LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Master's Thesis
ENTN19

Performance distribution of TikTok contentpreneurs: A conceptual replication study

Luiza Kulchetscki
Supervisor: Solomon Akele Abebe
Date of final Seminar: May 29th, 2024

Abstract

This study analyzes the performance distribution of TikTok contentpreneurs, along with its generative mechanism. It is a conceptual replication of Gala et al's (2024) "Star entrepreneurs on digital platforms: Heavy-tailed performance distributions and their generative mechanisms." It focuses on contentpreneurship as a form of entrepreneurship that takes place among content creators on social media platforms. The study tests Gala et al's hypotheses concerning the non-normality of entrepreneurial performance distributions on digital platforms, engendered by the prevalence of star entrepreneurs (or, contentpreneurs) who become outliers by outperforming the rest, skewing the platform's distribution into a non-normal shape and creating a heavy-tail effect. Using randomly selected data of 1,046 user accounts from TikTok, elements of distribution shape and fit are analyzed through statistical tests of correlation and goodness-of-fit tests. Longitudinal data from the same sample is used to identify the distribution's generative mechanism. Comparing the results of this study with its original, some support for the original hypotheses is found, mainly for non-normal shape and lognormal fit, but no support is found for knowledge intensity as a performance predictor, and only conditional support is found for proportional differentiation as a generative mechanism. Interpretation of the findings connect the discrepancy in results between the two papers to the distinct nature, evolution, and audience of the digital platforms investigated. This study contributes theory and empirical evidence for non-normal distributions to be embraced in the field of entrepreneurship, and provides theoretical and practical insights based on the results of the analysis.

Key terminology: contentpreneurship, social media platforms, non-normal distribution, lognormal, generative mechanism, proportional differentiation, preferential attachment, Pareto effect.

Acknowledgements

I would like to express my gratitude to Solomon Akele Abebe, whose guidance as my supervisor has been invaluable throughout the development of this thesis. His insightful feedback and unwavering support have greatly enriched the quality and direction of my work. I am also thankful to Hassan Hamadi, whose assistance with both the coding for the analysis as well as ideas for how to approach the research significantly contributed to the rigor and comprehensiveness of this study. His willingness to share his knowledge was instrumental in overcoming challenges during the data collecting process. I extend my appreciation to my peers from the Seminar group for their constructive feedback and constant encouragement. Furthermore, I am indebted to the authors of the articles that served as inspiration and motivation for this paper: Kaushik Gala, Andreas Schwab, and Brandon A. Mueller, authors of the paper I replicate in my thesis, and Harry Joo, Kyle J. Bradley, and Herman Aguinis, authors of the paper that inspired Gala, Schwab, and Mueller's. Their research provided the foundation upon which this study was built, and I am grateful for their important contributions to the field. A special thank you to Professor Aguinis in particular, who responded to my email inquiries and pointed me to useful resources for the distribution fitting method. Lastly, I would like to thank all individuals who have supported me in various capacities throughout this endeavor. Their encouragement and support have helped me achieve the successful completion of this project.

Table of Contents

List of Tables

List of Figures

1. Introduction

Meet the newest entrepreneur on the block: the contentpreneur. Equipped with a smartphone, laptop, and internet connection, contentpreneurs are content creating entrepreneurs who exploit ways to monetize their content on social media platforms (Johnson et al, 2022). Despite its economic relevance and its rise in popularity within the past few years (Frenkel, 2021) contentpreneurship is just barely finding its footing within academia (Johnson et al, 2022). An evolving phenomenon (Ibrar et al, 2022; Kullolli & Trebicka, 2023), digital content creation as a means of income lures many hopefuls looking for their big break on the internet (Gustafsson & Khan, 2017). This creates an intensely competitive environment with highly unequal outcomes in which relatively few succeed (Ashman et al, 2018; Srinavasan & Venkatraman, 2017). The gap between those who fall short of achieving specific milestones with their content creation and those who go on to reach social media celebrity status (Chen et al, 2020; Short & Short, 2023) is a reality scholars are only beginning to explore.

This contentpreneurial success gap is captured in a recent study published in 2024 by Gala, Schwab, and Mueller called "Star Entrepreneurs on digital platforms: Heavy-tailed performance distributions and their generative mechanisms." The study analyzes the performance distribution of lecturers (identified in the paper as digital entrepreneurs) who sell their courses on an instructional digital platform called Udemy. The study observes how a relatively small number of lecturers on Udemy consistently outperform the rest with an exponentially higher yearly income and number of enrolled students than the average lecturer on the platform (pg. 2). These high performing lecturers are the very star entrepreneurs mentioned in the title of the study, and represent the activity taking place in the tail end of Udemy's graphed distribution (pg. 2; 8). Compelled by their findings, this master's thesis is a conceptual replication of Gala et al's (2024) paper, and is meant to verify whether the same arguments and assumptions of the original study hold true for entrepreneurs of a different digital platform: TikTok.

TikTok is a dynamic social media platform that evolved into a valuable business and marketing tool from the entertainment platform it was primarily designed for (Cutolo & Grimaldi, 2023; Escolano, 2023). As such, it boasts a different and more informal engagement model, content type, algorithm, and audience (Gagliardi, 2024) compared to Udemy's formal educational content and structure (Udemy, 2023). Hence, one might observe a different

performance distribution and generative mechanism than the one found in the original study. Perhaps due to its newness TikTok remains underexplored in academia, despite its growing prevalence in our lives (Escolano, 2023), and especially in the lives of a younger generation that is now entering the workforce (Hazari & Sethna, 2023). These elements of novelty, growth, and entrepreneurial potential make TikTok an ideal digital platform for this analysis.

A replication study analyzing TikTok's performance distribution holds significant relevance in verifying the reliability and validity of previous findings (Lamal, 1990; Schmidt, 2009). It will help ensure that the results from the original study are not due to chance, but rather reflect true patterns or effects. In fact, the Udemy study's own authors call for replication pieces themselves (Gala et al 2024, pg. 19) because it allows researchers to confirm whether their observed patterns hold across different contexts and populations (Frank et al, 2010; Weismeier-Sammer, 2011), which aids in theory development. Additionally, given the rapid evolution of social media platforms like TikTok (Kullolli & Trebicka, 2023; Kushwaha, 2021), replication can help determine whether findings are robust and generalizable over time. This conceptual replication contributes to the cumulative body of evidence in the field by identifying conditions that may influence underlying mechanisms in TikTok's performance distribution; addresses possible limitations of the original study (Schmidt, 2009); and advances our understanding of contentpreneurial behavior on platforms like TikTok.

Moreover, a novel study centered on social media as a tool for entrepreneurship is timely. Statistics show that as of October 2023, out of 5.3 billion internet users, 4.95 billion are social media users, with a social network penetration rate of 64% (Statista, 2023). A recent Global Web Index survey revealed that social media dominates online time, with 98% of online users spending more than 2.5 hours on social media alone (Kemp, 2023). In a broad way these numbers reflect the extent to which social media has become an integral part of our lives (Chen et al, 2020), including how we communicate, connect, and create content. Consequently, this lifestyle transformation of social communities gathering in digital spaces has had a tremendous impact on the world economy. In fact, as the rise of digital platforms has disrupted established industries and even created new ones in the past couple of decades, today's digital economy alone makes up an estimated 15% of the global GDP, and has been growing at 2.5 times faster than the physical global economy (Hayat, 2022). For those seeking ways to exploit opportunities

online, the revenue acquisition potential imbued within digital platforms is undeniable and, it seems, only growing.

Against the backdrop of the above arguments, this paper undertakes the investigation of the performance distribution of TikTok contentpreneurs and its generative mechanism using a dataset of 1,046 randomly selected data samples for a statistical analysis that includes multiple nonparametric tests of correlation (Artusi et al, 2002) and goodness-of-fit distribution (D'Agostino et al, 1990). It finds support for some of the original study's hypotheses, mainly in what pertains to distribution non-normality and lognormal fit (Mitzenmacher, 2004), but no support for the hypothesis regarding knowledge and expertise (Dornekott et al, 2021; Morgeson & Humphrey, 2006) as a predictor of extreme performance (Hill, 1975), and no support for proportional differentiation as the distribution's generative mechanism (although caution is advised in interpreting the latter two findings, and further discussion is conducted later on for a more nuanced understanding (Mitzenmacher, 2004)). These findings contribute theoretical and practical insights to the fields of contentpreneurship (Johnson et al, 2022) and organizational performance (Andriani & McKelvey, 2009), as well as directives as to where the research could go from here.

1.2 Research problem

This paper tackles the challenge of investigating the prevalence of distribution models that lead to the occurrence of outstanding performers in the field of entrepreneurship (Aguinis et al, 2024; Andriani & McKelvey, 2009), especially since, traditionally, the assumption of normality has prevailed in past research in this field (Clark et al, 2023). It takes on this challenge by linking entrepreneurship research to two other topics of study that are relevant in their own fields but do not often overlap with entrepreneurial themes: content creation on social media platforms and the statistical implications of non-normal (also known as Paretian) performance distributions within teams and organizations (Andriani & McKelvey, 2009; Bannerjee & Yakovenko, 2010).

While some progress has been made in connecting non-normal distributions with digital entrepreneurship, such as Gala et al's 2024 piece upon which this paper is based, far fewer efforts have been extended towards exploring contentpreneurship more specifically (Johnson et al, 2022; Tang et al, 2023), along with its rapid evolution in recent years and the mechanism

behind its surge in popularity on social media platforms. Failing to address this gap in the literature limits our collective understanding of the entrepreneurial potential (Shane & Venkataraman, 2000) of content creation in exploiting online opportunities. It also prevents us from examining the nature of the platforms that engender contentpreneurial success, which is particularly salient for the up and coming generation (Breyer et al, 2019) whose demographic predominates among these platforms (Hazari & Sethna, 2022), and who are beginning to contribute to the global digital economy.

1.3 Research aim and question

The aim of this study is to put to test the assumptions posited by Gala et al (2024) in regards to digital entrepreneurship and performance distributions. It seeks to answer the question of non-normality and generative mechanisms pertaining to the performance distribution of the digital platform selected for this study in comparison to the findings for the digital platform from the previous study. More specifically, the question this study answers is whether TikTok's performance distribution is more likely to be lognormal (Joo et al, 2017) with a proportional differentiation generative mechanism (Mitzenmacher, 2004; Newman, 2005). The answer to this question is relevant because a lognormal distribution is indicative of the presence of star contentpreneurs on TikTok, and a proportional differentiation mechanism is indicative of the likelihood of becoming a star contentpreneur despite the competition. This paper also answers the question of whether greater knowledge in a field has a positive impact on generating star contentpreneurs on TikTok (Dornekott et al, 2021; Morgenson & Humphrey, 2006), as a way of predicting one's probability of success on the platform. If these questions are answered in the affirmative in this analysis, they corroborate the findings from the original study about digital platforms and contribute to their generalizability.

1.4 Context for the study

This thesis arises within a broader conversation surrounding top performers, or outliers, and how they appear across many fields and occupations where performance related data has been gathered (Aguinis & O'Boyle, 2012; Banerjee & Yakovenko, 2010; Joo et al, 2017). When data points are plotted on a graph they fall within a certain distribution pattern. For ease of calculation and understanding, the conventional practice among different fields of

study—including within the field of entrepreneurship (Booyavi & Crawford, 2023; Clark et al, 2023)—has been to drop the outliers from the data to obtain a more homogenous distribution fit (Andriani & McKelvey, 2009; Crawford et al, 2015). However, more recently this practice has come under greater scrutiny, with many arguing that the outliers that create the heavy-tail effect of a distribution ought to be studied more carefully (Clark et al, 2023). This is because these "throwaway[s]" (Andriani & McKelvey, 2009, pg. 1064) often represent the most outstanding performers within the data set, signaling some kind of unusual activity most organizations would likely want to take a closer look at. Heavy-tailed distributions are defined by asymmetry, non-linearity, interdependence, and a certain unpredictability (Hill, 1975). In statistical terms heavy-tailed distributions are referred to as non-normal distributions and vary in scale and shape. Examples of non-normal distributions can be seen in figure 1.

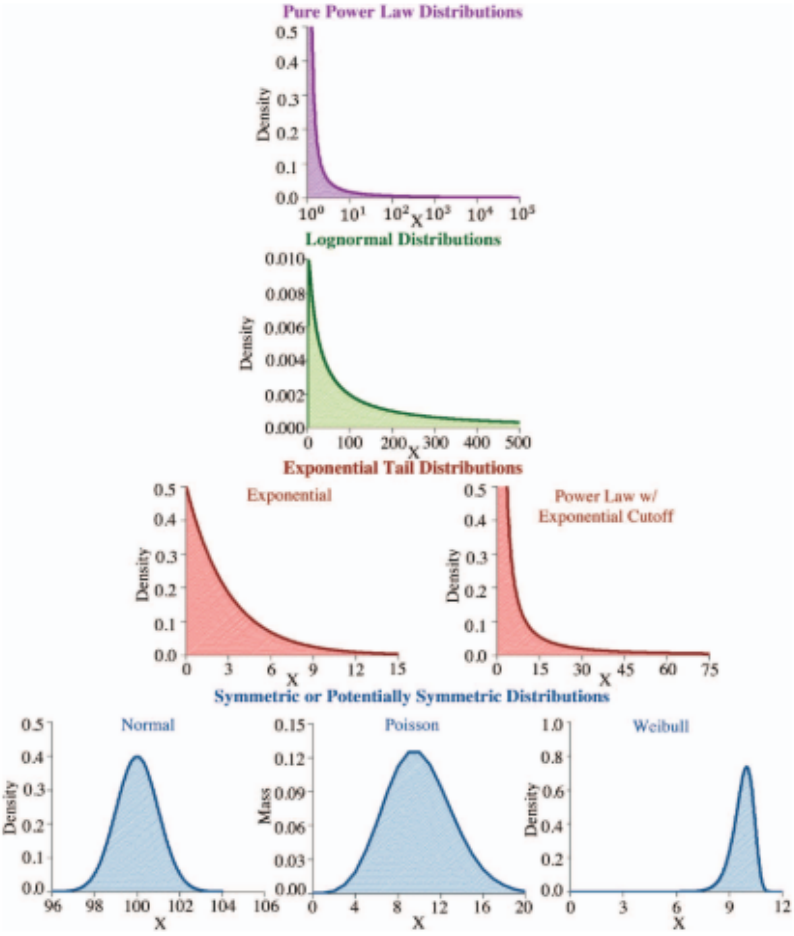


Figure 1: Joo et al's non-normal distribution taxonomy (2017)

Figure 1 was taken from Joo et al's (2017, pg. 1024) non-normal distribution taxonomy, upon which Gala et al (2024) based their "Star Entrepreneur" analysis. This taxonomy consists of a total of seven distributions, grouped into four categories: 1. Pure power law; 2. Lognormal; 3. Exponential tail; and 4. Symmetric or potentially symmetric. These particular categories were selected because one or more of the seven distributions within them tend to explain the vast majority of natural phenomena (Bak, 1996), and have been used in previous research across a variety of subjects (Aguinis et al, 2024; Banerjee & Yakovenko, 2010; Joo et al, 2017). In Gala et al's (2024) study, Udemy's distribution was found to fit the lognormal distribution the best (shown above in green), confirming its hypothesis that entrepreneurial performance on digital platforms is best represented by lognormal distributions. This replication study seeks to ascertain whether that holds true for TikTok. Table 1 on page 13 provides a brief summary of the four distribution categories from the taxonomy, including their most distinguishing characteristics and the generative mechanisms that tend to be associated with each one.

Examining the non-normal distributions of star performers in entrepreneurship is particularly salient when it comes to social media platforms, a prevalent topic of study in which stardom remains elusive for the vast majority (Chen et al, 2020). Here is a snapshot of the current reality of some popular content-creation based platforms: out of 114 million Youtube channels, only five have 100 million or more subscribers (Queen, 2022). Ten percent of all X (former Twitter) users are accountable for 92% of tweets shared on the platform (GilPress, 2023). While 75% of Instagram users have less than 10,000 followers, the top account has more than 600 million (Moore, 2024). And out of the ten most followed TikTok accounts (which already boast millions more than the average account), the top two outdo the other eight by a range of 50 to 90 million followers (Dean, 2023).

Out of all these popular platforms in today's social media landscape, TikTok stands out as an especially fitting context for this study because of its own unique rise from relative obscurity to becoming the most widely used content creation platform around the globe (Duarte, 2024). What originated as an outlet for individual self-expression (Cutolo & Grimaldi, 2023; Gustafsson & Khan, 2017) during worldwide lockdowns (Frenkel, 2022) quickly evolved into a self-marketing tool, becoming the social media platform users spend the most time on on a daily

Type of distribution	Characteristics	Generative Mechanism	Parameters
Pure power law	● long head ● very long right tail (seemingly infinite) ● top performers are very highly distinct	Self-organized criticality, preferential differentiation*: After some individuals's performance reaches critical states, any events trigger increases on their outcome; rich-get-richer effect.	(α) (>1) = rate of decay
Lognormal	● fat but finite right tail ● bell-shaped head ● top performers are highly distinct	Proportional differentiation: individual performance differs in terms of initial value and accumulation rate.	(μ) (>0) = mean (σ) (>0) = standard deviation
Exponential tail	● positively skewed tails that fall rapidly ● long head and somewhat heavy right tail ● top performers can be similar, highly distinct, or something in-between	Incremental differentiation: Accumulation rate may be subject to diminishing returns; outcome is only subject to accumulation rate, not initial value.	(α) (>1) and (λ) (>0) = rate of decay
Symmetric or potentially symmetric	● Symmetric or nearly symmetric tails (normal distributions, bell curve) ● values are normal to discrete	Homogenization: Reduces performance differences among individuals over time in terms of their values on an outcome.	(μ) (>0) = mean (σ) (>0) = standard deviation (λ) (>0) = the extent to which the distribution is pushed down and stretched sideways (β) (>0) = the extent to which the distribution's head is pushed to the right

Table 1: A summary of the four main categories of Joo et al's (2017) distribution taxonomy.

basis (Duarte, 2024; Kemp, 2023). Its rapid ascent to becoming the most downloadable social media app has far surpassed the much slower ascent of that of its counterparts in previous years (Dean, 2023), and today draws in more than 900 million users every day (Duarte, 2024). In addition, and more closely related to this study, TikTok's unique algorithm (Ehret et al, 2023) allows users to see content made by creators they do not have to be subscribed to, thus increasing

the possibility of views, likes, and shares, making for intriguing performance distribution possibilities.

1.5 Delimitation of the study

Admittedly, without more context the numbers mentioned in the paragraphs above do not tell us much about the reasons why some contentpreneurs do exponentially better on these platforms than others. Ultimately though, the reasons behind star contentpreneurship do not matter for this thesis as much as the fact that star contentpreneurs simply exist. Again, this is because the extreme gap in individual output on digital platforms signals a heavily right-tailed non-normal distribution similar to what has been noted across several occupations, industries, and disciplines in relation to individual performance (Aguinis & O'Boyle, 2014; Andriani & McKelvey, 2009; Clark et al, 2023; Newman, 2005). Thus, the interest of this study lies less in understanding why contentpreneurship yields non-normal distributions, and more in verifying whether generalizations can be made as to which type of non-normal distribution and generative mechanism best capture performance outcome on digital platforms, even for platforms as different from each other as Udemy and TikTok.

