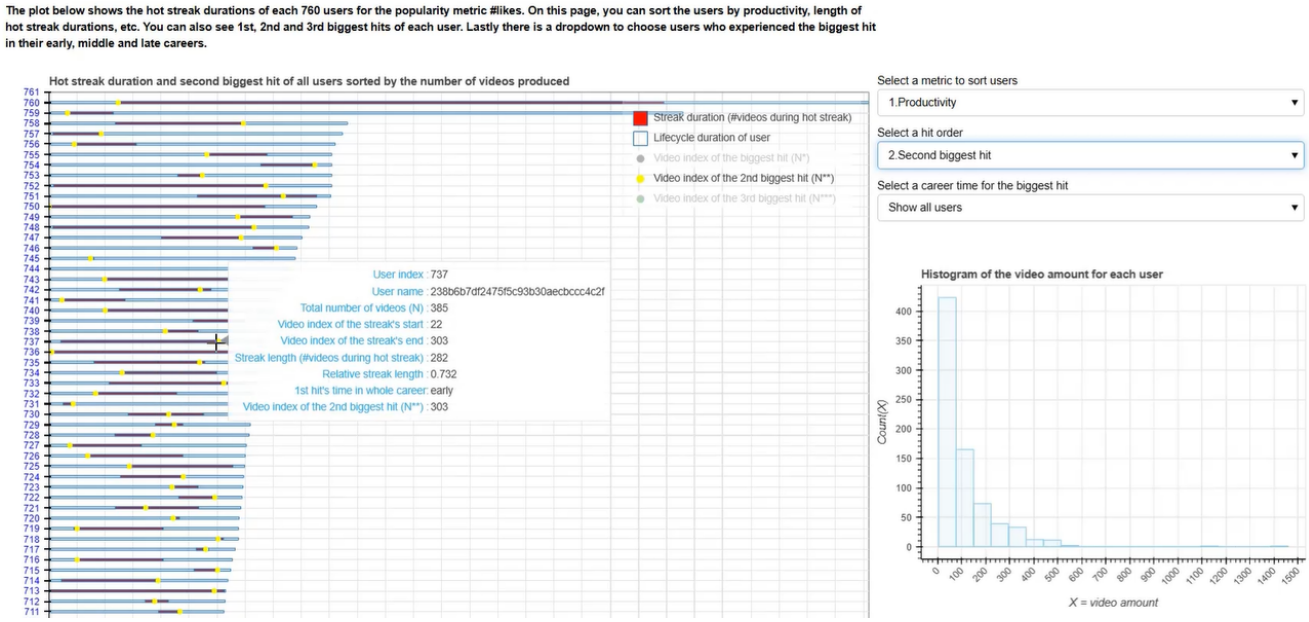



Figure 6.45: Tooltip shown when “Second biggest hit” is selected among the “hit order” options. It includes the video index of the second-biggest hit information in addition to the information shown on the tooltip when “Show no hit” is selected.

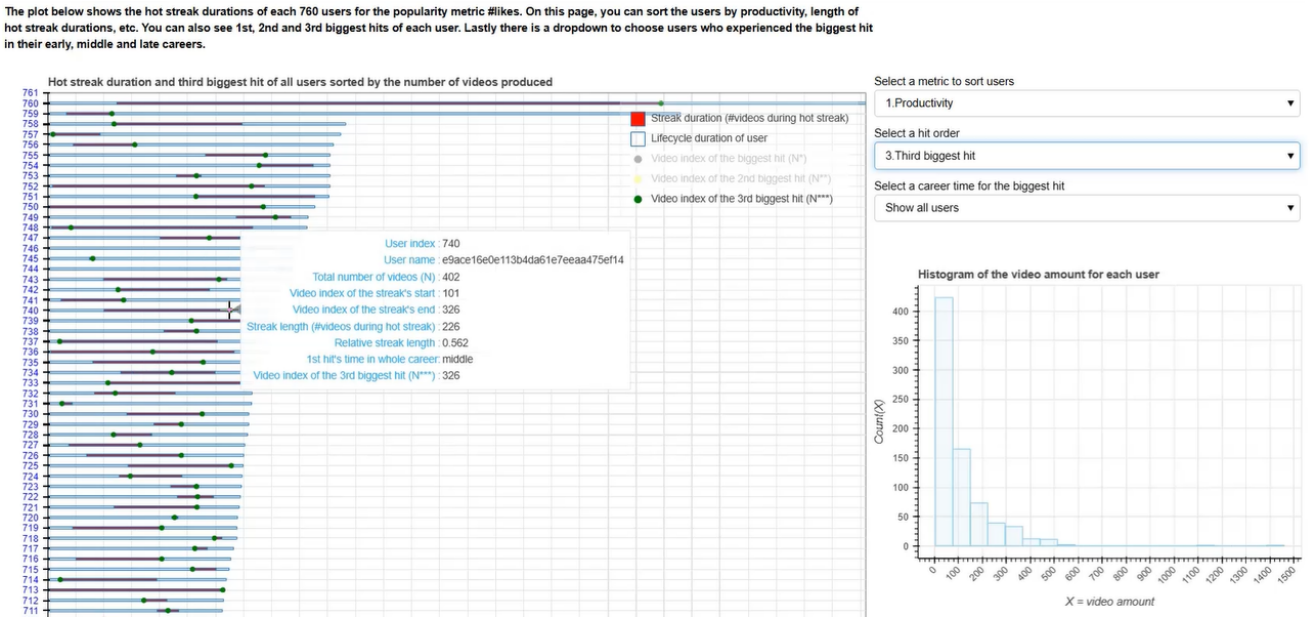



Figure 6.46: Tooltip shown when “Third biggest hit” is selected among the “hit order” options. It includes the video index of the third-biggest hit information in addition to the information shown on the tooltip when “Show no hit” is selected.

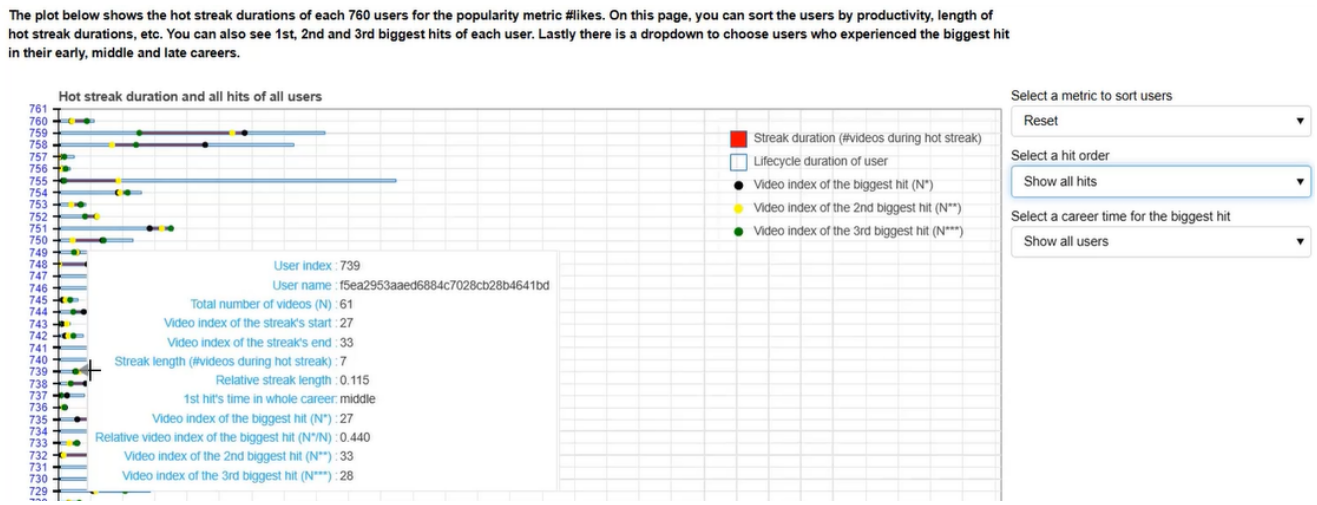



Figure 6.47: Tooltip shown when “Show all hits” is selected among the “hit order” options. It includes the video index of the first 3 biggest hits, and the relative video index of the biggest hit information in addition to the information shown on the tooltip when “Show no hit” is selected.