1.6 Contribution of the study

This verification is not simply a matter of curiosity, but carries with it a unique set of theoretical and practical implications for improved contentrepreneurial performance. This paper aims to contribute to the field of entrepreneurship in a number of ways. First, it responds to Anderson et al's (2019) call for more stringent and replicable quantitative research to be conducted within the field of entrepreneurship. Second, it strengthens existing research linking entrepreneurship with heavy-tailed performance analyses (Aguinis et al, 2024; Booyavi & Crawford, 2023; Clark et al, 2023; Crawford et al, 2015). Third, it focuses on a fresh and novel platform that carries unusually high entrepreneurial potential, and differs significantly from Udemy. For instance, one is formal and academic in nature, with content that must be purchased and can only be used with verified credentials, while the other is primarily used for entertainment (Rooney, 2020), boasts millions more users (Dean, 2023), has a unique algorithm (Ehret et al, 2023) in which content can be seen by any online user, focuses predominantly on short-video format (Ceci, 2023; Duarte, 2024; Kolsquare, 2023) and can be used as an audience builder by

content creators. Four, it advances the literature on contentpreneurship (Frenkel, 2021; Johnson et al, 2022), building on Nambisan's (2017) recommendation to incorporate "methodological approaches that reflect the incremental and nonlinear paths that digital artifacts and platforms facilitate in entrepreneurial initiatives" (pg. 1042), a theme underscored by non-normal distribution's nonlinearity and variance (Hill, 1975). And five, it uses its findings to inform both entrepreneurial theory and practice on late entrant (Mitzenmacher, 2004) content creators on digital platforms for increased chances of success.

1.7 Definition of key terms

A central piece of this paper is the discussion of topics that are well known within the realm of statistics but perhaps less known in the field of entrepreneurship. Within that discussion three key terms are used often throughout the paper, and thus understanding them is crucial. These terms are defined below:

Term	Definition
Non-normal distribution	A function of the probability that variables in a dataset do not occur in a normal (symmetric or bell-curve) pattern. A non-normal distribution may have skewness, kurtosis, or other characteristics that make it deviate from a symmetrical shape.
Lognormal distribution	A specific type of non-normal distribution consisting of a heavy but finite right tail with a bell-shaped head and extreme outlier values that fall rapidly at the highest values of observations. A non-normal distribution follows a lognormal fit if its logarithm results in a normal distribution.
Generative mechanism	A process that explains how data is produced by specifying underlying variables and their distributions. Specific generative mechanisms are often associated with certain distribution shapes. The generative mechanism typically associated with a lognormal distribution is called proportional differentiation.

1.8 Disposition of the thesis

Following the introduction, the remainder of this paper consists of five sections structured as follows: section 2 contextualizes the current academic conversation about contentpreneurship and performance distribution within a theoretical framework, and develops the five hypotheses to be tested. This is followed by section 3, the methodology, which delves into the data collection process, including ethical considerations and a comprehensive list of all variables selected for testing. Section 4 covers the results of the statistical analysis in detail, and section 5 interprets those results with insights for future practice and theory. The paper ends with section 6, the concluding thoughts.

2. Theoretical Framework

2.1 Contentpreneurship

Contentpreneurs can be defined as entrepreneurs whose business strategy centres on exploiting revenue streams (Baron, 2006; Shane & Venkataraman, 2000) derived from content created and disseminated on social media platforms (Frenkel, 2021; Gustafsson & Khan, 2017; Johnson et al, 2022) often in the form of text, photo, or video (Ashman et al, 2018; Chen et al, 2020). Monetization of content can take place on-platform (generated directly through the platform) and off-platform (revenue earned outside of the platform) (Cutolo & Grimaldi, 2023; Johnson et al, 2022) through a number of different venues: running ads; brand sponsorships; selling products and merchandise (Gagliardi, 2024); promoting existing businesses; collaborating with other influencers and contentpreneurs (Campos et al, 2023; Gagliardi, 2024; Srinivasan & Venkatraman, 2017); etc.

Largely underexplored as of yet (Johnson et al, 2022), contentpreneurship exists as a budding niche under the overarching umbrella of the more developed strand of literature, digital entrepreneurship (Nambisan, 2017; Sought et al, 2019; Yoo et al, 2012). Recent studies in digital entrepreneurship have delved into the unique characteristics of digital technologies (Breyer, 2019; Nambisan, 2017; Yoo et al, 2012), highlighting the collaborative nature (Campos et al, 2023; Srinivasan & Venkatraman, 2017) and fluid boundaries (Nambisan, 2017) of digital platforms, and how they enable the average person with no business or technical skills to become a content creator (Chen et al, 2020).

Indeed, digitization has reshaped the dynamics of value creation (Yoo et al, 2012) for contentpreneurs around the globe, with digital products, services, and interaction fostering greater market accessibility for the wider public (Chatradhi et al, 2023; Ebrahimi et al, 2023; Sahut et al, 2019). This has been a game changer for ordinary creatives (Frenkel, 2021; Gustafsson & Khan, 2017) who can share all kinds of appealing content that might attract a community of followers (Chen et al, 2020), blurring the lines between consumers and creators in a synergetic ecosystem of entrepreneurial activity (Campos et al, 2023). Further, digital platforms' ability to facilitate and reduce transaction costs (Ebrahimi et al, 2023) have not only disrupted conventional industries (Gala et al, 2024), but have also contributed to a digital economy hailed as one of the most significant economic developments since the industrial revolution (Breyer, 2019; Ibrar et al, 2022).

At the centre of this digital economy are contentpreneurs, often in "the quiet desperation" of their bedrooms (Ashman et al, 2018, pg. 481), leveraging their online personas to eke out an income (Dayan & Tafesse, 2023; Tang et al, 2023) while juggling a digital side hustle or two (Frenkel, 2021). Indeed, the emerging literature on today's contentpreneurs' identity reveals a certain pattern of scrappiness and resilience (Ashman et al, 2018; Garvey et al, 2023), perhaps best encapsulated by Li & Wang's (2024) usage of the term "underdog [cont]entrepreneurship." But to those aspiring for digital celebrity status (Chen et al, 2020), Ashman et al (2018) warn about the "cruel optimism" (pg. 475) that comes with this competitive landscape—one that promises much but often delivers little. After all, despite the openness and accessibility of the internet, relatively few will reach the level of stardom (Gala et al, 2024) needed on digital platforms to truly influence others' consumption (Chen et al, 2020). Missing in the extant literature is not only a better understanding of what engenders this performance disparity among contentpreneurs, but also some kind of guiding posts that could help level the playing field.

Still, many contentpreneurs battle on, catapulted by the challenges as well as opportunities of the recent pandemic (Crespo et al, 2024; Dornekott et al, 2021; Frenkel, 2021; Garvey et al, 2023; Ebrahimi et al, 2023), by continuing to exploit the accessibility of digital platforms to promote their businesses, build an audience, and, if the stars align, monetize content (Campos et al, 2023; Cutolo & Grimaldi, 2023; Dayan & Tafesse, 2023; Dornekott et al, 2021; Li & Wang, 2024). With million others doing the same, the current literature lacks a clear

visualization of the performance gap between the average contentpreneur and the top elite monetizing the most.

2.2.1 Social Media Platforms and Gen Z

Web 3.0 has been largely defined by the ascent of social media platforms (Ibrar et al, 2022), a paradigm shift from static web pages to immersive, community-driven interfaces (Escolano, 2023; Kullolli & Trebicka, 2023) that use sophisticated algorithms to enhance user engagement (Ehret et al, 2023). Chen et al (2020) argue that the most unique feature of social media platforms is the ability to "self-generate and self-publish content (pg. 35)," unlocking a certain democratization of an online space in which anyone can partake in content creation.

Within this current landscape, scholars have been particularly interested in the interaction patterns generation Z exhibits with social media (Kullolli & Trebicka, 2023), for a number of reasons. As the first generation to have been born into a digitized world, gen Z's ability to integrate digital platforms into their daily life is greater and more seamless than their millennial counterparts' (Kushwaha, 2021). Gen Z makes up one-third of the world's population (Hazari & Sethna, 2022) that is now beginning to contribute to the global economy, bringing its digital competencies and preferences along with them. This is particularly salient for the future of entrepreneurship, since digital competence has a positive correlation with entrepreneurial intention (Bachman et al, 2024).

Studies also reveal the younger generation's preference for peer-to-peer, visually rich, and video-based content, which would explain gen Z's drifting away from Facebook (Rooney, 2020) and gathering more predominantly on platforms like Instagram (Hazari & Sethna, 2022), Youtube (Chiang & Jang, 2023), and Tik Tok (Escolano, 2023; Ehret et al, 2023). These characteristics matter for this paper since it analyzes contentpreneurial performance specifically on Tik Tok, the social media platform most occupied by gen Z'ers. The platform's main audience likely influences the nature of the content most widely disseminated, as well as the following count of different content creators on the platform.

2.2 Performance Distributions

For the past several years a number of academic fields have turned their attention to the prevalence of non-normal distributions in individual performance (Aguinis & O'Boyle, 2012;

Aguinis & O'Boyle, 2014; Joo et al, 2017; Newman, 2005). Despite a general assumption that performance tends to fall within a normal, linear, and mostly symmetric distribution (Andriani & McKelvey, 2009)—one can envision a bell curve—growing evidence suggests that indeed it is a non-normal distribution that abounds in both natural (Bak, 1997) and man made phenomena (Newman, 2005), and prevails across different fields, industries, and professions (Aguinis & O'Boyle, 2014; Joo et al, 2017). Also known as the Pareto effect (Andriani & McKelvey, 2009; Mitzenmacher, 2004), these power law distributions refer to heavy-tailed plotted graphs where a disproportionate amount of the total output is captured by a small group of extreme performers (Andriani & McKelvey, 2009; Crawford et al, 2015). Whereas a bell curve or lognormal distribution is interpreted by its midpoint (the mean, or average μ) a Pareto or non-normal distribution has no predictable midpoint and is interpreted by its tail (Aguinis & O'Boyle, 2014; Andriani & McKelvey, 2009; Hill, 1975). The graph below, taken from Aguinis & O'Boyle, 2012 (pg. 80), provides a visual of both kinds of distributions in juxtaposition.

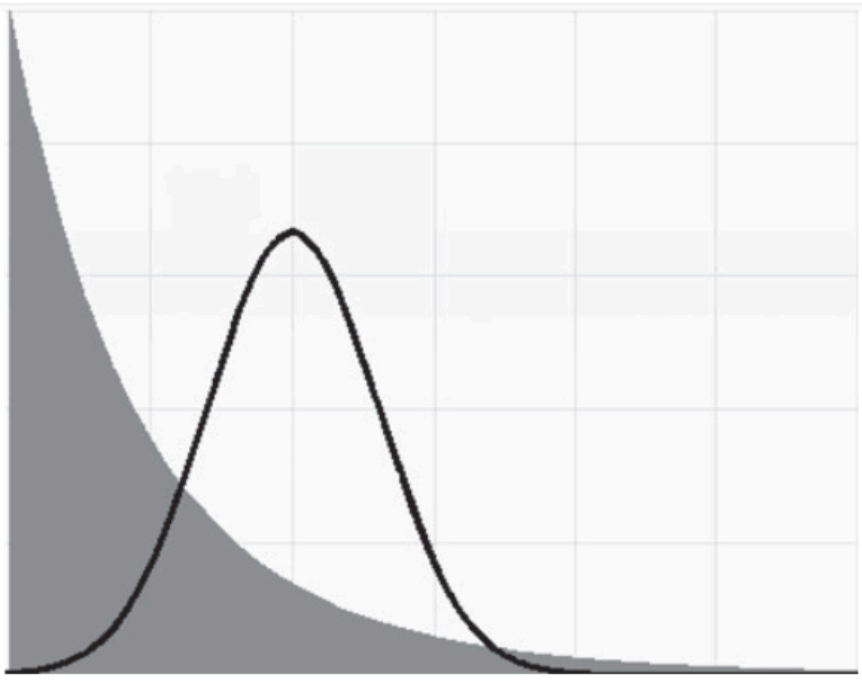


Figure 2: A normal distribution (black) overlaying a Paretian distribution (gray).

Many have highlighted the merits of studying the tail end of distributions (Andriani & McKelvey, 2009; Booyavi & Crawford, 2023; Clark et al, 2023; Crawford et al, 2015; Hill, 1975), arguing that not only is the Pareto effect of tail activity much more common across the

board, but is arguably more interesting than the cluster around a bell curve's average (Clauset et al, 2009). Andriani and McKelvey (2009) posit that tail ends are particularly salient to the field of entrepreneurship as they represent outliers (pg. 1056)—the black swans who stand out and disrupt the system through innovation (Clark et al, 2023), arguably the main goal of entrepreneurship. When we focus on the very outliers that often get eliminated from data sets in order to obtain more predictable distributions (Crawford et al, 2015), we essentially shine a spotlight on those who Booyavi and Crawford (2023) call the "rock star" entrepreneurs in their fields. This allows us to better understand high performing entrepreneurs' influence on their peers (Hendricks et al, 2022); analyze whether exceptional employees become better bosses (Kim & Makadok, 2022); more accurately reflect gender differences in new venture outcomes (Booyavi & Crawford, 2023); strengthen the source of an organization's greatest value creation (Cascio et al, 2022); and improve organizational theories and practice (Joo et al, 2017), among many other compelling performance related elements in entrepreneurship.

While some have voiced concerns about making widespread assumptions about power law effects (Clauset et al, 2009; Mitzenmacher, 2004), citing issues in replicability (Clauset et al, 2009; Beck et al, 2013), others have encouraged entrepreneurial research to embrace this outlier approach (Andriani & McKelvey, 2019; Clark et al, 2023). Afterall, the goal in identifying whether a data set follows a normal or non-normal distribution is to view the results as a continuum, not a dichotomy (Beck et al, 2013 pg. 539). In other words, when undertaking a study on entrepreneurial performance distribution (or, in this case, on contentpreneurial performance distribution), the objective is not to achieve absolute certainty of a normal vs non-normal distribution (Clauset et al, 2009), but to come as close as possible to finding the best fit for that particular data set (Beck et al, 2013) along with its particular generative mechanism (Joo et al, 2017). While some effort has been made to guide the field of entrepreneurship away from its dominant lens of normality and homogeneity (Aguinis et al, 2024; Andriani & McKelvey, 2009; Booyavi & Crawford, 2023; Clark et al, 2023; Crawford et al, 2015; Gala et al, 2024), there has been no comprehensive push for an outlier-driven approach to help the field systematically embrace outstanding performance going forward, let alone on a content creating social media platform.

2.3 Hypotheses development

This study is grounded in a positivistic ontology and adopts a deductive research approach, in which hypotheses are formulated based on existing theories and tested through empirical methods. Gala et al's (2024) original five hypotheses tested Udemy's distribution normality and fit, success prediction, and generative mechanism. The same hypotheses are tested for TikTok's performance distribution, but employ some slight modifications in terminology, in keeping with the key terms used in this paper thus far. More specifically, this study replaces Gala et al's (2024) oft used term "entrepreneurs on digital platforms" with the term contentpreneur (Frenkel, 2021; Johnson et al, 2022), with the understanding that any contentpreneurial performance being measured naturally takes place on a digital platform. This section presents the five hypotheses to be tested, with the updated terminology.

2.3.1 Contentpreneurial performance

The original study posits that the performance distribution of contentpreneurs is likely to be non-normal (pg. 4), because the presence of a few top performers who have the majority of subscribers (or followers) skews the distribution away from the mean creating the long tailed, non-normal Pareto effect. The authors argue that the same Pareto effect found in physical entrepreneurial contexts (Joo et al, 2017) should also be reflected in digital contexts, and may in fact be magnified given the nature of digital technologies in enabling venture creation (pg. 4). The low marginal costs and few physical constraints of digital platforms (Nambisan, 2017) enable more rapid growth and a broader market reach than those found in traditional market contexts. This accelerated momentum within the digital realm can be further reinforced by positive loops of early subscribers (ie.: customers who leave positive reviews and/or purchase again, or followers of one's page who comment and subscribe, etc) (Ashman et al, 2018; Donaker et al, 2019). When exploited, these elements act as a launching pad for digital success, creating the outlier effect of a non-normal distribution. Thus, the first hypothesis in Gala et al's 2024 study links entrepreneurial performance on digital platforms with a non-normal distribution.

Hypothesis 1: Performance distributions for contentrepreneurs are more likely to resemble non-normal distributions than normal distributions.

2.3.2 Lognormal distributions

If the first hypothesis holds true, then it follows that the particular non-normal performance distribution of contentpreneurs must be of a certain category, shape, and scale, given Joo et al's (2017) taxonomy described above and illustrated in figure 1. The authors predict that from that taxonomy, lognormal distributions best display the shape of contentpreneurial performance, especially when compared to the other four categories of distributions (Gala et al, 2024, pgs. 4-6). This is because lognormal distributions are characterized by particularly fat and finite tails (Andriani & McKelvey, 2009; Hill, 1975), indicating a greater prevalence of extreme performers than in normal or exponential tails, but also curtailed by the finite number of users on the platform, which differentiates it from power law fits with infinite variance. In practicality, this characteristic is exemplified by the disparity between those with exponentially more online customers or subscribers than their peers using the same platform, often because of the winner-take-all effect (Gala et al, 2024, pg. 2). A lognormal distribution, or a distribution with a particularly heavy tail, makes the most sense when marginal costs are low (Joo et al, 2017, pg. 1030), which is almost always the case with digital contexts. But these heavy tails are also finite, with more of a blunt cut-off, as opposed to a pure power law's tail which is slender but carries on seemingly indefinitely (Hill, 1975). The authors see the finite, blunt cut-off of the lognormal tail as fitting with the limitations of digital platforms' market size and rising costs (pg. 4). Therefore, the second hypothesis links entrepreneurial performance on digital platforms with lognormal distributions in particular.

Hypothesis 2: Performance distributions for contentpreneurs are more likely to resemble lognormal distributions than (a) pure power law, (b) exponential tail, (c) power law with exponential cutoff, or (d) symmetric distributions.

2.3.3 Knowledge intensity within a TikTok domain

The third hypothesis suggests tail extremity is positively correlated with knowledge intensity of a certain domain (Gala et al, 2024, pgs. 6-7). In the study, domains on a digital platform (such as "Business," "Art," or "Literature") represent different jobs and occupations, and highlight the importance of measuring a contentpreneur's skills and abilities within their domain as an element that might influence their performance. The authors draw upon an existing

definition of knowledge intensity by Morgenson and Humphrey (2006) which includes the five following components: complexity, information processing, problem-solving, skill variety, and specialization. These knowledge characteristics are especially relevant to digital platforms because the transferring of information from the content creator to the end user plays a central role in how a product or a service is created, communicated, and delivered (Chen et al, 2020; Gustafsson & Khan, 2017). Contentpreneurs with higher cognitive and intellectual resources can better exploit the vast amounts of online data to curate their content (Bachman et al, 2024), understand their customers and competitors, and collaborate with other stakeholders (Campos et al, 2023). These resources and collaborations will likely lead to greater success in one's role as online content creators (Dornekott et al, 2021; Tang et al, 2023), thus affecting the tail end of performance distribution. Each one of these components represents a unique facet of a professional's ability to perform well in their field, or domain (Gala et al 2024; Morgenson & Humphrey; 2006), and are taken into consideration later on when selecting the variables to measure knowledge intensity.

Hypothesis 3: Knowledge intensity of a domain is positively associated with the tail extremity of the performance distribution for contentpreneurs.

2.3.4 Generative Mechanism

Hypothesis four is split into (a) and (b) and both have to do with the correlation between a lognormal distribution's specific generative mechanism on performance. By way of reminder, a generative mechanism is the "process leading to the existence of the focal distributional shape for the phenomenon under investigation (Joo et al, 2017, pg. 1025)." Referring back to Table 1, each distribution type in the taxonomy tends to be associated with a certain generative mechanism for that particular non-normal distribution, although exceptions exist (Mitzenmacher, 2004). Lognormal distributions are often indicative of a generative mechanism called proportional differentiation (Banerjee & Yakovenko, 2010; Mitzenmacher, 2004), which encompasses two key components: initial value and accumulation rate (Joo et al, 2017, pg. 1030).