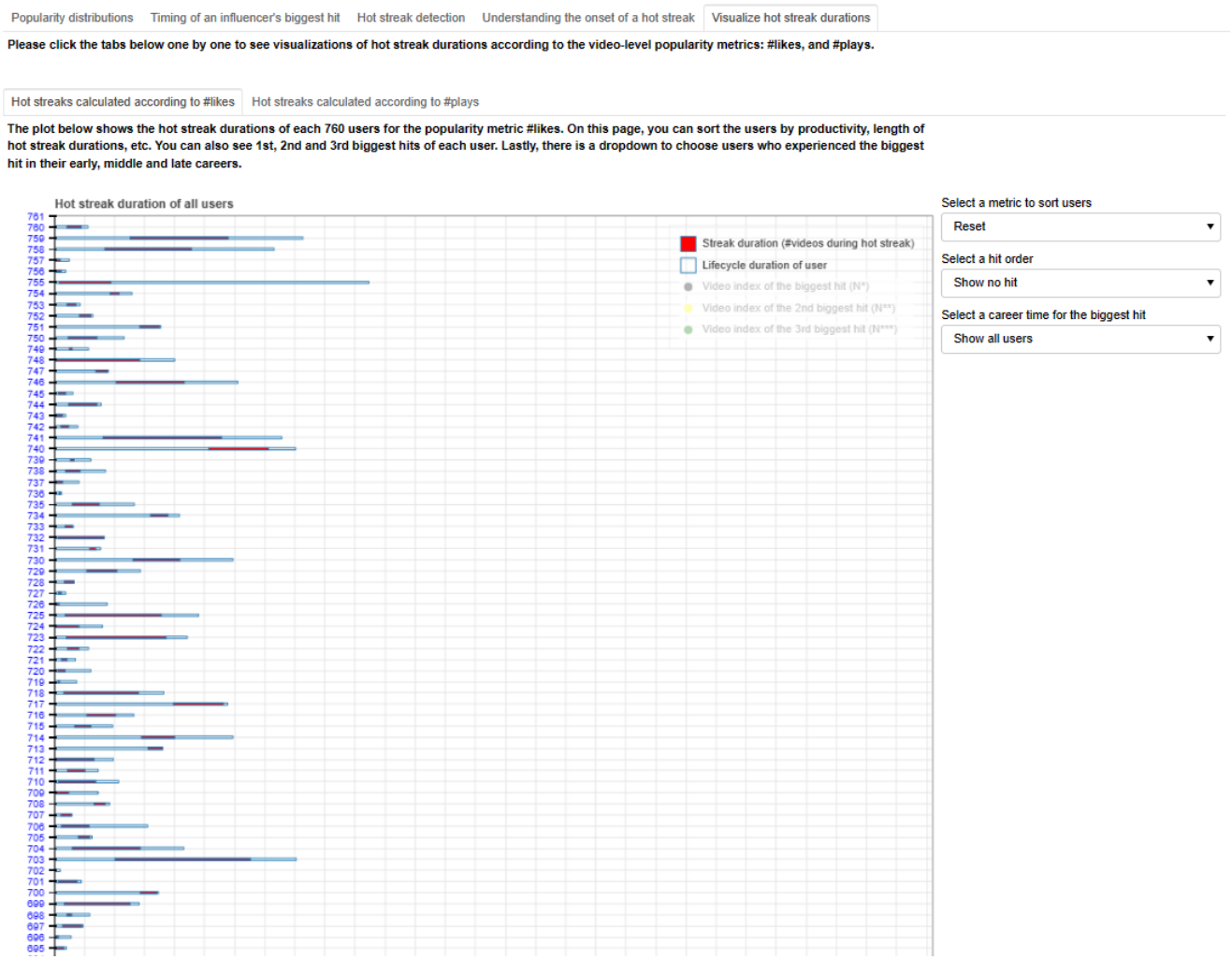



Figure 6.48: Visualize hot streak durations tab for #likes

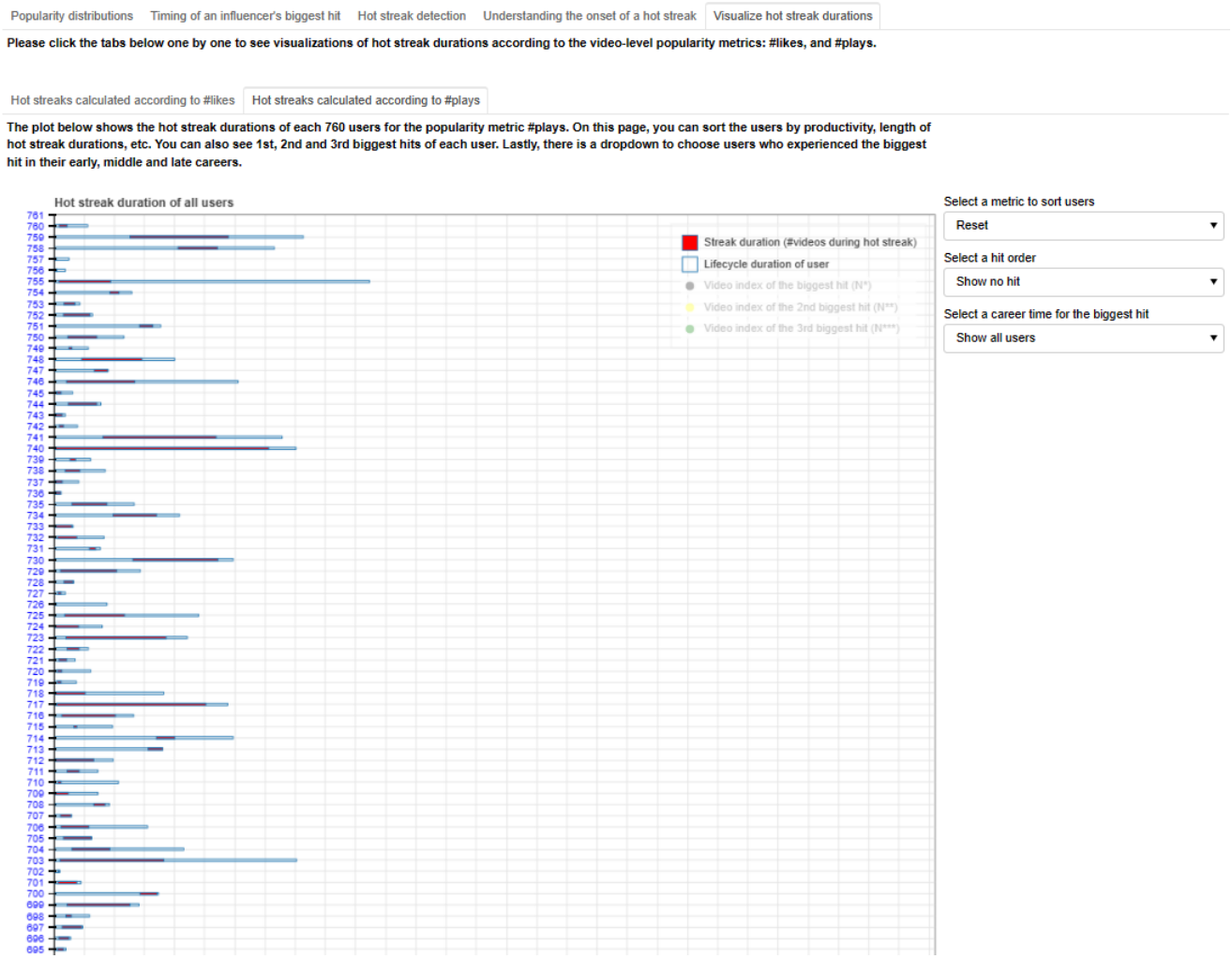



Figure 6.49: Visualize hot streak durations tab for #plays

# 7 Conclusion

**Data characterization** First, I identified that the number of videos was between 5 and 1460 for each author. The "total number of likes" per author has the widest range among other user-level popularity metrics, and the "total number of followers" has a range substantially higher than the "total number of followings", meaning that the authors in the dataset are the ones who preferred to follow less compared to their follower amounts, and they have received a great number of likes for their content.

Second, given the video-level popularity metrics' ranges, I discovered that if a video gets shared, it has a high possibility to be shared again. However, since the "total number of video plays", and the "total number of video likes" have the widest ranges among the other video-level popularity metrics from high to low, respectively, in my thesis, I decided to make conclusions by giving more credit to the patterns I've found regarding the data distributions according to the "number of plays", or "number of likes" metrics while analyzing the timing of the biggest hit, detecting hot streaks, understanding the onset of a hot streak, and examining hot streak durations.

**Timing of the biggest hit** Initially when the results are evaluated, it was observed that the biggest hit appears more in early careers, than in middle, or late careers. On the other hand, the randomized careers showed that the biggest hit can occur at any time through an influencer's lifecycle. However, as the probability distributions proved, there is not a significant difference for the timing of the biggest hit either to be in early, middle, or late careers both for the original and the shuffled dataset. In addition, the statistical test results between the original and shuffled datasets proved that the original and shuffled datasets could come from the same distribution. Therefore, one can conclude that the timing of the most popular video is random in TikTok users' lifecycles according to the dataset in my thesis study.

This conclusion also is parallel with the finding of Sinatra et al.'s study [21] who've discovered that the timing of the biggest hit in a scientist's career is random.

**Hot streak detection** As the original data and the average of 100 times shuffled data distributions are investigated for the video-level popularity metrics, it cannot be claimed when it is more likely for the second hit to occur: before or after the 1st biggest hit. However, one can assert that the timing of the first two biggest hits is close to each other for all video-level popularity metrics both in real and randomized careers.