2.3.4.1 The rich get richer effect

Proportional differentiation is a measure by which an initial value is multiplied by its accumulation rate (Andriani & McKelvey, 2009; Clark et al, 2023; Gala et al, 2024; Joo et al, 2017). In a venture setting, the initial value may be thought of as the initial number of customers, funds, resources, etc, which increase over time according to a certain accumulation rate (Mitzenmacher, 2004). In a digital venture setting, in which entry and market access are more fluid and low costs facilitate rapid scaling, a similar concept of accumulation rates can be applied to early entrants who benefit from public feedback loops (reviews, ratings, shares, likes) (Donaker et al, 2019), establishing what the authors call the "rich get richer" effect (Gala et al, 2024). Thus, hypothesis 4a essentially posits that initial success breeds more success.

Hypothesis 4.a: Higher initial performance has a positive effect on the future performance of contentpreneurs.

2.3.4.2 Accumulation rate effect

For a contentpreneur starting out later, though, or with lower initial value, the initial disadvantage can still be overcome through proportional differentiation's cumulative effect that allows one to eventually outperform a competitor over a sustained period of time (Aguinis et al, 2016; Gala et al, 2024, pg. 8). On a platform like TikTok, late entrants can still build a considerable following through consistent and strategic posting, which can increase the possibility of shares and likes because of the accumulation rate effect (Chen et al, 2020; Escolano, 2023) . The accumulation rate effect can also influence the platforms' algorithm (Ehret, 2023; Kullolli & Trebicka, 2023), increasing the chances of a late entrant's videos coming up in searches and in users' news feeds, positively affecting overall performance. This leads to hypothesis 4b, which suggests a positive correlation between the rate of performance accumulation on future performance.

Hypothesis 4.b: Relative rate of performance accumulation has a positive effect on the future performance of contentpreneurs.

3. Methodology

3.1 Data collection

To conduct this study, an application was submitted to TikTok's official research API in the beginning of February 2024 to access its data (tiktok.com, "Research API"), and within a few weeks the application was approved. Once permission from TikTok was granted, the necessary information was generated on Spyder, an open-source integrated software for scientific programming, using coding available online (Taboada-Villamarín, 2023) and manipulated to fit the study's specific parameters. Since the API had many limitations as to the type of information it could generate, some of the variables were also collected manually from publicly available information on TikTok. Data samples were generated randomly using search terms that were included into the coding language. These search terms were selected from "dictionaries"[1]—short lists containing different combinations of hashtags and grouped into categories (ie.: #easyrecipes for food related accounts; #growthmindset for psychology related accounts, etc. The table of dictionaries can be found in Appendix 2).

Once the daily limit for data generation would be reached using the API, some data sample collection would also be supplemented manually. This was necessary in order to adhere to the data collection timeframe of the study, as well as that of TikTok's API access, which was only granted for a relatively short period of time. The manual selection entailed entering the same search terms from the dictionaries into TikTok's search bar and mechanically selecting the first user accounts the search would generate in an attempt to mitigate potential biases and to ensure a diverse and representative sample selection. Due to the thirty-day limit that the API places on data searches, the variable "Shares" only captures the total number of videos shared from March 7th to April 6th per user account.

The data collection was structured in a way so as to replicate as closely as possible the way in which data was collected from Udemy for the original study (Gala et al, 2024, pg. 8). The authors treated each Udemy instructor as an entrepreneur, regardless of employment status, based on the definition of entrepreneurs as those who identify and exploit opportunities (Baron, 2006; Shane & Venkataraman, 2000). Following this same definition, this thesis also treats each

[1] A special thank you to Hassan Hamadi for his suggestion to create dictionaries to extract data samples more systematically.

content creator on TikTok as an entrepreneur (Frenkel, 2021; Johnson et al, 2022).

Udemy has thirteen categories with several subcategories, or domains. The authors randomly selected four domains within each category, making a total of 52 domains tested. The different domains represent jobs and occupations, so selecting data within each domain was important to ascertain whether star performers can be observed across different occupations on digital platforms (Gala et al, 2024, pg. 8). While the content on TikTok is not necessarily divided by categories and domains in the same way as Udemy (the content on TikTok is far too varied and unstructured (Ehret, 2023; Kullolli & Trebicka, 2023)), it does have an endless number of communities that can be found using hashtags (hence the dictionaries of hashtags). These communities focus on many different areas of interest, from hobbies to personal development to professions, and many partnerships between TikTok contentpreneurs and companies are often formed through the activity that takes place in these communities (Conrad, 2022).

While there are far too many of these groups to count, the top 21 most active and popular TikTok communities are listed below (Conrad, 2022). For the purpose of this study each one of these communities represents a domain similar to the 52 domains analyzed on Udemy. Fifty users from each community were randomly selected using the search terms from the dictionaries mentioned above, making up a total of 1,050 data samples. After deleting duplicates and private accounts a total of 1,046 contentpreneur profiles were used for the analysis.

#AdviceTok	#DanceTok	#MovieTok	#PlantTok	#TVTok
#ArtTok	#DIYTok	#MusicTok	#PrankTok	
#BeautyTok	#FashionTok	#NurseTok	#QueerTok	
#BookTok	#FitTok	#ParentTok	#SelfCareTok	
#CookTok	#KidTok	#PetTok	#SportsTok	

3.2 Variables

The original study analyzed two dependent variables, two independent variables, and three control variables (Gala et al, 2024, pgs. 9-11). This study attempts to replicate the variables as closely as possible with equivalent proxy measures selected for TikTok. The table below compares the measures selected by each study side by side, followed by an explanation of why these are a good representation of the variables, as well as why some modifications were needed.

		Original (Gala et al, 2024)	Replication (TikTok thesis, 2024)
Dependent Variables	Entrepreneurial Performance	Cumulative number of students enrolled across all paid courses offered by a Udemy instructor (alternatively, the cumulative number of reviews received across all paid courses offered by a Udemy instructor).	Cumulative number of followers across all TikTok user accounts (alternatively, the cumulative number of "likes" across all TikTok user accounts).
	Tail extremity of performance distribution	Scale (μ) and shape (σ)	Scale (μ) and shape (σ)
Independent Variables	Knowledge Intensity	Measured using five knowledge dimensions across courses in a domain: a) Complexity (avr. # of lectures) b) Information processing (avr. video length) c) Problem-solving (avr. # of downloadable resources) d) Skill variety (avr. # of instructors per course) e) Specialization (avr. # of informational articles)	Measured using five knowledge dimensions across hashtag communities: a) Complexity (avr. # of videos posted) b) Information Processing (avr. video length) c) Problem-solving (avr. # of "shares") d) Skill variety (avr. # of accounts being followed) e) Specialization (niche content)

	Initial Performance and performance accumulation rate	1) Initial value of instructor performance: the cumulative number of students at T1; 2) Accumulation rate: the number of students added over the following 16 weeks.	1) Initial value of contentpreneur performance: the cumulative number of followers at T1; 2) Accumulation rate: the difference in number of followers over the following 4 weeks.
Control Variables	Differences in: 1) Educational level 2) Professional commitment 3) Use of social media	1) Degree: presence of a degree (binary value) 2) Professional: presence of a website (binary value) 3) Social: number of social media platforms for each instructor	1) Degree: presence of a degree in bio (binary value) 2) Professional: presence of a website (binary value) 3) Social: number of social media platforms for each user 4) Extra control - Activity level: frequency of videos posted/month.

Table 2: Side by side comparison of variables from original study and the replication.

3.2.1 Dependent Variables

3.2.1.1 Contentpreneurial Performance

The first dependent variable is referred to as "contentpreneurial performance," and in the closest approximation to the original study is measured by both the number of followers per contentpreneur, as well as the number of likes per contentpreneur, on TikTok.

3.2.1.2 Tail Extremity

The second dependent variable employs distribution parameters of scale (μ) and shape (σ), as a way of measuring the heaviness of a tail (Hill, 1975; Joo et al, 2017). Scale represents the location, or central tendency of the distribution, while shape indicates its spread, or variability (Hill, 1975; Mitzenmacher, 2004). In other words, in a lognormal distribution the parameter of scale (μ) represents what would typically be the mean of a normal distribution, and shape (σ) is equivalent to a normal distribution's standard deviation (Hill, 1975). Following Gala et al's testing (2024, pg. 9) these parameters will be used for testing hypotheses 2 and 3, but preference will be given to scale as opposed to shape as it is deemed a better parameter for "fatty" rather than longer tails. In hypothesis 3 shape will be used as a control variable.

3.2.2 Independent Variables

3.2.2.1 Knowledge Intensity

As previously mentioned, knowledge intensity is a concept developed by Morgeson & Humphrey (2006) that encompasses five knowledge dimensions. In deploying knowledge intensity as one of their independent variables, Gala et al (2024) use proxy measures for each of the five knowledge dimensions based on information available to them on Udemy, then normalize and average the overall scores of each component into an overall measure. In this study, proxy measures available on TikTok are used to mimic the authors' proxy measures as closely as possible, but some changes were made to measures 3 (problem solving), 4 (skill variety), and 5 (specialization) due to the different nature of TikTok compared to Udemy.

In the original study, *problem-solving* is measured as the average number of downloadable resources included in a Udemy course to indicate a higher level of

problem-solving resources (pg. 9), but this is a non-existing feature in TikTok. Similarly, *skill variety* is measured as the average number of instructors per course across course domains, where the greater number of instructors reflects a greater variety of skills (pg. 9). TikTok does have a feature in which a user can "stitch" their video with someone else's (Okumoko, 2022), linking one's knowledge with another's and arguably having a similar effect as the multi-instructor variable from the original study. But stitched videos can be hard to identify, as they are not always prominently displayed on most accounts, and can sometimes be indistinguishable from a plagiarized video simply taken from another's account, without the stitching element. Lastly, *specialization* is originally measured by the average number of informational articles across courses within a domain, linking courses with a larger number of articles to more specialized domains (Gala et al, 2024, pg. 9). On TikTok, however, there is no equivalent to informational articles, as there can be no attachments to videos or on a user's profile in general.

For these reasons, the problem-solving dimension was replaced by average number of video "shares" across a community domain, based on the assumption that a high share count for a video can be correlated to how well the content of the video provides a highly sought solution to a problem. Skill variety registers as the average number of user accounts being followed by other users across a community domain, which reflects a certain level of deeper engagement on the platform with fellow contentpreneurs with a wide range of skills who post content on a variety of topics. And the specialization dimension was modified to a binary variable of "niche" topics, where a 1 indicates that the content created by a certain user displays specialty, uniqueness, and skill, and a 0 if those elements are not present.

The other two dimensions of knowledge intensity (1. complexity and 2. information processing) had closer equivalents between Udemy and TikTok. The number of videos posted across a community is used to reflect *complexity*, based on the idea that the more complex a topic, the more content is needed to explain it. *Information processing*, in turn, is computed by the average video length across a community (Ceci, 2023), because longer videos require followers to consume more content (Gala et al, 2024, pg. 9). The scores of all five dimensions were normalized by converting the raw scores for each dimension into a scale to ensure compatibility across different metrics and data samples. Thus, knowledge intensity was calculated by averaging the normalized scores across the five dimensions for each data sample,

resulting in an overall measure for this variable. This follows the original approach and tailors the measurement to the context of contentpreneurs on TikTok (Gala et al, 2024; Morgeson & Humphrey, 2006).

3.2.2.2 Initial performance and performance accumulation rate

The second independent variable, initial performance and performance accumulation rate, required collecting data about followers at two separate dates, for longitudinal purposes. Initial performance was measured by the cumulative number of followers (and alternatively of "likes") at the beginning of the observation period on March 7th (T1). Performance accumulation rate was measured by the cumulative number of followers and likes four weeks later, at the end of the observation period on April 6th (T2), and was calculated as the difference in performance metrics between T2 and T1. The observation time reflects the rate at which a contentpreneur gains additional metrics of success over time. The Gala et al (2024) used a 16-week period to measure performance on Udemy before and after because that is how long it takes a student to complete a course, on average (Udemy, 2023).

This study, however, uses a 4-week period to measure before and after performance on TikTok because of how dynamic and fast-paced the latter platform is in content creation and consumption (Kullolli & Trebicka, 2023). Indeed, in many cases a significant difference in the number of followers was noted in the data collected between the 4-week period. The 16-week timeframe for the Udemy study is appropriate given the level of time and commitment the platform demands from its subscribers. One has to pay to have access to the content, and then must dedicate enough time to have at least partially completed a course before being able to write a review as a verified subscriber. Becoming someone's follower on TikTok, on the other hand, is free and comes with no strings attached. Likewise,"liking" a short video on TikTok takes no time and requires no prior commitment (Ceci, 2023; Duarte, 2024; Kemp, 2023), and thus a shorter timeframe for the replication study is justified.

3.2.3 Control Variables

The following three variables control for differences in contentpreneurs' educational level, professional commitment, and use of social media, as per the original research. The first two are binary values, wherein the mention of an academic degree in a profile's bio, as well as a

link to a personal website is both set to a 1. The third control variable measures how many other social media platforms a contentpreneur is on. In the original study these three control variables were selected as instructor-level variables (Aguinis et al, 2016), on the basis that end users' search selection may be influenced by this kind of information available about the instructors on the platform (Gala et al, 2024, pg. 11).

One extra control variable was added to this study for greater robustness, and that is Activity Level, which is measured by the frequency of videos posted per month, taken as an average. This variable helps to account for the differences in nature and dynamics between TikTok and Udemy specifically in what pertains to the uploading of content. Video content posted on Udemy by an instructor occurs less frequently because it requires significantly more preparation and research per video (Udemy, 2023), whereas TikTok's very trademark are precisely short, informal videos that are usually uploaded much more frequently by content creators. Thus, it became relevant to control for activity level for a platform like TikTok when analyzing metrics of performance.

3.3 Ethical considerations

In handling TikTok's officially approved API, ethical considerations were paramount, and compliance with TikTok's terms of service and data usage policies was ensured to uphold user privacy and data security. In fact, part of the API application included submitting a data protection plan describing the terms of the research data and privacy governance, which can be found in Appendix 1. Throughout the research period the data was only handled by those with permission, there was no sharing of the data with third parties, and data was stored in a computer as well as in a secure hard drive. During the analysis, user data was aggregated to prevent the identification of individuals for privacy protection. All ethical guidelines set forth in the data protection plan were adhered to as closely as possible.

3.4 Analysis strategy and diagnostics

All tests were conducted on Stata, except for the Anderson-Darling tests which were done manually in Excel. On occasion, the aid of an AI tool called ChatCSV would be employed to help confirm the interpretation of results. Multiple tests were used for each hypothesis in order to

compare and strengthen results, as well as to help reduce bias in the findings (Carrol et al, 2006; D'Agostino et al, 1990). While multiple calculations were needed throughout the analysis, the primary statistical analysis focused on tests of correlation, normality, and prediction.

The correlation tests included Pearson and Spearman statistics for H1 (Artusi et al, 2002), and Somer's Delta (Somer, 1962) and Kendall Tau-A for H4 (Brossart et al, 2018; Newson, 2002). These tests are common ways of measuring linear (Pearson) and curvilinear (Spearman, Somer's D, and Kendall Tau) correlations, where when one variable changes the other variable changes in the same direction. The coefficients for these tests are values between -1 and 1, and represent varying levels of strength. Any value greater than 0.5 is generally considered strong for a Pearson correlation, whereas results above 0.7 are considered strong for Spearman, Somer's D, and Kendall Tau correlations (Artursi et al, 2002; Brossart et al, 2018).

Tests of normality included Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) goodness-of-fit tests. Both are powerful tools in detecting deviations from normality, but the A-D test is considered to be more sensitive to deviations in the tails of a distribution than the K-S test (D'Agostino et al, 1990). For both tests a low p-value indicates a rejection of the null hypothesis. For H2 and H3 the K-S and A-D were paired with a Maximum Likelihood Estimation (MLE), not a test of fit per se, but a method for estimating parameters that can be used in conjunction with a goodness-of-fit test to assess how well a distribution fits the observed data (Carrol et al, 2006). Values of skew, kurtosis, mean, and standard deviation were also often noted in observations related to distribution fit.

Lastly, nonparametric multivariate regressions were used as a test of prediction for testing H3. They are a more appropriate test than an OLS for non-normal distributions since it allows the parameters to relax the assumption of a linear relationship between the outcome and the predictor variables (Everitt & Hand, 1981). H3 also required residual procedures through variance partitioning (Banerjee & Yakovenko, 2010), which quantifies the amount of variance not explained by independent variables, and provides valuable diagnostic information in non-normal distribution contexts.

4. Results

4.1 Hypothesis 1 (H1): Non-normal distribution

The first hypothesis measured the non-normality of performance on digital platforms. This was tested using the number of followers for each TikTok account sampled. For H1 a series of tests were used that measure correlation and normality (Artusi et al, 2002). The Pearson correlation, between followers and likes was 0.66, and the Spearman correlation was 0.92. Therefore, there is a strong correlation between the number of followers and the number of likes when it comes to contentpreneurial performance, and it happens to be stronger for TikTok than it is for Udemy. This means that performance on TikTok is strongly correlated with one's number of followers and likes.

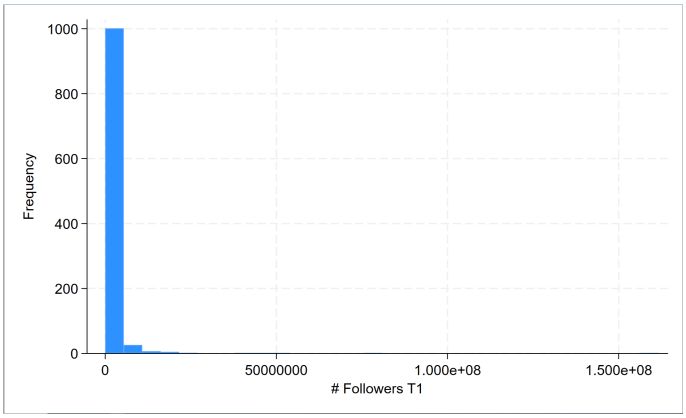
The table below shows some descriptive statistics analyzed for all data samples across all 21 hashtag communities, or domains. The numbers indicate that, just as in the original study, there is greater tail extremity in the number of likes than in the number of followers, which can be seen in the kurtosis and skew values, although the difference is not as extreme as in the original study. The kurtosis especially indicates how heavy the tails of a distribution are compared to a normal distribution (Arnau et al, 2013; Groeneveld & Meeden, 1084). Higher kurtosis and skew values signify more extreme outliers in the performance data, which is indicative of a heavy-tailed distribution.