When the original data and each of the 25 shuffled datasets' distributions are examined for the video-level popularity metrics, it cannot be identified if there is a pattern of whether the 2nd hit's timing is before or after the 1st hit's timing. But, it can be claimed that the timing of the first two hits is close to each other. Additionally, when the Rmax distribution is checked, one can assert that since there are peaks, the data distributions of original and shuffled data might be different from each other. However, the statistical test results between these distributions

show that the datasets come from the same distribution.

All in all, given the original dataset in my thesis, and the results summarized above, it can be affirmed that TikTok influencers experience average success close to their most popular videos. Similarly, in the study by Liu et al., they asserted that the biggest hit's timing is generally close to the second biggest hit's timing. In contrast to my results, they also observed in their dataset that the biggest hit can occur before or after the second biggest hit with almost equal probability [15].

**Understanding the onset of a hot streak** When the hashtag entropies of the original dataset and average of 100 times shuffled dataset were investigated according to the video-level popularity metrics, it was affirmed that TikTok influencers or content creators use different hashtags more during the hot streak phase than before it. As a result, TikTok influencers follow a different pattern than artists, film directors, and scientists who use diverse topics, or styles more before the hot streak period, as indicated in the study by Liu et al. [16].

**Visualizing hot streak durations** The results obtained from the plots of hot streak durations of all users show that there are typically fewer than 150 videos per author.

Moreover, there is another proof as denoted in table 6.3 that the timing of the biggest hit is random within an influencer's career lifecycle. This timing mainly occurs with an index of less than 50.

Additionally, hot streak lengths are generally below 40, while the relative hot streak length-success duration-of an influencer is usually between $\%12$ and $\%50$.

**Possible usage of the results in the industry** Companies and non-governmental organizations (NGOs) can use these observations in this thesis study to identify influencers who are in a hot streak period of their careers to reach them to promote their products or causes. For instance, if they notice that a TikTok influencer has had the biggest impact work recently, they can predict that this influencer may produce a second-biggest hit not in the distant future or has already experienced the second-biggest hit. They can also select authors who are in their early careers because this study's findings show that it is highly likely for influencers to experience the biggest hit in their early careers, and they generally have their biggest hit with a video index of less than 50.

Therefore, this research can help companies place their marketing campaigns by choosing the TikTok authors who are more likely to produce high-impact work or who are in their hot streak periods. Moreover, NGOs can increase the awareness of the causes they support by getting help from the TikTok influencers' followers' interactions with the content related to their causes.

# 8 Discussion and Future Work

**Discussion** While analyzing the topic, or style entropy in section 5.4, hashtags were used in my study. But, categories or challenge names can also be used to gauge the topic or style entropy. In this way, different results can be obtained regarding the tendency of exploration amount before the hot streak or during the hot streak periods. Then, the average of the three entropy values (hashtag, categories, challenge names), can be calculated to have a more estimated result for the exploration amount.

**Limitation on the dataset** The "share" button of videos on TikTok provides options such as copying the link or providing an embedded link for the video. Therefore, we don't know if the "number of shares" in the dataset corresponds to the shares as sending the link to friends or sharing on other social media platforms.

An additional video-level popularity metric could be "number of likes on comments". Meaning how many likes in total the comments of a video have could be another metric that should be analyzed to evaluate the popularity of a video.

Moreover, on the TikTok dataset, it is not known whether a video is a duet, stitch, or livestream. Therefore, video style's impact on the timing of the biggest hit video, or the start of a hot streak period couldn't be analyzed.

**Future Work** The methods utilized in this study and the results discovered can enlighten future research for other social media platforms that also include video content such as YouTube and Instagram. The same methods can be applied to the datasets obtained from these social media channels. In the end, the results can be evaluated to make predictions about the influencers' success periods, the start of their success periods, and the timings of their high-impact works.

# Abbreviations

| | |
|---|---|
| **AI** | artificial intelligence |
| **CCD** | complementary cumulative distribution |
| **CCDF** | complementary cumulative distribution function |
| **CD** | cumulative distribution |
| **CDF** | cumulative distribution function |
| **ETH** | Swiss Federal Institute of Technology Zurich |
| **PMD** | probability mass distribution |
| **PMF** | probability mass function |
| **NGOs** | non-governmental organizations |
| #**comments** | number of comments |
| #**author likes** | number of author likes |
| #**followers** | number of followers |
| #**followings** | number of followings |
| #**likes** | number of likes |
| #**shares** | number of shares |
| **SMIs** | social media influencers |
| #**plays** | number of plays |
| #**videos** | number of videos |
| **UZH** | University of Zurich |

# List of Figures

# List of Tables

# 9 Appendix

## 9.1 Data subset constructed in the "Data Characterization" phase

This section defines the columns in the data subset(data frame) that has been constructed in the phase described in the subsection "Constructing a subset of the data" 4.1.2.

### 9.1.1 User level and video level popularity metrics

**A. User level popularity metrics:**

**video_total (N):** Total number of videos

**author_num_likes:** Total number of likes

**author_num_followers:** Total number of followers

**author_num_followings:** Total number of followings

**B. Video level popularity metrics:**

**1. Metrics for "number of likes" videos have received:**

**video_order_1st_likes (N*):** Index of the video that received the most likes in an author's career lifecycle

**video_order_2nd_likes (N**):** Index of the video that received the 2nd most likes in an author's career lifecycle

**video_order_3rd_likes (N***):** Index of the video that received the 3rd most likes in an author's career lifecycle

**max_likes:** maximum number of likes that a video received in an author's career lifecycle

**max_2nd_likes:** 2nd maximum "number of likes" that a video received in an author's career lifecycle

**max_3rd_likes:** 3rd maximum "number of likes" that a video received in an author's career lifecycle

**ratio_max_likes (N*/N):** relative index of the maximum "number of likes" that a video received in an author's career lifecycle

**ratio_max_2nd_likes (N**/N):** relative index of the 2nd maximum "number of likes" that a video received in an author's career lifecycle

**ratio_max_3rd_likes (N***/N):** relative index of the 3rd maximum "number of likes" that a video received in an author's career lifecycle

**ratio_delta_likes ((N*-N**)/N):** relative distance between the indexes of the videos that received the maximum "number of likes" and that received the 2nd maximum number of likes in an author's career lifecycle