Performance	Mean	Median	Skew	Kurtosis	SD	Min	Max
No. of followers	1,282,177	128,600	17.3	383.1	6,505,002	13	1.62e+08
No. of likes	3.55e+07	2,600,000	17.8	403.9	2.02e+08	7	5.10e+09

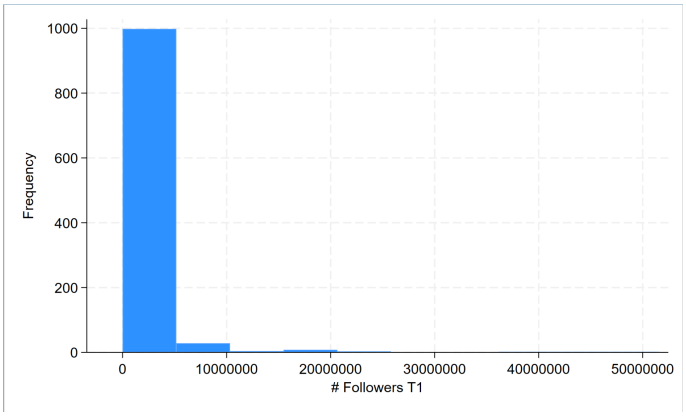
Table 3: Descriptive statistics for performance across all hashtag communities. Note: n=1,046

Figure 3 shows histograms with the number of TikTok contentpreneurs (frequency) plotted against the number of followers per contentpreneur (#Followers T1), across all domains. The histogram is zoomed in six times, with the first panel representing a distribution for the entire range of performance (from 1 to 1,342,657,033 followers in total); the second panel representing a distribution range from 1 to 50,000,000 followers; the third from 1 to 10,000,000;

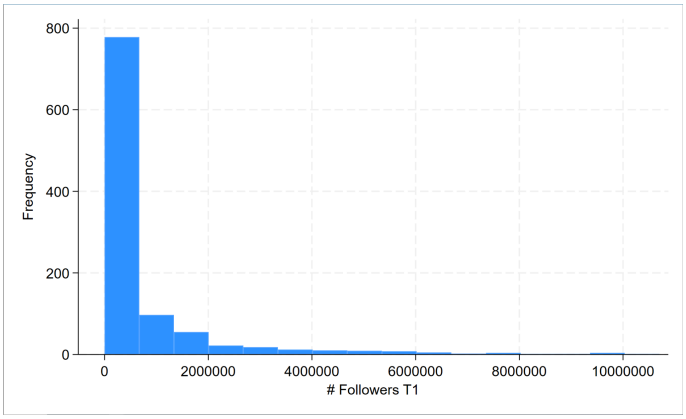
Figure 3: Histograms of performance across hashtag communities



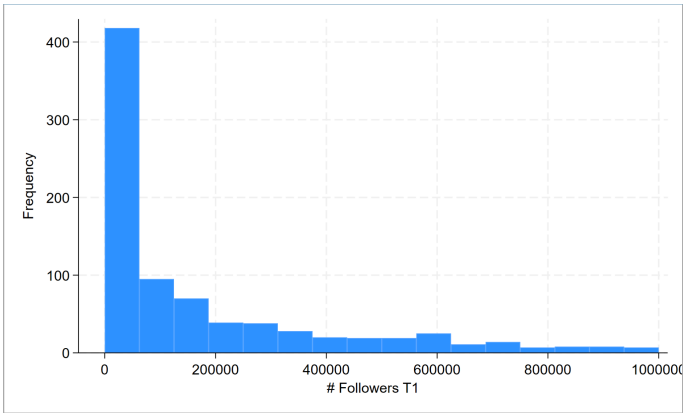
Panel 1: 1 to 1,342,657,033 followers



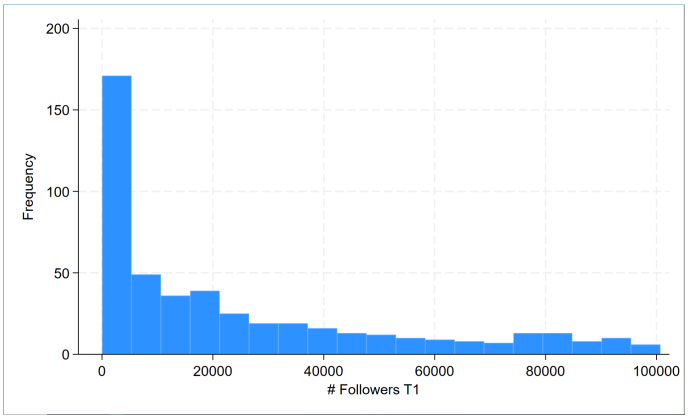
Panel 2: 1 to 50,000,000 followers



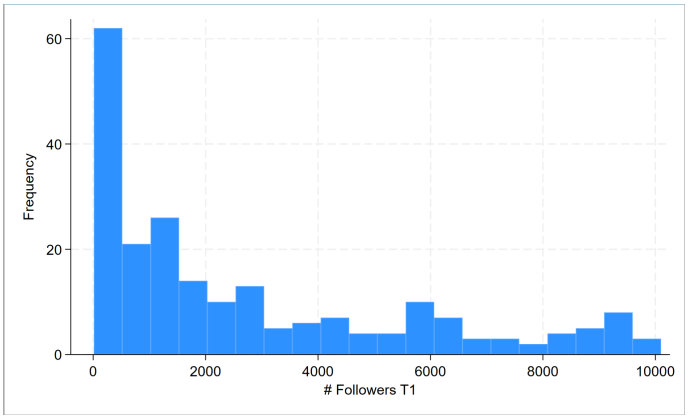
Panel 3: 1 to 10,000,000 followers



Panel 4: 1 to 1,000,000 followers



Panel 5: 1 to 1,000,000 followers



Panel 6: 1 to 10,000 followers

fourth from 1: 1,000,000; fifth 1:1,000,000; and sixth from 1:10,000. This was done to assess whether tail extremity activity can be noted at different ranges, including smaller ranges (Gala et al, 2024, pg. 12). All panels, at varying ranges, display non-normal and heavy-tailed distributions.

The Kolmogorov-Smirnov (K-S) and the Anderson-Darling (A-D) tests confirm the non-normality of the distribution shown in figure 3, resulting in D=0,345 and p=0,000 for the K-S, and A=280,24 and p=1,3882e-60 for the A-D. The extremely low p-value in both cases indicates that the null hypothesis of a normal distribution can be rejected (D'Agostino et al, 1990). Additionally, histograms were plotted for contentpreneurial performance for each specific domain, and a cursory view alone reveals non-normal, heavy-tailed distributions for all 21 communities, which is further confirmed by the respective low p-values of the K-S and A-D tests conducted for each one. All 21 histograms and test values can be found in Appendix 3.

In conclusion, the tests conducted for correlation and normality, as well as the histograms plotted in this section are consistent with H1 of a non-normal and heavy-tailed performance distribution on TikTok. By way of interest, in the domain with the highest follower count (PrankTok) the top 1 content creator has 80.2% of all followers in that community, and in the domain with the lowest follower count (QueerTok) the top 5 content creators have 56% of all followers in that community.

4.2 Hypothesis 2 (H2): Lognormal distributions

For H2 the data was tested for its distribution fit using Joo et al's (2017) taxonomy as a reference (see Figure 1). The aggregate performance of all 1,046 TikTok user accounts was fitted to a lognormal distribution with the parameters scale (μ= 11.4), shape (σ=2.8), with a low log likelihood of -14510.522, which suggest the lognormal fit is not a good fit (Carrol et al, 2006). The K-S goodness-of-fit test was then employed, yielding a D statistic of 0.1245 (above the critical value of 0.042 at a significance level of 0.05, computed as $1.36/\sqrt{1,046}$ for α= 0.05) and a low p-value of 1.31e-14. As the p-value is less than 0.05 the null hypothesis of a lognormal distribution can be rejected. Likewise, a D-value greater than the critical value allows one to reject the null hypothesis that the data follows a lognormal distribution. For further confirmation an added Anderson-Darling test was conducted (D'Agostino et al, 1990) yielding a statistic of A=230.28, which far surpassed the critical values at all significance levels (ranging from 1.008 at

0.5% significance to 0.425 at 25% significance). This also allows the null hypothesis of a lognormal distribution to be rejected, and indicates that the lognormal distribution generally may not be a compatible fit for the pooled data.

Since this deviated from the original study another histogram was plotted for a simple visual inspection, this time with a density plot. At a glance, one might suspect a power law with exponential cutoff fit (Clauset et al, 2009), the closest shape visually to any in Joo et al's (2017) taxonomy. However, testing confirmed that the power law with an exponential cutoff shape is also not a good fit for the plotted data, with a K-S goodness-of-fit statistic of 0.094, which is above the conservative critical value of 0.0418, and a p-value far too small.

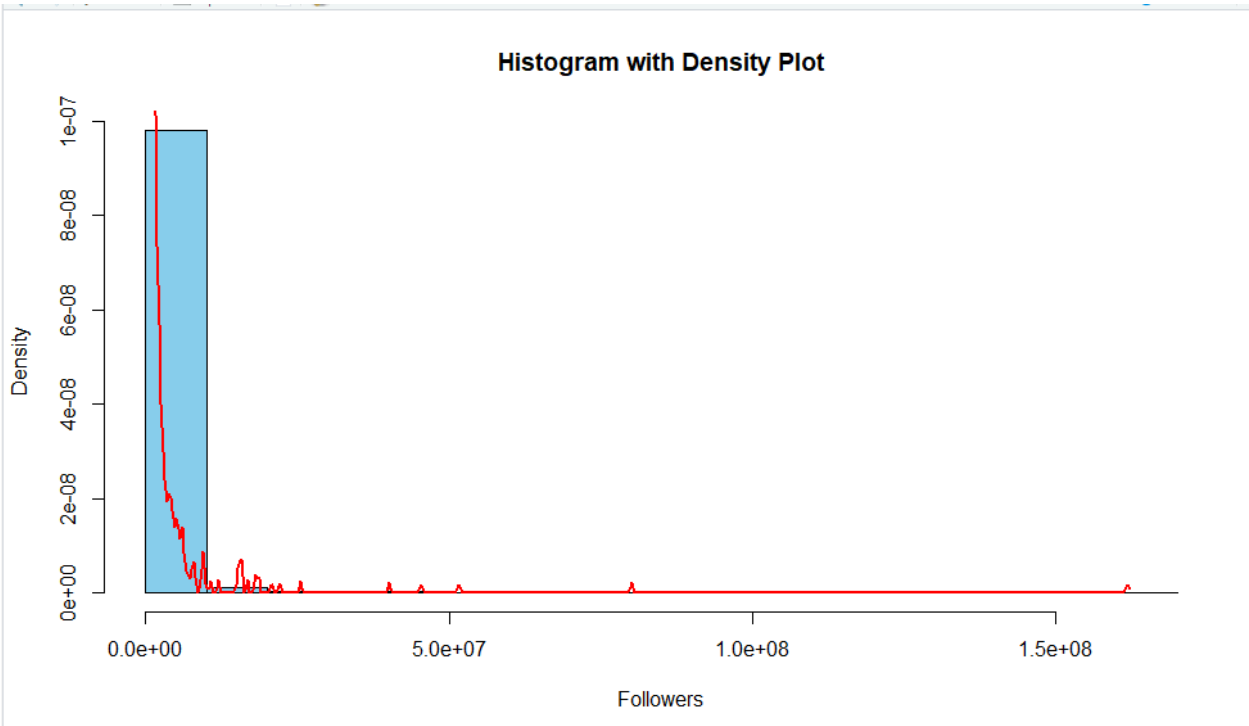


Figure 4: Histogram of performance distribution with density plot.

Finally, a distribution fitting comparing all seven distribution types from figure 1 was done to eliminate the worst fits based on log-likelihood ratios (LR) and p-values (D'Agostino et al, 1990). Interestingly, none of the distributions represented a good fit for the aggregate data (see Appendix 4). All seven had statistics above the critical value as well as low p-values, lacking support for any of them as a dominant distribution fit. Of all the seven distributions, the Weibull distribution computed as, for lack of a better term, the least bad fit, as it had the smallest

K-S statistic and the largest p-value. However, its p-value was still below 0.05, suggesting that the finding is not statistically significant and the data does not perfectly follow a Weibull distribution either. As a result, none of the measures taken provide evidence to support H2.

In accounting for the possible reasons for a lack of a dominant fit for the pooled data, one could consider the nature of the data itself. Despite Joo et al's (2017) assertion that most phenomena can be represented by one of the seven distributions in their taxonomy (pg. 1025), real-world data can often be complex and may not perfectly follow any standard statistical distribution (Beck et al, 2013; Everitt & Hand, 1981; Mitzenmacher, 2004; Newman, 2005). A greater data size, like in the original study, may have provided a clearer picture of performance distribution on TikTok, or at least increased the chances of finding greater compatibility with one of the distribution shapes. But even with a larger dataset statistical analyses can become convoluted (Aguinis et al, 2024), involving multiple interactions among variables and parameters that do not always conform to a simple distribution in a straightforward manner (Beck et al, 2013).

Additionally, outlier numbers on TikTok soar into the millions, on occasion into the billions (Dean, 2023), and a certain level of difficulty with fit can be expected when extreme outliers are present. The sheer magnitude of the gap between the average user accounts and the extreme outliers stretches the tail so far to the right (Hill, 1975) that it can confuse the computation of the fit parameters. Incidentally, when speaking to a statistics PhD student on the test results and the clear lack of a good distribution fit for the data, his advice was to drop the outliers to achieve more manageable numbers that might fit a distribution better—precisely what this study argues against (Andriani & McKelvey, 2009; Clark et al, 2023; Crawford et al, 2015). While it is true that outliers can complicate fit, ultimately the objective of distribution fitting is not to find a perfect fit, which is highly unlikely, but to come as close as possible to a distribution that is compatible to one's sample (Clauset et al, 2009), and no compatibility is also a possible outcome.

Even though at the aggregate level there is no support for H2, distribution fittings conducted for each individual domain yielded different results, as seen in Table 4 below. Following a LR=0 (a null hypothesis of no difference) between each pairwise distribution comparison, and a p-value cutoff of 0.1 (Clauset et al, 2009), the distribution type identified as the dominant distribution representing a specific domain's performance was the one that was

never estimated as the worst fit compared to the other six. If two distributions were equally or inconclusively dominant, Joo et al's (2017, pg. 1034) principle of parsimony was employed to determine dominance based on which one had fewer parameters (refer to table 1 for all parameters). If the same number of parameters were identified, the result constituted a co-dominant pair. None of the domains resulted in a no dominant situation.

The distribution comparisons reveal the lognormal distribution as the dominant distribution for 14 (66.6%) out of the 21 domains. Of the remaining 7, lognormal was co-dominant for 5 (4 were lognormal and power law, and 1 was lognormal and Weibull); 1 had power law and Weibull as co-dominants; and 1 was power law dominant. Thus, out of a total of 21 domains, only two did not have a lognormal distribution as either its dominant or co-dominant distribution, which can be noted in table 4 on page 40. It may appear strange that the pooled distribution of all 1,046 data samples did not follow the fit of any particular distribution, while the individual domains did, but it is not uncommon to see compatibility fit more clearly within a subset of a dataset (Carroll et al, 2006; Everitt & Hand, 1981). In fact less noise and greater heterogeneity of subgroups (Everitt & Hand, 1981) within a dataset often allow for a different or even a tighter distribution fit than that of the pooled data sample they originated from.

Although not as encompassing as in the original study, in which all domains had lognormal as either a dominant or co-dominant fit, these results do provide strong evidence for H2, that contentpreneurial performance tends to be best represented by a lognormal distribution with a heavy tail and with the presence of outstanding performers (Andriani & McKelvey, 2009). Furthermore, the fact that the Poisson and the normal distributions computed as the worst and second worst fit for each domain provides further support for H1. Overall however, given the lack of evidence in the first part of this section, it is reasonable to conclude that these findings provide moderate support for H2.

		Domain count	Percentage
Dominant Distributions	Exponential Tail	0	0
	Lognormal	14	66.6%
	Normal	0	0
	Power Law	1	4.76%
	PL with cutoff	0	0
	Poisson	0	0
	Weibull	0	0
Co-dominant Distributions	Lognormal and PL	4	19%%
	Lognormal and Weibull	1	4.76%
	PL and Weibull	1	4.76%
No dominant or co-dominant distribution	N/A	0	0
Total		21	100%

Table 4: Frequency of performance distribution by domain.

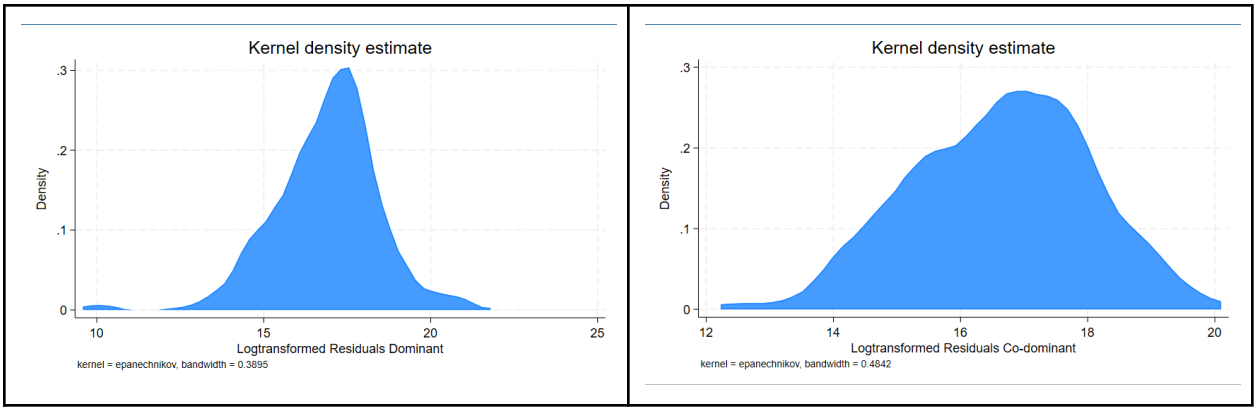


Figure 5: Density plot of all 14 dominant lognormal domains on the top, and of the 4 co-dominant domains on the bottom, logtransformed for improved visualization. Note: the effect of logtransforming a non-normal distribution can skew the original right tail to the left, as seen in the images. This is an attempt at normalizing the distribution.

4.3 Hypothesis 3 (H3): Knowledge intensity

This hypothesis posited a positive correlation between a domain's knowledge intensity and the tail extremity of its performance distribution, with focus on the scale parameter (μ) as one of the dependent variables, while the shape parameter (σ) was used as a control variable. Tests were conducted on 19 of the 21 domains, including only those with either a dominant or co-dominant lognormal fit from H2. Before conducting the tests for H3, all five variables for *knowledge intensity* were normalized and averaged in Stata, and the control variables Degree, Professional, Social, and Frequency were factored in through the residual procedure for variance partitioning (Banerjee & Yakovenko, 2010).

The residuals generated from the variance partitioning procedure were then tested for a lognormal distribution fit. From these 19 residual fittings, the lognormal distribution was the dominant fit for 6 domains, and co-dominant for 8. For these 14 domains with a lognormal distribution of residuals, the scale (μ range: 12.8 to 14.6) and shape (σ range: 1 to 1.76) parameters were estimated, and the Spearman correlation between scale vs shape, shape vs knowledge intensity, and scale vs knowledge intensity were taken. The results can be seen in table 5 below.

	Lognormal shape (σ) parameter	**Knowledge Intensity**
Lognormal scale (μ) parameter	0.4035	-0.1718
Lognormal shape (σ) parameter	–	-0.0396

Table 5: Correlation analysis for knowledge intensity and shape and scale parameters. Note: n=14; *p < 0.05, **p < 0.01

Unlike the original study, the Spearman statistic shows a weak correlation between scale and knowledge intensity. In fact, the negative sign indicates that knowledge intensity, as a predictor of the activity in the tail-end of the lognormal distribution, moves in opposite directions as the scale, and might negatively impact tail-end activity (Artusi et al, 2002). It is worth mentioning that knowledge intensity does have a positive association with the log transformed residuals, particularly at domain level. This is noted at varying degrees of strength within the Spearman statistics between the log transformed residuals and knowledge intensity (range 0.0346 to 0.7792), as well as in the log transformed R^2 value for the 14 domains (range: 0.00 to 0.60). While these ranges do not affect tail-end activity, they tell us that for the 14 domains compatible

with a lognormal distribution fit, knowledge intensity helps account for up to 60% of the unexplained variance of contentpreneurial performance, and that for the same domains knowledge intensity positively correlates with the unexplained variance of contentpreneurial performance within a 0.03-0.77 range, with values ≥ 0.7 considered strong.