**2. Metrics for "number of shares" videos have received:**

**video_order_1st_shares:** Index of the video that received the most shares in an author's career lifecycle

**video_order_2nd_shares:** Index of the video that received the 2nd most shares in an author's career lifecycle

**video_order_3rd_shares:** Index of the video that received the 3rd most shares in an author's career lifecycle

**max_shares:** maximum number of shares that a video received in an author's career lifecycle

**max_2nd_shares:** 2nd maximum "number of shares" that a video received in an author's career lifecycle

**max_3rd_shares:** 3rd maximum "number of shares" that a video received in an author's career lifecycle

**ratio_max_shares:** relative index of the maximum "number of shares" that a video received in an author's career lifecycle

**ratio_max_2nd_shares:** relative index of the 2nd maximum "number of shares" that a video received in an author's career lifecycle

**ratio_max_3rd_shares:** relative index of the 3rd maximum "number of shares" that a video received in an author's career lifecycle

**ratio_delta_shares:** relative distance between the indexes of the videos that received the maximum "number of shares" and that received the 2nd maximum number of shares in an author's career lifecycle

**3. Metrics for "number of plays" videos have received:**

**video_order_1st_plays:** Index of the video that received the most plays in an author's career lifecycle

**video_order_2nd_plays:** Index of the video that received the 2nd most plays in an author's career lifecycle

**video_order_3rd_plays:** Index of the video that received the 3rd most plays in an author's career lifecycle

**max_plays:** maximum number of plays that a video received in an author's career lifecycle

**max_2nd_plays:** 2nd maximum "number of plays" that a video received in an author's career lifecycle

**max_3rd_plays:** 3rd maximum "number of plays" that a video received in an author's career lifecycle

**ratio_max_plays:** relative index of the maximum "number of plays" that a video received in an author's career lifecycle

**ratio_max_2nd_plays:** relative index of the 2nd maximum "number of plays" that a video received in an author's career lifecycle

**ratio_max_3rd_plays:** relative index of the 3rd maximum "number of plays" that a video received in an author's career lifecycle

**ratio_delta_plays:** relative distance between the indexes of the videos that received the maximum "number of plays" and that received the 2nd maximum number of plays in an author's career lifecycle

#### 4. Metrics for "number of comments" videos have received:

**video_order_1st_comments:** Index of the video that received the most comments in an author's career lifecycle

**video_order_2nd_comments:** Index of the video that received the 2nd most comments in an author's career lifecycle

**video_order_3rd_comments:** Index of the video that received the 3rd most comments in an author's career lifecycle

**max_comments:** maximum number of comments that a video received in an author's career lifecycle

**max_2nd_comments:** 2nd maximum "number of comments" that a video received in an author's career lifecycle

**max_3rd_comments:** 3rd maximum "number of comments" that a video received in an author's career lifecycle

**ratio_max_comments:** relative index of the maximum "number of comments" that a video received in an author's career lifecycle

**ratio_max_2nd_comments:** relative index of the 2nd maximum "number of comments" that a video received in an author's career lifecycle

**ratio_max_3rd_comments:** relative index of the 3rd maximum "number of comments" that a video received in an author's career lifecycle

**ratio_delta_comments:** relative distance between the indexes of the videos that received the maximum "number of comments" and that received the 2nd maximum number of comments in an author's career lifecycle