In sum, to varying degrees knowledge intensity does account for some of the variance in the outcome of the distribution fit for certain hashtag communities (Banerjee & Yakovenko, 2010), but it does not positively affect specifically the tail-end activity of the distribution fit of all 14 log transformed domains combined (Hill, 1975; Mitzenmacher, 2004). Further, albeit not as relevant, a negative association between shape and knowledge intensity was also found, and a positive but somewhat weak relationship between shape and scale was also found. This tells us that among the 14 domains analyzed, the distribution's shape is a better, albeit weak, predictor of tail-end activity on TikTok than knowledge intensity. It can be concluded that these results do not provide evidence to support H3.

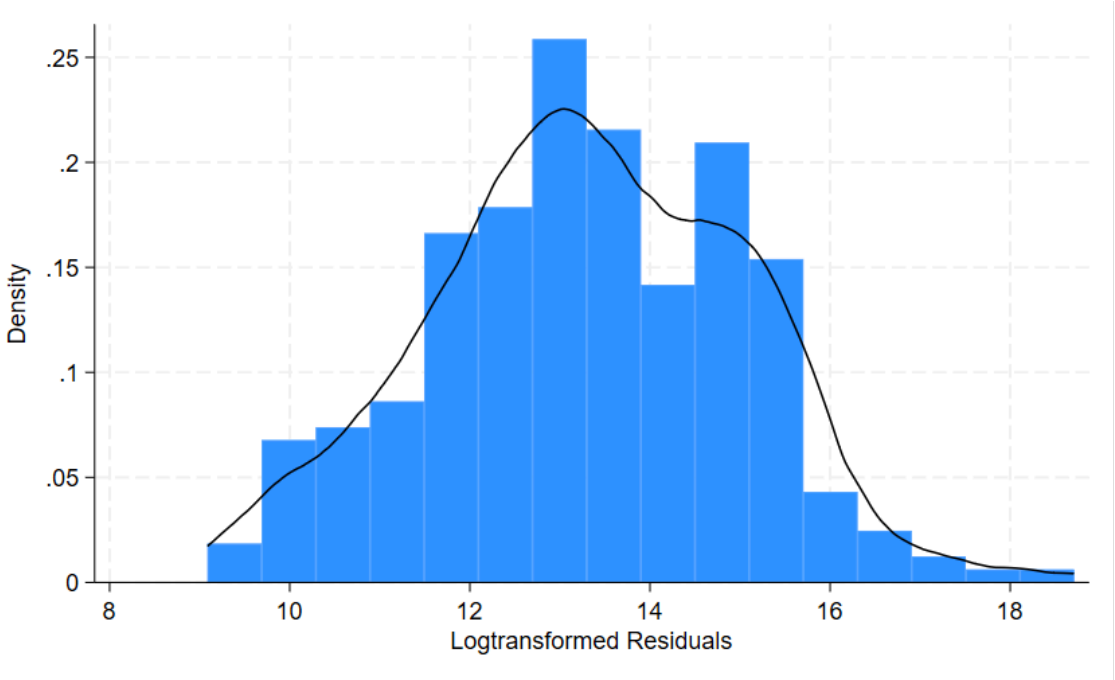


Figure 6: Histogram with kernel density of combined performance distribution for the 14 domains with lognormal residuals. Note the right skewed tail, not as "fat" due to no association with knowledge intensity.

Further variance analyses were conducted to seek added insights into the relationship between knowledge intensity and the distribution's tail end activity. Two nonparametric multivariate regressions (Everitt & Hand, 1981) were run in Stata, one with only the control variables (shape and a combination of the Degree + Professional + Social + Activity level variables, normalized and averaged together as the variable "Contentpreneur") as part of the regression, and a second with the addition of the independent variable of knowledge intensity, to compare the difference between the two. In the second regression the negative KI coefficient and negative t-statistic, coupled with a lower F-statistic, higher Pr(>F) value, and lower Adj. R^2, indicate inconsistency with H3. KI's p-value associated with the t-statistic is also statistically insignificant at all significance levels.

NP regression	Resid. Df	Df	Coefficient	t	P > | t |	F	Pr(>F)	Adj. R^2
σ + Cont.	11	2	σ: 0.808	σ: 0.81	σ: 0.435	2.36	0.140	0.1731
			Cont.: -7.24	Cont.:-1.9	Cont.: 0.081			
σ + Cont+ KI	10	3	σ: 0.65	σ: 0.61	σ: 0.556	1.56	0.2596	0.1144
			Cont.: -6.83	Cont.:-1.72	Cont.: 0.117			
			KI: -4.53	KI: -0.52	KI: 0.614			

Table 6: Nonparametric multivariate regression of lognormal scale parameter (μ). Note: n=14; *p < 0.05, **p < 0.01

4.4 Hypotheses 4a (H4a) and 4b (H4b): Proportional differentiation

The last two hypotheses pertain to the generative mechanism by which the lognormality of this distribution comes to be. Here, hypotheses H4a and H4b, while tested separately, are combined because both must be true to provide evidence that the generative mechanism for this lognormal distribution is proportional differentiation (Mitzenmacher, 2004). This is because both elements are key in making up the main characteristics of proportional differentiation, whereas one without the other can be noted in other generative mechanisms (Banerjee & Yakovenko, 2010; Hill, 1975; Mitzenmacher, 2004). Even though one might expect a generative mechanism to automatically be associated with a specific distribution fit (Joo et al, 2017), variation can be observed (Mitzenmacher, 2004), especially in this case where the distribution shape does not

follow any fit in particular. Thus, support for H4a alone but not for H4b shows evidence of a preferential attachment mechanism, also referred to as the rich-get-richer effect. Similarly, support for H4b but not for H4a reveals incremental differentiation as the predominant generative mechanism (Gala et al, 2024, pg. 15). The table below helps illustrate this concept.

Generative mechanism	**H4a (Initial Value)**	**H4b (Accumulation rate)**
Proportional differentiation	✔	✔
Preferential attachment	✔	**x**
Incremental differentiation	**x**	✔

Table 7: Generative mechanisms pertaining to Hypotheses H4a and H4b

To test H4a one must demonstrate to what extent past performance predicts future performance. The four week longitudinal data about follower count collected at T1 and T2 serve as predictor X (T1) and outcome Y (T2). To test the directional association between these two variables Somer's Delta is used (Crawford et al, 2015; Somer, 1962). H4b tests the extent to which the accumulation rate, which is the difference in follower count from T1 to T2, predicts future performance, and in this case accumulation rate serves as the predictor X while T2 remains the outcome Y. Somer's Delta is also used to test H4b. Lastly, the Spearman rank correlation is computed between T1 and the accumulation rate to ascertain the degree of correlation between the two predictors, revealing a range of 0.02 to 0.34, a low to moderate degree of correlation between the predictors (this range excludes the Spearman correlation of 4 domains that had a negative value). The results of these tests can be seen in table 8.

Something to note in H4b is that in the original study the accumulation rate is defined as the *increase* in Udemy subscribers between T1 and T2 (Gala et al, 2024, pg. 11), because on that platform subscriber count can only go up or stay the same. Once a subscriber pays for and enrolls in a course, the record of their enrollment remains within the platform even after the course is completed or even if the subscriber no longer uses the platform, and additional subscribers get added to that count (Udemy, 2023). On TikTok, however, follower count can increase, stay the same, or actually decrease, as users of the platform can just as easily stop following someone's account (Freehan, 2023). Out of the 1,046 user accounts investigated in this study, 175 had a negative value for their accumulation rate, meaning those users had fewer followers after the

four-week period. Another 286 had no change in follower count at all, and the remaining 585 saw an increase in follower count. For this reason, in this replication study the accumulation rate is defined as the *difference* in follower count between T1 and T2.

Somer's D	**All domains (n=21)**	**Lognormal dominant domains (n=14)**	**Lognormal co-dominant domains (n=5)**
Initial value and Outcome *(avr)* CI 95% (lowest to highest)	0.99 [0.877, 1.005]	0.99 [0.877, 1.005]	0.6 [0.954, 1.004]
Accumulation Rate and Outcome *(avr)* CI 95% (lowest to highest)	0.13 [-0.009, 0.511]	0.37 [-0.009, 0.511]	0.06 [-0.092, 0.397]

Table 8: Somer's D measure across domains (performance= number of followers). Note: CI - 95% Confidence Interval.

Observing the numbers in table 8, it should be noted that no negative values are present in the initial value (number of followers at T1) and outcome (number of followers at T2) combination, and the confidence intervals do not include zero. This eliminates the possibility of there being either a negative correlation or no correlation (Newson, 2002) between the initial value and the outcome, indicating a positive association between the two variables across all 21 domains, in all three categories (n=21, n=14, n=5). The high Somer's D statistic (shown as an average) of 0.98 represents a strong correlation between the two variables, and provides evidence against incremental differentiation as the generative mechanism. These findings show support for H4a.

However, the same cannot be said about the accumulation rate (the difference between follower count from T1 to T2) and the outcome (number of followers at T2) combination. In all three categories of the second Somer's Delta test, the confidence intervals do have negative values, and at some point the intervals do include zero, indicating instances of either a negative or no correlation at all between the variables. This reflects the fact that some user accounts had fewer followers at T2, resulting in a negative value for the accumulation rate. And while there is a positive association between the accumulation rate and the outcome across all domains and in all three categories, the Somer's Delta statistic is low and not statistically significant (Somer,

1962; Newson, 2002). This provides evidence against proportional differentiation as the generative mechanism, and does not support H4b.

Referring back to table 7, support for both H4a and H4b is required to offer proof of proportional differentiation as the distribution's generative mechanism, but in this case only support for H4a was found. Support for H4a but not for H4b indicates that preferential attachment is the generative mechanism for the non-normal distribution of contentpreneurial performance on TikTok. With this generative mechanism, the catch-up effect is essentially null, and it is highly unlikely that someone with a lower initial value will be able to surpass a top performer on this platform (Andriani & McKelvey, 2009; Bak, 1997; Newman, 2005). (It is worth noting that a preferential attachment generative mechanism is often associated with a power law distribution, and the visual resemblance of the pooled data's distribution to a power law type fit, despite there being no statistical evidence supporting a power law fit.)

4.5 Robustness

Additional tests were conducted for each hypothesis to ensure robustness in the results (Carrol et al, 2006). In the original study, H1 and H2, which assess normality, used reviews as an alternative performance variable to the number of subscribed students. In this paper, the alternative permanence measure used is the number of *likes*, which is a form of review. The same tests from H1 were conducted using this new variable, showing a strong Spearman (0.92) and Pearson (0.66) statistics, high skew (17.8) and kurtosis (403.9) values, and low p-values for the K-S (0.00) and A-D tests (1.456e-65), which allow for the null hypothesis of normality to be rejected. Similar findings were found for each individual domain, and support H1. A histogram of the distribution and more information can be found in Appendix 7.

In order to test H2 using "Likes" as the new measure of performance, the aggregate data was fitted to a lognormal distribution on Stata using MLE commands (Carrol et al, 2006), and parameters of scale (μ= 14.3) and shape (σ= 3.21) were estimated. The K-S goodness-of-fit statistic value (0.0347) was below the critical value of 0.0421 (calculated as $1.36/\sqrt{1046}$ for α= 0.05) with a fairly high p-value=0.1745 >0.05. This indicates that a lognormal distribution may be a reasonable fit for the data, since we fail to reject the null hypothesis (D'Agostino et al, 1990). Furthermore, a lognormal fit was dominant for 17 (80.95%) out of 21 domains, and co-dominant for another 3 (14.28%). Only one was found to be more compatible with a different

distribution fit, namely a power law fit (Mitzenmacher, 2004). These results provide evidence in favor of H2. Histograms and domain specific information can be found in Appendix 7.

For H3, the control variable *shape* (σ) was replaced with *kurtosis*, a measure that captures the tail extremity of a performance distribution (Arnau et al, 2013; Groeneveld & Meeden, 1984). After factoring the control variables into the 20 lognormal domains from H2, residuals were generated using the variance partitioning method (Banerjee & Yakovenko, 2010) in Stata and tested for fit. Six out of the 20 domains were found to be lognormal dominant, and 8 co-dominant. The Spearman correlation was taken yielding a positive but weak association between the residual tail-end activity of the combined lognormal dominant and co-dominant domains and knowledge intensity (0.2827). Overall, when investigating "Likes" as the predicted outcome for performance, and when analyzing the association of kurtosis with knowledge intensity, knowledge intensity does play a bigger role than in the tests using followers and scale parameter, but only marginally (see table in Appendix 7).

Further, two OLS regressions were taken of kurtosis, one testing the control variables only, and one adding knowledge intensity. Knowledge intensity was not found to have any statistical significance in predicting kurtosis at any significance level on the aggregate data, and the table with this information can be found in Appendix 7. As a point of interest, though, at domain level knowledge intensity was statistically significant in explaining some of the residual outcome variance for #BookTok (p-value=0.002, **p<0.01) (Ehret et al, 2023), #MusicTok (p-value=0.000, **p<0.01), and #TVTok (p-value= 0.031, *p<0.05), indicating that within those TikTok communities some of the difference between average and top performers can be accounted for by one's knowledge within those fields. Overall, though, the control variable *Activity level* (how often videos were posted on a TikTok account in a month) computed as being more statistically significant (Dayan & Tafesse, 2023) in explaining performance outcome than all other predictors and controls throughout all the testing. In sum, some support for H3 is present in these results, but still rather negligible.

Kendall Tau-A was used instead of Somer's D to test H4a and H4b to estimate ordinal association (Brossart et al, 2018; Newson, 2002). The number of likes per account was collected longitudinally on the same dates as the number of followers, and were labeled as Likes1 and Likes2. For this test, Likes1 was used as the alternate measure of initial value, or predictor (X), and Likes2 was used as the outcome (Y). The second round of Kendall Tau-A used the

Accumulation Rate as the predictor, and Likes2 as the outcome. In collecting the longitudinal data for number of likes, it was observed that negative values were also present, meaning, a user could "retract" their like, but the amount of negative values was far fewer compared to the negative values found in the accumulation rate between followers at T1 and T2, and the increase percentage of likes during the same time period was overall higher than that for followers.

The test was run on the aggregate, lognormal dominant, and lognormal co-dominant levels. No negative values within any of the confidence intervals were found, eliminating the possibility of no association, or of an inverse association between the variables. Table A7.4 in Appendix 7 shows a strong correlation between initial value and outcome, with an average value of 0.946, which provides evidence against incremental differentiation as the generative mechanism and support for H4a. A moderate correlation between accumulation rate and outcome is also noted, with an average value of 0.41, providing evidence against preferential attachment as the generative mechanism and support for H4b. All together, there is support for proportional differentiation as the generative mechanism (Mitzenmacher, 2004) when number of likes is used as a measure of performance.

5. Discussion

By way of reminder, Gala et al's (2017) "Star Entrepreneurs" analysis on Udemy posited that entrepreneurial performance distributions on digital platforms are more likely to be non-normal (H1); that the dominant type of non-normal performance distributions on digital platforms is likely to be lognormal (H2); that there is a positive correlation between knowledge intensity and the tail end of the distribution (H3); and that proportional differentiation is the generative mechanism that brings about the lognormal distribution on digital platforms (H4a and H4b) (pgs. 3-8). In this replication study, which tests the same hypotheses but uses data from TikTok for the analysis, strong evidence is found to support H1; moderate evidence is found to support H2; weak evidence is found to support H3; and strong evidence is found to support H4a but not H4b. The robustness tests, which use alternative variables (Beck et al, 2013) for performance and tail end activity, provide additional support for H2 and H4b. Overall, similar results from the original paper were partially replicated by the analysis undertaken in this study, and thus partially strengthen Gala et al's (2017) postulates about performance distribution on digital platforms.

Having tested the hypotheses, a moment to synthesize the results and what they mean in a real life contentpreneurial context is warranted. The advent of content creation (Dayan & Tafesse, 2023; Escolano, 2023) and its monetization (Gustafsson & Khan, 2017) on social media platforms in recent years has capitalized on the unique attributes of digitization (Chatradhi et al, 2023; Yoo et al, 2012), which include digital artifacts at nearly zero marginal costs of production and distribution (Breyer et al, 2019), and almost boundless physical constraints (Campos et al 2023; Nambisan, 2017) in terms of market outreach. The ease of digital content creation (Sahut et al, 2019) has served as a springboard to propel the online careers (Li et al, 2024), ventures, and personas (Ashman et al, 2018) of many contentpreneurs at a relatively rapid pace (Frenkel, 2021; Johnson et al, 2022), enabling a certain propensity for star performers on content creating platforms (Cutolo & Kenney, 2021; Gala et al, 2024). The presence of these high performing outliers is the element that generates non-normal performance distributions with a heavy tail (Aguinis et al, 2024) thus, test results of non-normality for H1 essentially demonstrate that there are indeed star contentpreneurs on TikTok.

Social media platforms like TikTok have their limitations (Dean, 2023), however, and the prevalence of star contentpreneurs is ultimately circumscribed by a finite number of users on the platform. While a heavy but finite tail should indicate a lognormal fit (Joo et al, 2017; Mitzenmacher, 2004), ascertaining the best distribution fit on the TikTok dataset was not as straightforward. Even though the lognormal distribution is *generally* the best fit for the performance distribution of contentpreneurs on TikTok, as noted within the smaller community subsets, the pooled data sample collected for this analysis may not be representative enough (Everitt & Hand, 1981) of that "big picture" distribution fit. Despite this, it is reasonable to conclude that the test results for H2 indicate a high likelihood of a relatively few outstanding and highly distinct star contentpreneurs on TikTok (Clark et al, 2023; Freehan, 2023) whose prevalence drops off dramatically when the most extreme number of followers on the platform are reached (Joo et al, 2017).

In addition to the number of followers (and likes), surely other factors help predict the outcome of star contentpreneurs' performance. In the original piece for Udemy one such factor was knowledge intensity, a five-dimensional variable consisting of elements of complexity, information processing, problem-solving, skill variety, and specialization (Morgenson & Humphrey, 2006; Gala et al, 2024, pg. 9). The TikTok analysis, however, shows knowledge

intensity as not playing a significant role on how well top performers do on the platform (Bachmann et al, 2024), even though to a certain extent it does help explain some of the variance in outcome at domain level, and might be of greater contribution within communities (Conrad, 2024) with a more narrowed focus. Overall, for H3 knowledge intensity operates quite differently on TikTok than it does on Udemy, a finding that is consistent with the informal nature (Ceci, 2023; Kemp, 2023) of the platform.

Lastly, the mechanism behind this competitive environment is not as straightforward for TikTok as it is for Udemy, because it depends on how performance is being measured. In the original study, longitudinal data of both measures of performance established that Udemy's generative mechanism is proportional differentiation (Gala et al, 2024, pg. 8), where initial performance hurdles can be overcome over time by the means of a multiplicative effect akin to interest rate (Mitzenmacher, 2004). A proportional differentiation generative mechanism enables late entrants or contentpreneurs with low initial value to succeed and eventually surpass top performers (Aguinis et al, 2016; Joo et al, 2017).

Longitudinal data for this analysis, however, demonstrates that if the number of followers is used as the measure of performance being analyzed, TikTok's generative mechanism is preferential attachment (Bak, 1996; Mitzenmacher, 2004) in which success breeds success, the rich get richer, and those at the top will likely not be surpassed regardless of time and effort (Andriani & Mckelvey, 2009; Gala et al, 2024; Joo et al, 2017). But if using the number of likes for performance instead of followers, proportional differentiation is found to be the generative mechanism (Beck et al, 2013). Considering that the number of followers is a much weightier measure (Dayan & Tafesse, 2023; TikTok, 2024) if monetization of one's account is the ultimate goal (Campos et al, 2023; Gagliargi, 2024), on a practical level the number of followers likely prevails as a better measure of performance, establishing the less forgiving (Mitzenmacher, 2004) preferential attachment as the main generative mechanism for TikTok's performance distribution.