## 9.2 CCDF plots for log-log axis scale constructed in the “Data Characterization” phase

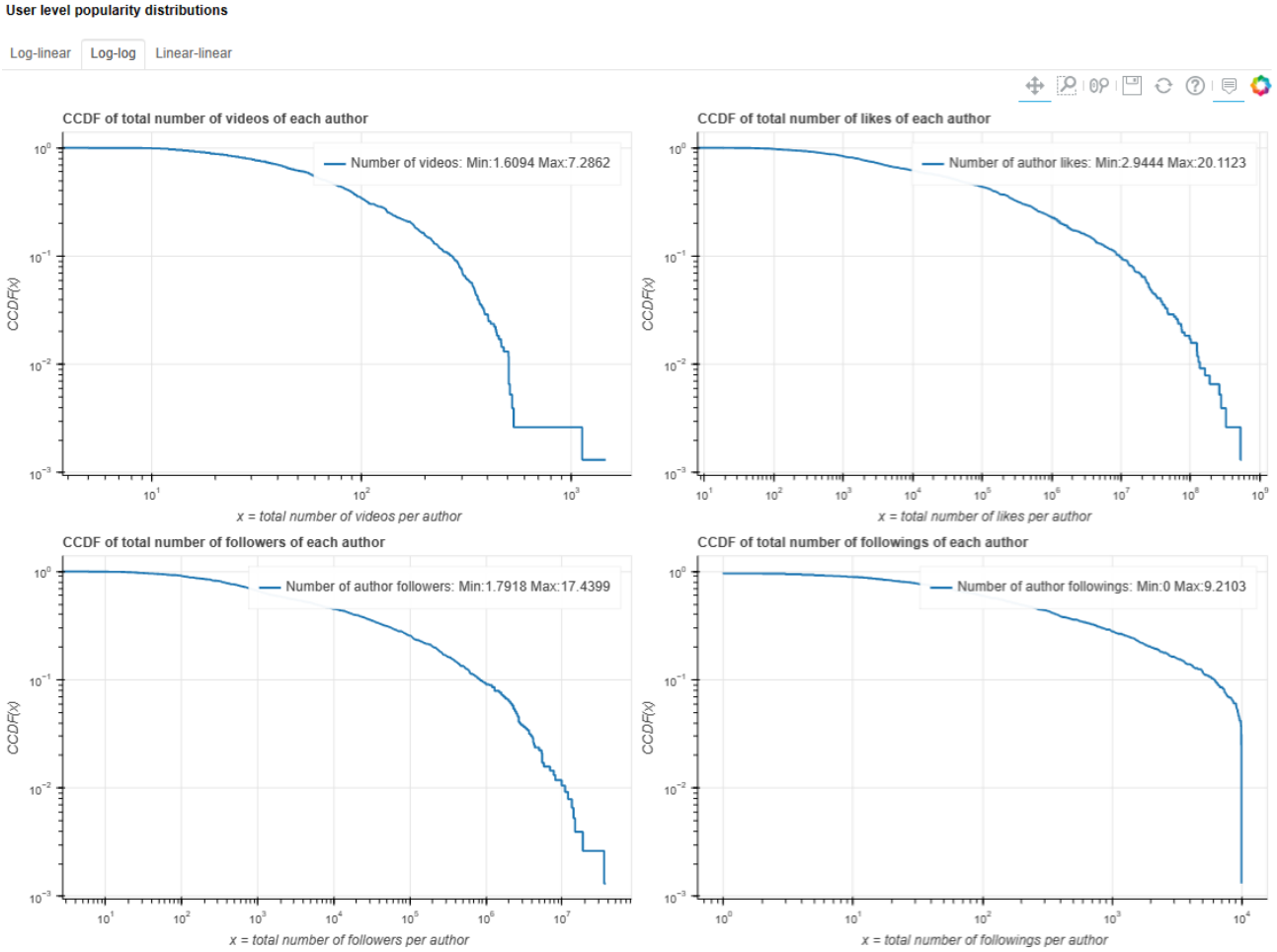



Figure 9.1: CCDF of user-level popularity metrics in log-log scale

CCDF plots for user-level popularity metrics in log scale on both axes. The metrics are the total number of videos per author, the total number of likes per author, the total number of followers per author, and the total number of followings per author, from left to right, respectively.

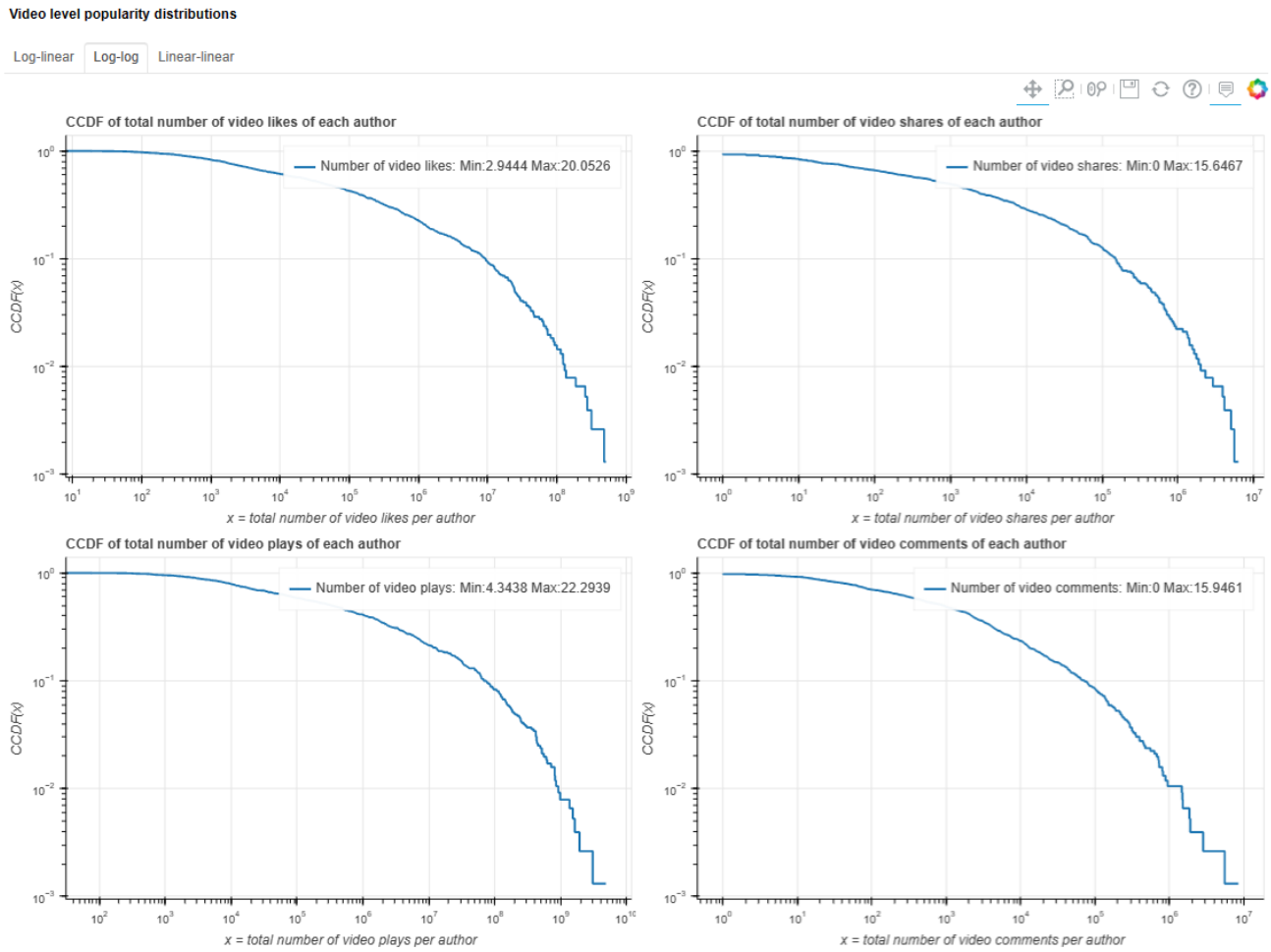



Figure 9.2: CCDF of video-level popularity metrics in log-log scale

CCDF plots for video level popularity metrics in log scale on both axes. The metrics are the total number of video likes per author, the total number of video shares per author, the total number of video plays per author, the total number of video comments per author, from left to right, respectively.

## 9.3 Data subset constructed in the "Timing of an influencer's hit" phase

This section defines the columns in the data subset(dataframe) that has been constructed in the phase described in the subsection "Data"4.2.3 of chapter 4's section-"Timing of an influencer's hit" section4.2 from the average of 100 times shuffled dataset and from the 25 separate shuffled datasets.

### 9.3.1 User level and video level popularity metrics

#### A. User level popularity metrics:

**video_total:** Total number of videos

**author_num_likes:** Total number of likes

**author_num_followers:** Total number of followers

**author_num_followings:** Total number of followings

#### B. Video level popularity metrics:

#### 1. Metrics for "number of likes" videos have received:

**sh_video_order_1st_likes:** Index of the video that received the most likes in an author's career lifecycle

**sh_video_order_2nd_likes:** Index of the video that received the 2nd most likes in an author's career lifecycle

**sh_video_order_3rd_likes:** Index of the video that received the 3rd most likes in an author's career lifecycle

**sh_max_likes:** maximum number of likes that a video received in an author's career lifecycle

**sh_max_2nd_likes:** 2nd maximum "number of likes" that a video received in an author's career lifecycle

**sh_max_3rd_likes:** 3rd maximum "number of likes" that a video received in an author's career lifecycle

**sh_ratio_max_likes:** relative index of the maximum "number of likes" that a video received in an author's career lifecycle

**sh_ratio_max_2nd_likes:** relative index of the 2nd maximum "number of likes" that a video received in an author's career lifecycle

**sh_ratio_max_3rd_likes:** relative index of the 3rd maximum "number of likes" that a video received in an author's career lifecycle

**sh_ratio_delta_likes:** relative index of the distance between the video that received the maximum "number of likes" and that received the 2nd maximum number of likes in an author's career lifecycle

#### 2. Metrics for "number of shares" videos have received:

**sh_video_order_1st_shares:** Index of the video that received the most shares in an author's career lifecycle

**sh_video_order_2nd_shares:** Index of the video that received the 2nd most shares in an author's career lifecycle

**sh_video_order_3rd_shares:** Index of the video that received the 3rd most shares in an author's career lifecycle

**sh_max_shares:** maximum number of shares that a video received in an author's career lifecycle

**sh_max_2nd_shares:** 2nd maximum "number of shares" that a video received in an author's career lifecycle

**sh_max_3rd_shares:** 3rd maximum "number of shares" that a video received in an author's career lifecycle

**sh_ratio_max_shares:** relative index of the maximum "number of shares" that a video received in an author's career lifecycle

**sh_ratio_max_2nd_shares:** relative index of the 2nd maximum "number of shares" that a video received in an author's career lifecycle

**sh_ratio_max_3rd_shares:** relative index of the 3rd maximum "number of shares" that a video received in an author's career lifecycle

**sh_ratio_delta_shares:** relative index of the distance between the video that received the maximum "number of shares" and that received the 2nd maximum number of shares in an author's career lifecycle

**3. Metrics for "number of plays" videos have received:**

**sh_video_order_1st_plays:** Index of the video that received the most plays in an author's career lifecycle

**sh_video_order_2nd_plays:** Index of the video that received the 2nd most plays in an author's career lifecycle

**sh_video_order_3rd_plays:** Index of the video that received the 3rd most plays in an author's career lifecycle

**sh_max_plays:** maximum number of plays that a video received in an author's career lifecycle

**sh_max_2nd_plays:** 2nd maximum "number of plays" that a video received in an author's career lifecycle

**sh_max_3rd_plays:** 3rd maximum "number of plays" that a video received in an author's career lifecycle

**sh_ratio_max_plays:** relative index of the maximum "number of plays" that a video received in an author's career lifecycle

**sh_ratio_max_2nd_plays:** relative index of the 2nd maximum "number of plays" that a video received in an author's career lifecycle

**sh_ratio_max_3rd_plays:** relative index of the 3rd maximum "number of plays" that a video received in an author's career lifecycle

**sh_ratio_delta_plays:** relative index of the distance between the video that received the maximum "number of plays" and that received the 2nd maximum number of plays in an author's career lifecycle

#### 4. Metrics for "number of comments" videos have received:

**sh_video_order_1st_comments:** Index of the video that received the most comments in an author's career lifecycle

**sh_video_order_2nd_comments:** Index of the video that received the 2nd most comments in an author's career lifecycle

**sh_video_order_3rd_comments:** Index of the video that received the 3rd most comments in an author's career lifecycle

**sh_max_comments:** maximum number of comments that a video received in an author's career lifecycle

**sh_max_2nd_comments:** 2nd maximum "number of comments" that a video received in an author's career lifecycle

**sh_max_3rd_comments:** 3rd maximum "number of comments" that a video received in an author's career lifecycle

**sh_ratio_max_comments:** relative index of the maximum "number of comments" that a video received in an author's career lifecycle

**sh_ratio_max_2nd_comments:** relative index of the 2nd maximum "number of comments" that a video received in an author's career lifecycle

**sh_ratio_max_3rd_comments:** relative index of the 3rd maximum "number of comments" that a video received in an author's career lifecycle

**sh_ratio_delta_comments:** relative index of the distance between the video that received the maximum "number of comments" and that received the 2nd maximum number of comments in an author's career lifecycle

## 9.4 Metrics defined and calculated in the "Understanding the onset of a hot streak" phase

#### A. User level metrics:

**unique_hashtag_count:** (m): Number of unique hashtags of a user

**video_total:** (n): Total number of videos of a user

**hashtag_freq:** (pi): Hashtag frequency of a hashtag of a user

#### B. Video level metrics:

#### 1. Metrics associated with "number of likes":

**Nonset_likes:** Minimum of the video index of the first 3 biggest hits. First video index of the hot streak period.

**Nend_likes:** Maximum of the video index of the first 3 biggest hits. Last video index of the hot streak period.

**Hbefore_likes:** (Hb_likes): Normalized hashtag entropy of a user before the hot streak period starts

**Hduring_likes:** (Hd_likes): Normalized hashtag entropy of a user during the hot streak period

#### 2. Metrics associated with "number of shares":

**Nonset_shares:** Minimum of the video index of the first 3 biggest hits. First video index of the hot streak period.

**Nend_shares:** Maximum of the video index of the first 3 biggest hits. Last video index of the hot streak period.

**Hbefore_shares:** (Hb_shares): Normalized hashtag entropy of a user before the hot streak period starts

**Hduring_shares:** (Hd_shares): Normalized hashtag entropy of a user during the hot streak period, associated with the video level popularity metric -number of shares