5.1 Implications for theory

This conceptual replication provides some empirical support for Gala et al's (2024) findings linking performance on digital platforms with non-normal distributions. It strengthens our understanding of the prevalence of top contentpreneurs who outperform the rest (Aguinis et

al, 2012; Cascio et al, 2022) and create a heavy tail in the distribution (Aguinis et al, 2024; Joo et al, 2017; Newman, 2005). Far from a passing interest, these findings send a message to scholars of the repercussions of ignoring outliers (Andriani & McKelvey, 2009) in entrepreneurial performance data, and of failing to normalize heavy-tailed distributions in this field (Clark et al, 2023). Doing so can severely limit our understanding of the nature of content creating platforms as enabling tools in entrepreneurial success (Escolano, 2023; Li et al, 2024; Rooney, 2020). Focusing most theory on average performance and normal distributions (Crawford et al, 2015) can also hinder scholars' ability to estimate variance in success (Carrol et al, 2006) and to predict its likelihood within these environments. Thus, this study extends the original study's call to "revisit long-held assumptions of normal distribution of performance" (Gala et al, 2024, pg. 18) in entrepreneurial research (Clark et al, 2023).

This paper also narrows the scope of Gala et al's (2024) conversation on digital entrepreneurship by focusing more specifically on social media platforms that enable contentpreneurship (Frenkel, 2021; Johnson et al, 2022)—a nascent field within digital entrepreneurship that is here to stay. Particularly prevalent in the lives of gen Z'ers (Chiang & Jang, 2023; Escolano, 2023; Kullolli & Trebicka, 2023; Kushwaha, 2021), who are now starting to contribute to the world economy (Breyer et al, 2019; Hazari & Sethna, 2022; Ibrar et al, 2022), contentpreneurship as a topic of study is ripe for further research. A better understanding of contentpreneurial performance distribution lays the foundation for additional theory development on topics that not only shape, inform, and impact contentpreneurial success (Johnson et al, 2022), but intersect with other streams of literature for original and robust research (Anderson et al, 2019). These might include an investigation of emerging technologies (Ebrahimi et al, 2023), such as AI-driven content creation tools (Short & Short, 2023) and blockchain-based monetization platforms (Ibrar et al, 2022), or ethical implications of contentpreneurship, including issues of transparency, authenticity, and user privacy (Breyer, 2019); etc. Understanding these dynamics becomes increasingly relevant for policymakers, educators, and industry stakeholders seeking to support the next generation of entrepreneurs (Breyer et al, 2019) and foster sustainable economic growth in an age where content creation on social media platforms has become a daily occurrence (Kullolli & Trebicka, 2023).

For social media platforms with such widespread followers like TikTok, the results from this paper illustrate the importance of working with more comprehensive sample sizes to draw

better conclusions about distribution fit (Aguinis et al, 2012; Booyavi & Crawford, 2023) and generative mechanisms. A sample size like that of the original paper's, for instance, can provide a clearer picture of distribution and mechanism. Nonetheless, the data available for this paper, albeit limited, still demonstrates that performance on TikTok at domain level is best characterized by a lognormal distribution, a corroboration of Gala et al's (2024) insights and a deviation from some of the more commonly found distributions in work and organizational research (such as Poisson, exponential tail, and power law with exponential cutoff (Aguinis et al, 2014)). Two new insights that stray from the original paper's findings are also revealed by the data, namely that a) the element of knowledge intensity does not play the same kind of role on TikTok's distribution's tail end activity as it does on Udemy's, and b) the generative mechanism for TikTok's performance distribution, if follower count is used as the measure of performance, is not proportional differentiation, where time and persistence can eventually overcome initial disadvantages, but rather preferential attachment (Andriani & McKelvey, 2009; Bak, 1996; Mitzenmacher, 2004), which makes leveling out the playing field quite difficult once top performers are entrenched on the platform.

Some important insights can be gleaned from these findings. First, the elements of this study that deviated from the results from the previous study demonstrate the need to consider the nature of the digital platform (Hazari & Sethna, 2022) being analyzed when generalizing assumptions in entrepreneurial research. Gala et al's (2024) hypotheses about entrepreneurial performance on digital platforms appear not to take into consideration the dynamics of social media platforms (Rooney, 2020) more specifically, where entrepreneurial activity does take place but the informal nature of the platform does not always necessitate specialized knowledge to ensure success (Bachman et al, 2024). One of the highest performing contentpreneurs on TikTok, for instance, has built a following of 162 million people with content consisting of short funny videos with no talking in them.

Similarly, the evolution of social media platforms and how their audience evolves with them (Kullolli & Trebicka, 2023) must also be taken into account when investigating elements of digital performance. Whereas Udemy was born as a paid educational platform for an educated and adult population, TikTok began as a free and relatively obscure platform for entertainment purposes (Cutolo & Grimaldi, 2023) for a younger audience (Hazari & Sethna, 2022); surged in popularity during the pandemic (Crespo et al, 2024; Dornekott et al, 2021; Garvey et al, 2023) as

a community builder; then ultimately evolved into a space where contentpreneurship can flourish (Escolano, 2023) through the monetization of self-promotion (Ashman et al, 2018; Dayan & Tafesse, 2023; Tang et al, 2023). Today, the sheer magnitude of the gap between average and top contentpreneurs on TikTok not only helps to explain the dataset's general incompatibility with any given distribution fit (Beck, 2013; Clauset et al, 2009; Everitt & Hand, 1981; Mitzenmacher, 2004; Newman, 2005), but is also a reflection of the platform's evolution over time (Kullolli & Trebicka, 2023), which allowed early entrants to eventually profit (Donaker et al, 2019) from a large, active, and well established audience. Thus, stemming from the discrepancies between the original study and its replication, this thesis posits that elements of a platform's nature, purpose, audience, and evolution must be considered when developing theories of digital entrepreneurship.

Insights can also be drawn from the similarities between both studies. The non-normality and lognormality of both datasets highlight the need to adapt current quantitative methodologies in entrepreneurial research to incorporate non-normal paradigms within scholarly guidelines of statistical analyses (Andriani & McKelvey, 2009; Booyavi & Crawford, 2023; Cascio et al, 2022; Clark et al, 20023). This can be done by embracing the pervasiveness of non-normality within entrepreneurial variables (Andriani & McKelvey; 2009; Aguinis et al, 2024), and streamlining the process of non-normal data analysis within academic literature (Joo et al, 2024). Specifically, to properly assess non-normal data certain steps must be taken to identify the best distribution fit for the data, assess the impact of its outliers on the whole of the distribution, and test hypotheses all while employing the analytical methods that are most sensitive to non-normal data (Joo et al, 2024). This study calls for this approach to become standardized procedure in quantitative studies in this field. Doing so can not only bolster the research on the taxonomy of distribution shapes, leading to new generative mechanisms (Newman, 2005), but can also enrich the field of entrepreneurship by expanding its realm of interdisciplinary crossover with stringent and replicable quantitative research (Anderson et al, 2019).

5.2 Implications for practice

For those that are either in or are considering entering the world of contentpreneurship, the high variance in performance (Campos et al, 2023) on a social media platform such as TikTok, as revealed in this study, should be noted in order to temper expectations (Gagliardi,

2024; Hendricks et al, 2023; Nambisan, 2017). The extremeness of TikTok's tail end activity, coupled with its preferential attachment (rich-get-richer) generative mechanism, render it highly unlikely to reach influencer status (Chen et al, 2020) even after a sustained period of time, especially for a late entrant (Newman, 2005). Far from discouraging participation, this admonishment should be taken as an invitation to reframe what success might mean (Ashman et al, 2018) for one's specific venture. If monetization from the platform itself seems unrealistic even after some dedicated time (Ashman et al, 2018; Cutolo & Grimaldi, 2023; Dornekott et al, 2021), using these types of platforms mainly as an audience builder can still prove to be an efficient tool in product or service promotion, brand differentiation, or persona crafting (Ashman et al, 2018; Tang et al, 2023), especially considering the market outreach potential online compared to that of the physical world (Nambisan, 2017; Srinivasan & Venkatraman, 2017; Yoo et al, 2012). This study's findings about TikTok's tail-end activity can thus assist contentpreneurs' entry or scale-up strategies with a better informed outlook of their odds of success in a deeply competitive landscape (Li et al, 2024).

While it might be hard to compete at the aggregate level, the odds of succeeding can increase at domain level. This study's analysis reveals that the performance variance is usually smaller within a community, highlighting the significance of a more narrowed and defined scope (Conrad, 2024; Ehret et al, 2023) when interacting on the platform for increased chances of growth. Essentially, the numbers tell us to pick a niche. This might also be relevant when it comes to the element of knowledge intensity. The findings show that knowledge intensity as a predictor does not have a big effect (and may in fact have a negative effect) on the tail end of the distribution, but it plays a role in explaining some of the residual variance (Dornekott et al, 2021; Morgeson & Humphrey, 2006) in the overall shape of the distribution. This means that greater knowledge intensity will not improve the status for the star contentpreneurs already entrenched at the top (Dean, 2023; Freehan, 2024), but may well help to improve the odds of success for an individual or an organization at the beginning or middle portions of their TikTok experience.

Lastly, the findings in this analysis indicate a certain level of residual variance in outcome that remains unexplained, which means that there must be exogenous factors (Andriani & McKelvey, 2009) not analyzed by the study that affect performance on TikTok. It is possible that some of these factors include resource level, account type, and celebrity status. Unlike on Udemy, where every entrepreneur is an instructor, TikTok accounts vary greatly. Netflix, for

example, has a TikTok account of 38.5 million followers and easily posts several times a day, because their content is ready to go. ESPN's account boasts 45.6 million followers, and posts an average of 1,500 videos a month from their already existing archive of sports footage. In addition, out of the top 50 TikTok accounts, 15 belong to world renowned celebrities (Dean, 2023). For a platform like TikTok, these kinds of factors may serve as better predictors of the unexplained portion of performance outcome than some of the variables analyzed in the study.

The average contentpreneur may not be able to compete with those levels of resources or status, but ultimately they may not have to. The purpose of studying star contentrepreneurs and their performance distribution has less to do with prescribing how to become one (Aguinis & O'Boyle, 2014), and more to do with informing participants about the realities of co-existing alongside them (Hendricks et al, 2023; Kim & Makadok, 2022) and how this may impact their own performance. The practical insights provided by the results of this analysis can help contentpreneurs on TikTok approach their activity on the platform more realistically in order to bolster their performance. They can also proactively identify up and coming platforms relevant to their brand where star entrepreneurs may not be as entrenched yet (Donaker et al, 2019), to exploit early entry advantages and leverage feedback loops and network effects.

5.3 Limitations of the study

The theoretical and practical implications of this study highlight the ways in which it is constrained by certain limitations, also of a theoretical and practical nature. As a replication piece, it follows certain parameters imposed by the structure and design of the original study. Since only one digital platform was analyzed in Gala et al's (2024) piece, this study followed suit and likewise focused on one platform only. Other studies on star performers, however, collect data from several organizations from various fields and backgrounds (Aguinis & O'Boyle, 2012; Aguinis et al, 2024; Booyavi & Crawford, 2023; Clauset et al, 2009; Crawford et al, 2015; Joo et al, 2017) and perform a cross-sectional analysis of multiple datasets all at once. Not only does a broader scope contribute to generalizability (Clauset et al, 2009), but it also increases the chances of exploring new distribution shapes (Joo et al, 2017), or perhaps a combination of distribution shapes. This, in turn, also leaves open the door to either new or combined generative mechanisms (Andriani & McKelvey, 2009; Newman, 2005), with associated implications for

further theory development. Future research on star contentpreneurs could compare data from a number of different digital platforms, replicating some components of this study, but perhaps also including exogenous factors such as new technology, cultural contexts, business model sustainability, etc.

A lack of exogenous factors is in fact a limitation in this study in particular. TikTok grew in popularity precisely because of a worldwide exogenous factor—the 2020 pandemic (Crespo et al, 2024; Garvey et al, 2023). The phenomenon of sudden content virality, celebrity accounts, and the popularity of "trends" are other examples of shock factors (Andriani & McKelvey, 2009, pg. 1058; Joo et al, 2017, pg. 1025) notable to this social media app that one might expect would impact performance and tail-end extremity (arguably more accurately than some of the variables in this study, selected so as to replicate the variables in the original paper). Even though the study's objective entailed testing the generalizability of findings pertaining to entrepreneurial performance on digital platforms, performance assessment on TikTok in particular could have been improved with more relevant variables (Beck et al, 2013).

Lastly, one of the more practical limitations of this study came about during the data collection stage of the analysis. Despite obtaining official permission from TikTok to use its research API, the data the API generates is considerably constrained (Taboada-Villamarín, 2023), both in nature and in amount. These constraints simultaneously increased the amount of time required to collect the data samples, while also limiting the sample size that was ultimately obtained for the study. A more comprehensive dataset to conduct the tests on would have likely yielded more definitive conclusions, particularly to H2.

6. Conclusion

This study set out to conceptually replicate the findings from Gala et al's 2024 "Star Entrepreneurs" on heavy-tailed performance distributions on digital platforms. It did so by analyzing randomly selected data from TikTok, a content creating social media platform where contentpreneurs can monetize (Cutolo & Grimaldi, 2023) their content usually once a big enough following (Dayan & Tafesse, 2023; Tang et al, 2023) has been established. The statistical analysis of the data reveals that the performance distribution on TikTok is not normal (Aguinis et al, 2024; Joo et al, 2017), and that at a subset level its performance distribution is predominantly lognormal (Gala et al, 2024; Mitzenmacher, 2004). This confirms the original study's hypotheses

that performance on digital platforms tend to be the most compatible with the lognormal distribution fit. Further, the results show no indication that knowledge intensity positively affects how well top TikTok contentpreneurs do on the platform (Dornekott et al, 2021), a significant deviation from the original study in which one's degree of knowledge and expertise in their field played an important role in top performance. Lastly, the study also found that once entrenched, the likelihood of matching or surpassing a star contentpreneur on TikTok is extremely low (Bak, 1996; Mitzenmacher, 2004), unlike Udemy, in which proportional differentiation acts as a mechanism that enables a "catch-up" effect over time.

In sum, these results partially support the findings from the original study it replicated, and offer new insights into the results that did not match the results from Gala et al's (2024) paper. In doing so it advances empirical research within the fields of entrepreneurship (Anderson et al, 2019; Shane & Venkataraman, 2000), organizational theory of non-normal distributions (Andriani & McKelvey, 2009; Clark et al, 2023), and social media platforms. This thesis posits that top performers that represent the outliers in a sample ought to be investigated more carefully and not discarded from the data (Aguinis & O'Boyle, 2012; Andriani & McKelvey, 2009), as has been the conventional practice in the field of entrepreneurship. Failing to do so can and will limit our collective understanding of the catalysts for outstanding performance (Booyavi & Crawford, 2023), as well as our explanatory capacity and forecasting reliability, specifically within the realm of digital platforms (Ibrar et al, 2022; Kullolli & Trebicka, 2023) and contentpreneurship. This replication also cautions against the generalization of assumptions regarding non-normal distributions on digital platforms (Beck et al, 2013; Joo et al, 2017), as the nature of a platform and its distinctive characteristics (Chiang & Jang, 2023; Rooney, 2020) may yield unique findings pertaining to performance distributions and generative mechanisms (Mitzenmacher, 2004) that must be taken into account when developing theories of digital entrepreneurship. Lastly, the findings in this paper introduce possibilities for future research, and motivate scholars to expand the conversation on the important topic of contentpreneurship.

References

Aguinis, H. & O'Boyle, E., Jr., & (2012). The Best and the Rest: Revisiting the norm of normality of individual performance. *Personnel Psychology,* vol. 65, no. 1, pp. 79-119, https://doi.org/10.1111/j.1744-6570.2011.01239.x

Aguinis, H., & O'Boyle Jr., E. (2014). Star performers in twenty-first century organizations. *Personnel Psychology,* vol. 67, no. 2, pp. 313-350, https://doi.org/10.1111/peps.12054

Aguinis, H., O'Boyle, E., Jr., Gonzalez-Mulé, E., Joo, H. (2016). Cumulative Advantage: Conductors and Insulators of Heavy-Tailed Productivity Distributions and Productivity Stars. *Personnel Psychology*, vol. 69, no. 1, pp. 3-66, https://doi.org/10.1111/peps.12095

Anderson, B. S., Wennberg, K., & McMullen, J. S. (2019). Enhancing quantitative theory-testing entrepreneurship research. *Journal of Business Venture*, vol. 34, no. 5, p. 105928, https://doi.org/10.1016/j.jbusvent.2019.02.001

Andriani, P. & McKelvey, B. (2009). Perspective—From Gaussian to Paretian Thinking: Causes and Implications of Power Laws in Organizations, *Organization Science*, vol. 20, no. 6 pp. 1053-1071, https://doi.org/10.1287/orsc.1090.0481

Arnau, J., Blanca, M. J., López-Montiel, D., Bono, R., and Bendayan, R. (2013). Skewness and Kurtosis in Real Data Samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, vol. 9, no. 2, pp.78-84, doi:10.1027/1614-2241/a000057

Artusi, R., Verderio, P., Marubini, E. (2002). Bravais-Pearson and Spearman Correlation Coefficients: Meaning, test of hypothesis and confidence interval. *The International Journal of Biological Markers*, vol. 17, no. 2, pp. 148-151, doi:10.1177/172460080201700213

Ashman, R., Patterson, A., & Brown, S. (2018). 'Don't forget to like, share and subscribe': digital autopreneurs in a neoliberal world. *Journal of Business Research,* vol. 92, pp. 474-483, https://doi.org/10.1016/j.jbusres.2018.07.055

Bachmann, N., Hölzle, K., Maul, V. & Rose, R. (2024). What makes for future entrepreneurs? The role of digital competencies for entrepreneurial intention. *Journal of Business Research,* vol. 174, p. 114481, https://doi.org/10.1016/j.jbusres.2023.114481

Bak, P. (1996). How nature works: The science of self-organized criticality. New York, NY: Copernicus.

Banerjee, A., & Yakovenko, V. M. (2010). Universal patterns of inequality. *New Journal of Physics,* vol. 12, p. 075032, doi: 10.1088/1367-2630/12/7/075032

Baron, R.A. (2006). Opportunity Recognition as Pattern Recognition: How entrepreneurs "connect the dots" to identify new business opportunities. *Academy of Management Perspectives*, vol. 20, no. 1, pp. 104-119

Beck, J. W., Beatty, A. S., & Sackett, P. R. (2013). On the distribution of job performance: The role of measurement characteristics in observed departures from normality. *Personnel Psychology,* vol. 67, pp. 531-566

Breyer, Y., Dumay, J. & Zaheer, H. (2019). Digital entrepreneurship: An interdisciplinary structured literature review and research agenda. *Technological Forecasting and Social Change*, vol. 148, p. 119735, https://doi.org/10.1016/j.techfore.2019.119735

Booyavi, Z. & Crawford, G.C. (2023). Different, but same: A power law perspective on how rock star female entrepreneurs reconceptualize "gender equality." *Journal of Business Venturing Insights*, vol. 19, E00374, https://doi.org/10.1016/j.jbvi.2023.e00374

Brossart, D. F., Laird, V. C., Armstrong, T. W., & Walla, P. (2018). Interpreting Kendall's Tau and Tau-*U* for single-case experimental designs. *Cogent Psychology*, vol. 5, no. 1, https://doi.org/10.1080/23311908.2018.1518687

Campos, R.D., Chimenti, P., & Da Fonseca, A.L.A. (2023). 'Take my advice': Entrepreneurial consumers and the ecosystemic logics of digital platforms. *Technological Forecasting and Social Change*, vol. 193, p.122601, https://doi.org/10.1016/j.techfore.2023.122601

Carrol, R. J., Crainiceanu, C. M., Ruppert, D., Stefanski, L. A. (2006). Measurement Error in Nonlinear Models. A Modern Perspective, 2nd edn, New York: Chapman and Hall/CRC, https://doi.org/10.1201/9781420010138

Cascio, W. F., Collings, D. G., Kehoe, R. R. (2022). Simply the best? Star performers and high-potential employees: Critical reflections and a path forward for research and practice. *Personnel Psychology,* vol. 76, no. 2, https://doi.org/10.1111/peps.12558

Ceci, Laura (2023). Average TikTok video length from March 2023 to August 2023, by number of video views. Statista. https://www.statista.com/statistics/1372569/tiktok-video-duration-by-number-of-views/ [Accessed 9 February 2024]

Chatradhi, N., Jha, S., Paliwal, M., & Tripathy, S. (2023). Growth of Digital Entrepreneurship in Academic Literature: A Bibliometric Analysis. *International Journal of Sustainable Development and Planning,* vol. 18, no. 6, pp. 1929-1942, https://doi.org/10.18280/ijsdp.180629

Chen, T. Y., Lee, F. Y., Yeh, T. L. (2020). The impact of Internet celebrity characteristics on followers' impulse purchase behavior: the mediation of attachment and parasocial interaction. *Journal of Research in Interactive Marketing*, vol. 15, no. 3, pp. 483-501, doi:10.1108/JRIM-09-2020-0183

Chiang, IT. & Jang, YT. (2023). Incorporating desire and persistence into understanding Gen Z learners' continuance intention toward using Youtube for learning in digital learning context. *Educ Inf Technol*, https://doi.org/10.1007/s10639-023-12202-9

Clark, D.R., Crawford, G.C. & Pidduck, R.J. (2023). Exceptionality in entrepreneurship: Systematically investigating outlier outcomes. *Journal of Business Venturing Insights*, vol. 20, p. e00422, https://doi.org/10.1016/j.jbvi.2023.e00422

Clauset, A., Shalizi, C. R. & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review,* vol. 51, pp. 661-703

Conrad, Sarah (2024). A complete guide to TikTok content categories. GRIN. https://grin.co/blog/complete-guide-to-tiktok-content-categories/ [Accessed 8 February 2024]

Crawford, G. C., Aguinis, H., Lichtenstein, B., Davidsson, P., & McKelvey, B.. (2015). Power law distributions in entrepreneurship: Implications for theory and research', *Journal of Business Venturing*, vol. 30, no. 5, pp. 696–713, https://doi.org/10.1016/j.jbusvent.2015.01.001

Crespo, N.F., Crespo, C.F., & Silva, G.M. (2024). Every cloud has a silver lining: The role of business digitalization and early internationalization strategies to overcome cloudy times. *Technological Forecasting and Social Change*, vol. 200, p. 123084, https://doi.org/10.1016/j.techfore.2023.123084

Cutolo, D., & Grimaldi, R. (2023). I wasn't expecting that: How engaging with digital platforms can turn leisure passion into entrepreneurial aspirations. *Journal of Business Venturing Insights,* vol. 20, p. e00404, https://doi.org/10.1016/j.jbvi.2023.e00404

Cutolo, D., & Kenney, M. (2021). Platform-dependent entrepreneurs: Power asymmetries, risks, and strategies in the platform economy. *Academy of Management Perspectives,* vol. 35, no. 4, pp. 584-605, https://doi.org/10.5465/amp.2019.0103

D'Agostino, R. B., Belanger, A., and D'Agostino, R. B. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. *The American Statistician*, vol. 44, no. 4, pp. 316-321, https://doi.org/10.1080/00031305.1990.10475751

Dayan, M. & Tafesse, W. (2023). Content creators' participation in the creator economy: Examining the effect of creators' content sharing frequency on user engagement behavior on digital platforms. *Journal of Retailing and Consumer Services*, vol. 73, p. 103357, https://doi.org/10.1016/j.jretconser.2023.103357

Dean, Brian (2023). TikTok Statistics you need to know in 2024. https://backlinko.com/tiktok-users [Accessed 31 January 2024]

Donaker, G., Kim, H., Luca, M. and Weber, M. (2019). Designing better online review systems. *Harvard Business Review*, vol. 97, no. 6, pp.122-129

Dornekott, D., Holder, U.& Wollborn, P. (2021). Entrepreneurial Efforts and Opportunity Costs: Evidence from Twitch Streamers. *International Entrepreneurship and Management Journal, Forthcoming*. http://dx.doi.org/10.2139/ssrn.380101

Duarte, Fabio (2024). Average time spent on TikTok 2024. Exploding Topics. https://explodingtopics.com/blog/time-spent-on-tiktok [Accessed 20 March 2024]

Ebrahimi, P., Dustmohammadloo, H., Kabiri, H., Bouzari, P., & Fekete-Farkas, M. (2023). Transformational Entrepreneurship and Digital Platforms: A Combination of ISM-MICMAC and Unsupervised Machine Learning Algorithms, *Big Data and Cognitive Computing*, vol. 7, no. 2, p. 118, https://doi.org/10.3390/bdcc7020118

Ehret, C., Hagh, A. & Low, B. (2023). Algorithmic imaginings and critical digital literacy on #BookTok. *New Media & Society*, vol. 0, https://doi.org/10.1177/14614448231206466

Escolano,V. J. C. (2023). How Do Influencers Create Value for SMEs? A TikTok Review in the Philippines. 8th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, pp. 414-419, doi:10.1109/ICBIR57571.2023.10147643

Everitt, B. S., & Hand, D. J. (1981). Finite Mixture Distributions. London, Chapman and Hall

Frank, H., Kessler, A., & Fink, M. (2010). Entrepreneurial orientation and business performance—a replication study. *Schmalenbach Business Review*, vol. 62, pp. 175-198, https://doi.org/10.1007/BF03396804

Freehan, Blair (2023). 2023 TikTok Benchmark Report. Rival IQ. https://www.rivaliq.com/blog/tiktok-benchmark-report/#title-key-takeaways [Accessed 8 February 2024]

Frenkel, Jonathan (2021). The rise and importance of contentpreneurs. Israel Forbes. https://forbes.co.il/e/rise-of-contentpreneurs/ [Accessed 14 March 2024]

Gagliardi, Alyssa (2024). Eleven ways influencers & creators can make money in 2024. Later. https://later.com/blog/how-content-creators-make-money/ [Accessed 14 March 2024]

Gala, K., Schwab, A. & Mueller, B.A. (2024). Star entrepreneurs on digital platforms: Heavy-tailed performance distributions and their generative mechanisms. *Journal of Business Venturing*, vol. 39, no. 1, p.106347, https://doi.org/10.1016/j.jbusvent.2023.106347

Garvey, E., Liguori, E.W. & Santos, S.C. (2023). How digitalization reinvented entrepreneurial resilience during COVID-19. *Technological Forecast Social Change*, vol. 189, p. 122398, https://doi.org/10.1016/j.techfore.2023.122398

GilPress (2023). Twitter Statistics for 2024: Users, Demographics, Trends. https://whatsthebigdata.com/twitter-statistics/ [Accessed 31 January 2024]

Groeneveld, R. A., Meeden, G. (1984). Measuring Skewness and Kurtosis. *Journal of the Royal Statistical Society Series D: The Statistician,* vol. 33, no. 4, pp. 391-399, https://doi.org/10.2307/2987742

Gustafsson, V., Khan, M. S. (2017). Monetising blogs: enterprising behaviour, co-creation of opportunities and social media entrepreneurship. *Journal of Business Venturing Insights*, vol. 7, pp. 26-31, https://doi.org/10.1016/j.jbvi.2017.01.002

Hayat, Zia (2022). World Economic Forum. https://www.weforum.org/agenda/2022/08/digital-trust-how-to-unleash-the-trillion-dollar-opportunity-for-our-global-economy/#:~:text=The%20World%20Bank%20estimates%20that,faster%20than%20physical%20world%20GDP [Accessed 31 January 2024]

Hazari, S., & Sethna, B.N. (2023). A Comparison of lifestyle marketing and brand influencer advertising for generation Z Instagram users. *Journal of Promotion Management*, vol. 29, pp. 491-534, https://doi.org/10.1080/10496491.2022.2163033

Hendricks, J.L., Call, M.L. & Campbell, E.M. (2023). High Performer Peer Effects: A Review, synthesis, and agenda for future research, *Journal of Management*, vol. 49(6), pp. 1997-2029, https://doi.org/10.1177/01492063221138225

Hill, Bruce M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, vol. 3, no. 5, pp. 1163-1174

Ibrar, H., Mushtaq, A. & Riaz, S. (2022). Content Generation in Web 3.0 and Blockchain-Based Decentralized Social Networks: A Theoretical Adoption Framework. TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON), Hong Kong, Hong Kong, pp. 1-6, doi: 10.1109/TENCON55691.2022.9977762

Johnson, N.E., Short, J.C., Chandler, J.A. & Jordan, S.L. (2022). Introducing the contentpreneur: Making the case for research on content creation-based online platforms. *Journal of Business Venturing Insights*, vol. 18, p.e00328, https://doi.org/10.1016/j.jbvi.2022.e00328

Joo, H., Aguinis, H. & Bradley, K.J. (2017). Not all nonnormal distributions are created equal: Improved theoretical and measurement precision. *Journal of Applied Psychology*, vol. 102, no. 7, p.1022, doi: 10.1037/apl0000214

Kemp, Simon (2023). Digital 2023 deep-dive: how much time do we spend on social media? Data Reportal. https://datareportal.com/reports/digital-2023-deep-dive-time-spent-on-social-media [Accessed 20 March 2024]

Kim, J., & Makadok, R. (2022). Where the stars still shine: Some effects of star-performers-turned-managers on organizational performance. *Strategic Management Journal*, vol. 43, no. 12, pp. 2629-2666, https://doi.org/10.1002/smj.3398

Kolsquare: All you need to know about short-form video content. https://www.kolsquare.com/en/blog/all-you-need-to-know-about-short-form-video-content#:~:text=What%20are%20the%20main%20short,exclusively%20for%20short%2Dform%20content. [Accessed 29 January 2024]

Kullolli, T. & Trebicka, B. (2023). Generation Z and the evolution of social media: a two-decade analysis of impact and usage trends. *Interdisciplinary Journal of Research and Development*, vol. 10, p. 77, https://doi.org/10.56345/ijrdv10n311

Kushwaha, B. P. (2021). Paradigm shift in traditional lifestyle to digital lifestyle in Gen Z: a conception of consumer behaviour in the virtual business world. *International Journal of Web Based Communities (IJWBC)*, vol. 17, no. 4, doi:10.1504/IJWBC.2021.119472

Lamal, P., A. (1990). On the importance of replication. *Journal of Social Behaviour and Personality*, vol. 5, no. 4, p. 31.

Li, C., Li, D., Liang, Y., Wang, Z. (2024). Underdog entrepreneurship in the digital era: The effect of digital servitization on household entrepreneurship in China. *Heliyon*, vol. 10, p. e24154, https://doi.org/10.1016/j.heliyon.2024.e24154

Mitzenmacher, M. (2004). A Brief history of generative models for power law and lognormal distributions, *Internet Mathematics*, vol. 1, no. 2, pp. 226–251, https://doi.org/10.1080/15427951.2004.10129088

Moore, Thomas (2024). What is the average amount of Instagram followers? https://viralyft.com/blog/average-amount-of-followers-on-instagram [Accessed 31 January 2024]

Morgeson, F.P. & Humphrey, S. E. (2006). The work design questionnaire (WDQ): developing and validating a comprehensive measure for assessing job design and the nature of work. *Journal of Applied Psychology*, vol. 91, no. 6, p. 1321

Nambisan, S. (2017). Digital entrepreneurship: Toward a digital technology perspective of entrepreneurship. *Entrepreneurship theory and practice*, vol. 41, no. 6, pp.1029-1055, https://doi.org/10.1111/etap.12254

Newman, M.E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, vol. 46, no. 5, pp. 323-351

Newson, R. (2002). Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences. *The Stata Journal*, vol. 2, no. 1, pp. 45-64

Rooney, T. & O'Carroll, R. (2020). Uses and gratifications of generation Z within social networks: a dialectical investigation into the Facebook domain. *Journal of Promotional Communications*, vol. 8, no. 1, pp. 1-27, https://doi.org/10.21427/8TRC-G695

Okumoko, Joy (2022). How to stitch a video on TikTok. Make Use Of. https://www.makeuseof.com/how-to-stitch-tiktok-videos/ [Accessed 28 April 2024]

Queen, Tim (2024) How many Youtube channels are there in 2024? https://timqueen.com/youtube-number-of-channels/ [Accessed 31 January 2024]

Sahut, J.M., Iandoli, L. & Teulon, F. (2019). The age of digital entrepreneurship. *Small Business Economics*, vol. 56, pp.1159-1169, https://doi.org/10.1007/s11187-019-00260-8

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, vol. 13, no. 2, pp. 90-100

Shane, S., & Venkataraman, S. (2000). The Promise of Entrepreneurship as a Field of Research. *The Academy of Management Review*, vol 25, no. 1, pp. 217-226, https://doi.org/10.2307/259271

Short, C. E. & Short, J. C. (2023). The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights*, vol. 19, https://doi.org/10.1016/j.jbvi.2023.e00388

Somers, R.H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, pp.799-811

Srinivasan, A. & Venkatraman, N. (2017) Entrepreneurship in digital platforms: A network-centric view. *Strategic Entrepreneurship Journal*, vol 12, no. 1, pp. 54-71, https://doi.org/10.1002/sej.1272

Statista report (2023): Social media usage worldwide. Statistics report about global social network usage. https://www-statista-com.ludwig.lub.lu.se/study/12393/social-networks-statista-dossier/ [Accessed 30 January 2024]

Taboada-Villamarín, A. (2023). A guide to collecting data from the TikTok Research API for academics using Python. Medium. https://medium.com/@albatabo81/a-guide-to-collecting-data-from-the-tiktok-research-api-for-academics-using-python-19223328f55e [Accessed 28 March 2024]

Tang, L., Xu, Z., Lyu, X. (2023). More popular, more listings? Effects of popularity on Airbnb host expansion. *International Journal of Contemporary Hospitality Management,* vol. 35, pp. 1650-1669, https://doi.org/10.1108/IJCHM-03-2022-0276

TikTok for developers. Research API. https://developers.tiktok.com/products/research-api/ [Accessed 7 February 2024]

Udemy, 2023. Online courses - learn anything, on your schedule | Udemy. https://udemy.com [Accessed 06 April 2024]

Weismeier-Sammer, D. (2011). Entrepreneurial behavior in family firms: A replication study. *Journal of Family Business Strategy*, vol. 2, no. 3, pp. 128-138.

Yoo, Y., Boland, R.J., Lyytinen, K.J., & Majchrzak, A. (2012). Organizing for Innovation in the Digitized World. *Organization Sci.,* vol. 23, pp. 1398-1408, https://doi.org/10.1287/orsc.1120.0771

Appendix 1: Data Protection Plan

1. Introduction

1.1 Purpose of the Data Protection Plan

The purpose of this Data Protection Plan is to outline the measures taken to protect the privacy and confidentiality of user data obtained through TikTok's Research API for the purpose of conducting quantitative research.

2. Data description

2.1. Method and type of data collection

Existing data will be collected using TikTok's Research API, if and when access to such data is granted by TikTok. The material collected will be digital text, and will be saved as an Excel file. The data will then be imported into SPSS for data analysis.

3. Documentation and data quality

3.1. The data collected for this research will be documented and described with comprehensive metadata to facilitate understanding and interpretation. The collection method, encompassing details on the acquisition process through TikTok's Research API, will be outlined. The metadata will define the content, specifying variables such as user demographics, engagement metrics, and any other relevant information.

3.2. The quality of the data will be assured through a combination of rigorous methodologies and documentation practices. To ensure accuracy, repeated measurements will be conducted during data collection, reducing the likelihood of errors or outliers. Data entry validation mechanisms, including range checks and consistency checks, will be implemented to identify and rectify any discrepancies during the input phase.

4. Storage and backup

4.1. Storage and backup procedures will be designed to safeguard against potential loss or corruption. The integrity of the storage will be maintained through secure storage systems. Multiple backups, stored in geographically diverse locations, will be performed at predetermined intervals to minimize the risk of data loss. These backups will include both raw data and associated metadata. Access controls will be implemented to prevent unauthorized alterations.

4.2. Information security and access to data will be controlled through a combination of technical and procedural measures. Access to the research data will be restricted to authorized personnel only. User authentication will be implemented. Sensitive and personal data will be identified and classified, and additional layers of security will be applied to these datasets.

5. Legal and ethical requirements

5.1. To safeguard the data, all processing activities will strictly adhere to relevant data protection laws and ethical guidelines. In particular, compliance with regulations such as GDPR (General Data Protection Regulation) or other applicable data protection laws will be ensured. Confidentiality of data will be maintained through access controls and anonymization methods. Only authorized personnel will have access to the data. Intellectual property rights will be respected, and proper attribution will be given to existing works or data sources. Transparency in reporting and communication will be prioritized to address any concerns related to personal record handling, confidentiality, or intellectual property rights, demonstrating a commitment to ethical research practices.

5.2. Data will be handled ethically through frameworks that prioritize transparency, informed consent, and respect for privacy. Strict adherence to relevant ethical guidelines, institutional review board approvals, and data protection regulations will be maintained throughout the research process. The research will be conducted with integrity, ensuring proper attribution of intellectual property and responsible communication of findings.

6. Data sharing and long-term preservation

6.1. In accordance with applicable regulations, efforts will be made to share the data and metadata to contribute to the transparency and reproducibility of the research. However, if there are legal or ethical constraints, such as privacy concerns or intellectual property issues, appropriate measures will be implemented to balance the imperative of data sharing with the need to protect privacy and adhere to ethical standards.

6.2. Sharing of the data will be facilitated through reputable repositories, subject to the constraints of any legal and ethical limitations. Any identifiable or sensitive information will be appropriately anonymized or aggregated to protect individual privacy. Efforts will be made to ensure compliance with applicable data protection laws and ethical guidelines.

6.3. A unique and persistent identifier (PID) such as a Digital Object Identifier (DOI) will be employed. This practice aligns with best practices in data sharing and promotes transparency and reproducibility in research.

6.4. Lund university is subject to The Swedish National Archives Regulation on preservation and disposal, which stipulates how archiving of research data is to be performed. The research at a university is a public authority activity. The documents produced in research belong to the University, and are therefore subject to the University's archive preservation standards.

7. Responsibilities and resources

7.1. Each individual member of the research team will be responsible for, and will assist with, the data management during the project. Lund University will be responsible for the data management and the keeping of the records long-term, once the project is finished.

7.2. We do not expect there to be any excessive costs associated with the data management for this research. However, some allocation of resources for data management will encompass:

1. Long-term Preservation: covered by the university, costs may include fees associated with repository services or establishing institutional archives.

2. Labor: already covered by the university, paid personnel will aid in data management tasks, including documentation, interpretation, and quality assurance.

8. Conclusion

This Data Protection Plan serves as a guide to ensure the responsible and ethical handling of user data obtained through TikTok's Research API. By implementing these measures, the research aims to contribute valuable insights while safeguarding the privacy of individuals.