#### 3. Metrics associated with "number of plays":

**Nonset_plays:** Minimum of the video index of the first 3 biggest hits. First video index of the hot streak period.

**Nend_plays:** Maximum of the video index of the first 3 biggest hits. Last video index of the hot streak period.

**Hbefore_plays:** (Hb_plays): Normalized hashtag entropy of a user before the hot streak period starts, associated with the video level popularity metric -number of plays

**Hduring_plays:** (Hd_plays): Normalized hashtag entropy of a user during the hot streak period, associated with the video level popularity metric -number of plays

#### 4. Metrics associated with "number of comments":

**Nonset_comments:** Minimum of the video index of the first 3 biggest hits. First video index of the hot streak period.

**Nend_comments:** Maximum of the video index of the first 3 biggest hits. Last video index of the hot streak period.

**Hbefore_comments:** (Hb_comments): Normalized hashtag entropy of a user before the hot streak period starts, associated with the video level popularity metric -number of comments

**Hduring_comments:** (Hd_comments): Normalized hashtag entropy of a user during the hot streak period, associated with the video level popularity metric -number of comments

# Bibliography


[1] Katie Elson Anderson. "Getting acquainted with social networks and apps: it is time to talk about TikTok". In: *Library Hi Tech News* 37.4 (Feb. 2020), pp. 7–12. DOI: https://doi.org/10.1108/LHTN-01-2020-0001.

[2] Gil Appel et al. "The future of social media in marketing". In: *Journal of the Academy of Marketing Science* 48 (Oct. 2019), pp. 79–95. DOI: 10.1007/s11747-019-00695-1.

[3] Sergio Barta et al. "Influencer marketing on TikTok: The effectiveness of humor and followers' hedonic experience". In: *Journal of Retailing and Consumer Services* 70 (2023), p. 103149. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2022.103149. URL: https://www.sciencedirect.com/science/article/pii/S0969698922002429.

[4] *Bokeh documentation*. https://docs.bokeh.org/en/latest/. Accessed: 2023-01-01.

[5] Laura Cervi. "Tik Tok and generation Z". In: *Theatre, Dance and Performance Training* 12.2 (July 2021), pp. 198–204. DOI: 10.1080/19443927.2021.1915617. URL: https://doi.org/10.1080/19443927.2021.1915617.

[6] Danica Facca et al. *Academic TikTok Report*. Tech. rep. Knowledge Media Design Institute, Faculty of Information, University of Toronto, Aug. 2022, pp. 1–41. URL: https://hdl.handle.net/1807/124170.

[7] Kendra Fowler and Veronica L. Thomas. "Influencer marketing: a scoping review and a look ahead". In: *Journal of Marketing Management* (2023), pp. 1–32. DOI: 10.1080/0267257X.2022.2157038. URL: https://doi.org/10.1080/0267257X.2022.2157038.

[8] Kiran Garimella and Robert West. "Hot Streaks on Social Media". In: *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)* (2019), pp. 170–180. DOI: 10.48550/ARXIV.1904.03301. URL: https://arxiv.org/abs/1904.03301.

[9] Jacob Goldenberg et al. "The Research Behind Influencer Marketing". In: *Journal of Marketing Research-Impact at JMR* (Feb. 2021). URL: https://www.ama.org/2022/02/16/the-research-behind-influencer-marketing/.

[10] Michael Haenlein et al. "Navigating the New Era of Influencer Marketing: How to be Successful on Instagram, TikTok, & Co." In: *California Management Review* 63 (2020), pp. 5–25. DOI: 10.1177/0008125620958166.

[11] Carly Hill. *What is TikTok: The complete platform guide for 2023*. Oct. 2022. URL: https://sproutsocial.com/insights/what-is-tiktok/.

[12] Milan Janosov, Federico Battiston, and Roberta Sinatra. "Success and luck in creative careers". In: *EPJ Data Science* 9 (Dec. 2020), pp. 9–20. DOI: 10.1140/epjds/s13688-020-00227-w.

[13] Rafi Letzter. *A teenager on TikTok disrupted thousands of scientific studies with a single video*. Sept. 2021. URL: https://www.theverge.com/2021/9/24/22688278/tiktok-science-study-survey-prolific.

[14] Fine Leung et al. "Influencer Marketing Effectiveness". In: *Journal of Marketing* 86.6 (2022), pp. 93–115. DOI: 10.1177/00222429221102889.

[15] Liu Lu et al. "Hot streaks in artistic, cultural, and scientific careers". In: *Nature* 559.7714 (July 2018), pp. 396–399. URL: https://doi.org/10.1038%5C%2Fs41586-018-0315-8.

[16] Liu Lu et al. "Understanding the onset of hot streaks across artistic, cultural, and scientific careers". In: *Nature Communications* 12.1 (Sept. 2021). URL: https://doi.org/10.1038%2Fs41467-021-25477-8.

[17] Newman MEJ. "Power laws, Pareto distributions and Zipf's law". In: *Contemporary Physics* 46.5 (Sept. 2005), pp. 323–351. DOI: 10.1080/00107510500052444. URL: https://doi.org/10.1080%2F00107510500052444.

[18] *OpenRefine user manual*. https://openrefine.org/docs. Accessed: 2022-09-15.

[19] Bandinee Pradhan, Kaushal Kishore, and Nilesh Gokhale. "Social Media Influencers and Consumer Engagement: A Review and Future Research Agenda". In: *International Journal of Consumer Studies* (Feb. 2023), pp. 1–25. DOI: 10.1111/ijcs.12901.

[20] *Python documentation*. https://www.python.org/doc/. Accessed: 2022-09-15.

[21] Sinatra Roberta et al. "Quantifying the evolution of individual scientific impact". In: *Science* 354.6312 (2016), pp. 596–604. URL: https://www.science.org/doi/abs/10.1126/science.aaf5239.

[22] Oliver Williams, Lucas Lacasa, and Vito Latora. "Quantifying and predicting success in show business". In: *Nature Communications* 10 (June 2019), pp. 2256–2263. DOI: 10.1038/s41467-019-10213-0.