Appendix 2: Dictionaries

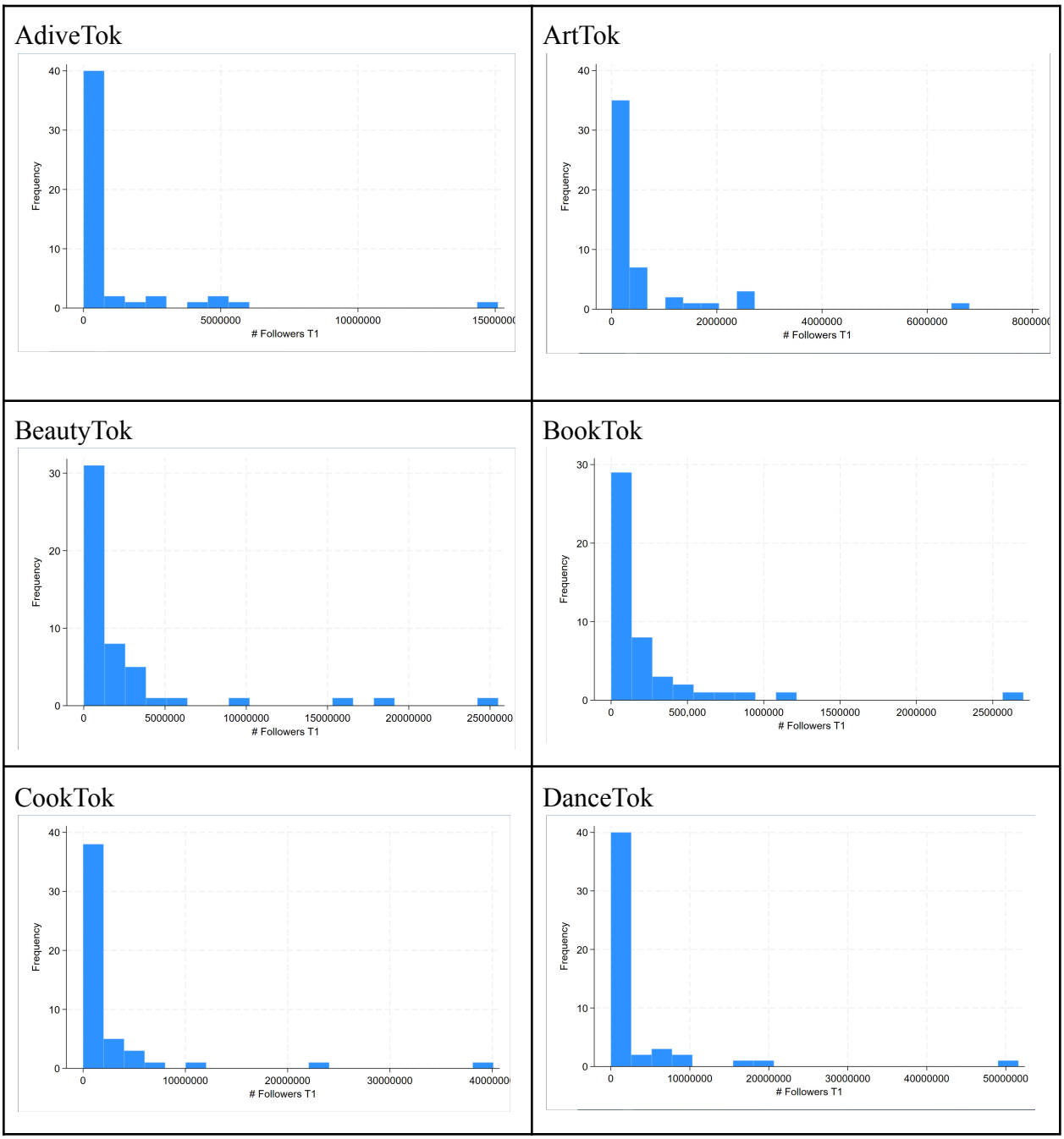
Table A2: Dictionaries for each one of the 21 most active hashtag communities on TikTok.

#AdviceTok #datingadvice #psychology #advicetiktok #lifehacks #motivation #inspirational #psychologyadvice #lifelessons	**#ArtTok** #artistsoftiktok #arttips #howtodraw #tiktokart #art #painting #artist #artwork #mixedmedia	**#BeautyTok** #makeuptips #makeup #makeuphacks #lipstick #beauty #beautyproduct #skincare #makeuptutorial	**#BookTok** #bookquotes #bookrecommendations #books #goodreads #bookstories #bookclub #reading #bookworm
#CookTok #recipes #cooking #foodie #foodhack #baking #easyrecipes #snacks #asmrfood #food #trendingrecipes	**#DanceTok** #tiktokdance #dancetutorial #dancechallenge #professionaldancers #dance #choreography #foryou #dancegirl #dancingtiktoks	**#DIYTok** #diyproject #diyideas #dyiflip #ikeatok #diycraft #ikeahack #renovation #diymakeover	**#FashionTok** #styletips #fashionhacks #outfitideas #styletransformation #fashinmakeover #outfitideas #fashiontrends #mensfashion #womensfashion
#FitTok #gym #fitness #health #workout #mindfulness #meditation #armworkout #legday #gymgirl #shoulders	**#KidTok** #toddlersoftiktok #toys #toddlertok #funnybaby #family #homeschool #scienceforkids	**#MovieTok** #moviereviews #goodfilms #bestmovies #movieactors #independentfilms #moviesuggestions #whattowatch #mustsee	**#MusicTok** #singers #bestmusicians #songreviews #amazingvoices #singingchallenge #covers #viralsongs #accousticcover

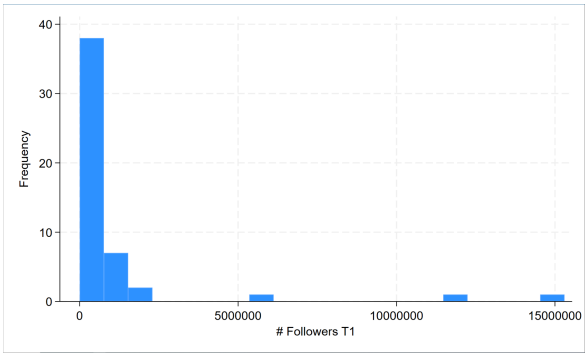
#NurseTok #nursetiktok #nursesoftiktok #nursing #nurselife #nursingstudent #medicalhumor #nurseproblems #healthcareworker	**#ParentTok** #parentsoftiktok #dadsoftiktok #momsoftiktok #singleparenting #parentingtips #parenting #parentingadvice #pet	**#PetTok** #funnypets #petrescue #petsoftiktok #tiktokpets #cutebabyanimals #adorablepet #dogsoftiktok #shelter	**#PlantTok** #plantcaretips #houseplantsoftiktok #houseplants #planthacks #plantsindoor #plantlover #plantparent #repotting
#PrankTok #scareprank #funnyvideos #prank #prankvideo #jumpscare #funnytiktoks #troll #aprilfools	**#QueerTok** #lgbt #lgbtq #blackqueertok #comingout #queertiktok #wlw #gay #transtiktok #queercontent #queerwomen #bi	**#SelfCareTok** #selfcaretiktok #selfcareproducts #mindfulness #skincareroutine #healthyskin #selfgrowth #mindsetshift #growthmindset #selflove #selfgrowthquotes	**#SportsTok** #olympics #sports #athlete #gymnastics #training #sportstiktokchalleng e #soccer #basketball #tennis #volleyball #sportsmotivation
#TVTok #tvtowatch #whattowatch #fromtvtoshow #top10 #netflix #series			

Appendix 3: Hypothesis 1

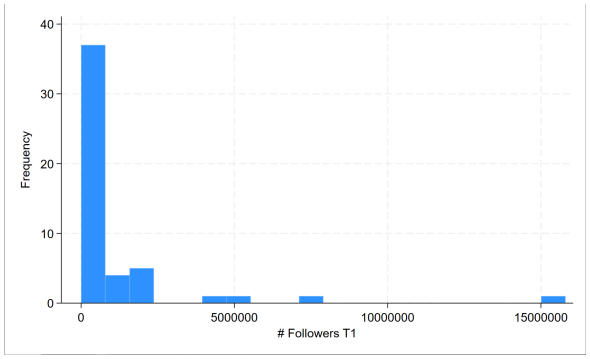
Figure A3: Histogram of performance distribution of each hashtag community using “followers” as the measure of performance.



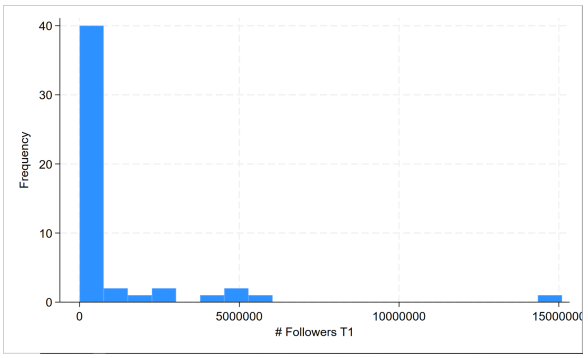
DIYTok



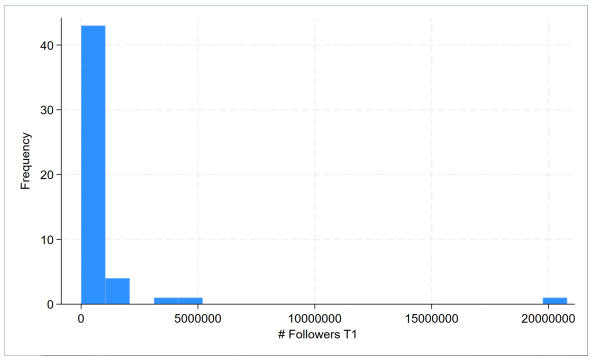
KidTok



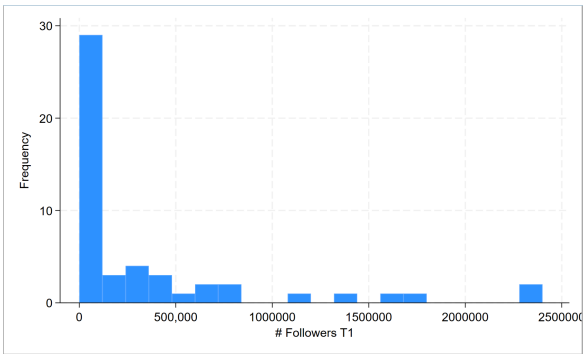
FashionTok



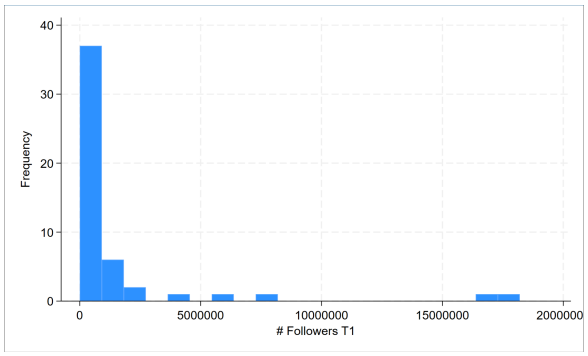
FitTok



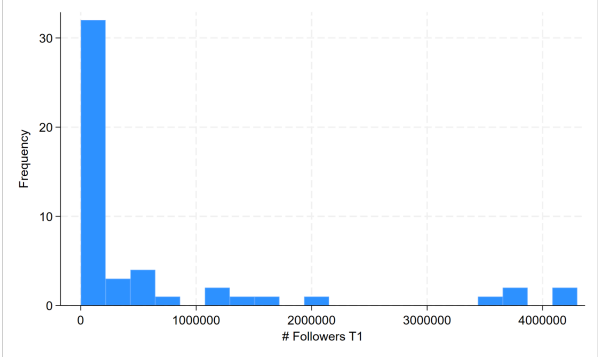
MovieTok



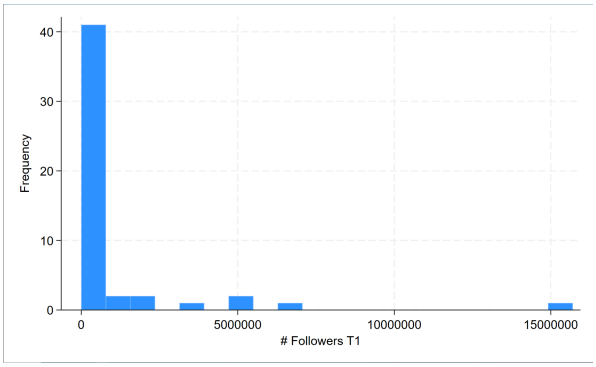
MusicTok



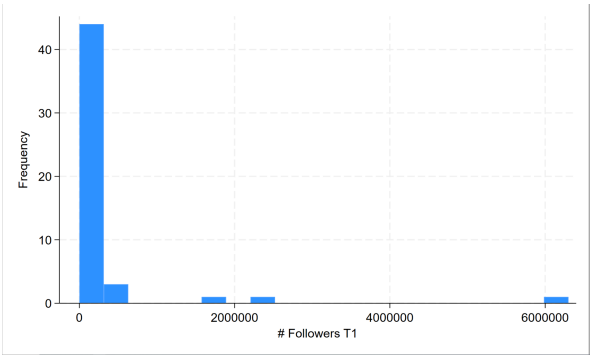
NurseTok



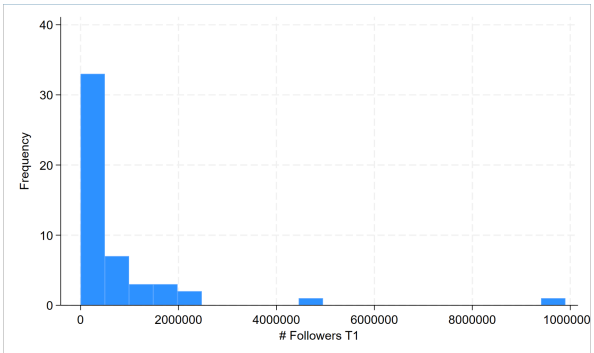
ParentTok



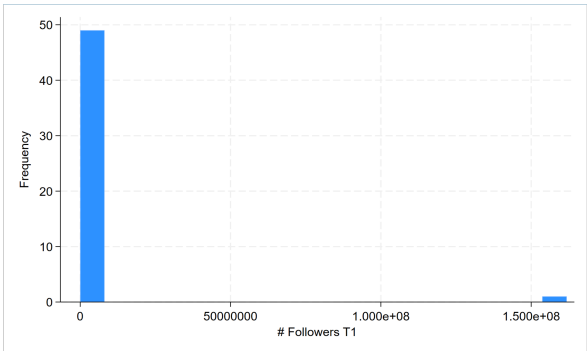
PetTok



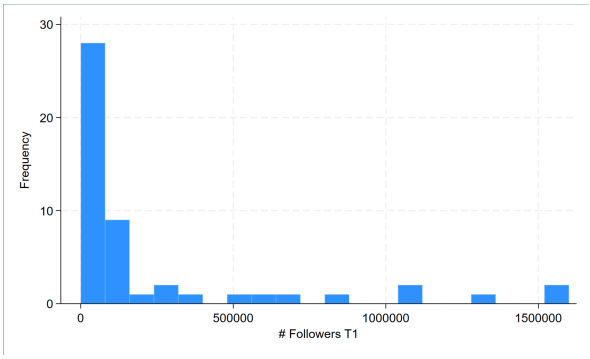
PlantTok



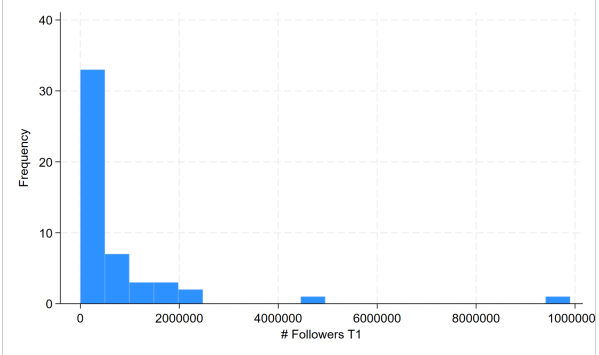
PrankTok



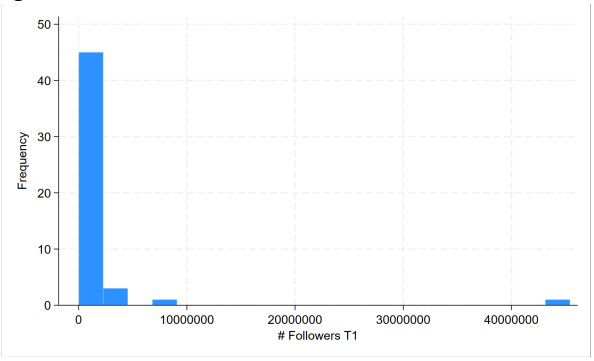
QueerTok



SelfCareTok



SportsTok



TVTok

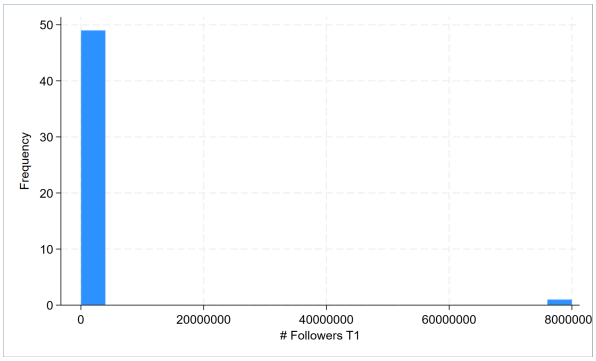


Table A3: Kilmogorov-Smirnov and Anderson-Darling results for each domain, along with total follower count per domain

AdviceTok	**ArtTok**	**BeautyTok**	**BookTok**
User accounts: 50 Followers: 51,442,217 Top 8 content creators have 82.3% of all followers K-S: D: 0.3509; *p*:0.000 D-A: A:1063; *p*: 1.54e-69	User accounts: 50 Followers: 25,858,998 Top 7 content creators have 78.4% of all followers K-S: D: 0.300; *p*:0.000 D-A: A:1022; *p*: 1.23e-53	User accounts: 50 Followers: 117,846,850 Top 3 content creators have 50.2% of all followers K-S: D: 0.2759; *p*:0.000 D-A: A: 993; *p*: 1.12e-45	User accounts: 47 Followers: 10,363,293 Top 4 content creators have 54.4% of all followers K-S: D: 0.2862; *p*:0.000 D-A: A:1039; *p*: 1.34e-62
CookTok	**DanceTok**	**DIYTok**	**FashionTok**
User accounts: 50 Followers: 131,730,993 Top 3 content creators have 55.4% of all followers K-S: D: 0.3130; *p*:0.000 D-A: A: 1080; *p*: 1.59e-67	User accounts: 50 Followers: 143,219,565 Top 7 content creators have 83.9% of all followers K-S: D: 0.3283; *p*:0.000 D-A: A:1165; *p:* 1.31e-71	User accounts: 50 Followers: 50,945,695 Top 3 content creators have 65.2% of all followers K-S: D: 0.3316; *p*:0.000 D-A: A: 1184; *p*: 2.10e-31	User accounts: 50 Followers: 53,112,205 Top 10 content creators have 82.7% of all followers K-S: D: 0.3098; *p*:0.000 D-A: A: 1066; *p*: 1.56e-69
FitTok	**KidTok**	**MovieTok**	**MusicTok**
User accounts: 50 Followers: 56,549,447 Top 13 content creators have 84% of all followers K-S: D: 0.2897; *p*:0.000 D-A: A: 862; *p*: 1.10e-35	User accounts: 50 Followers: 41,133, 566 Top 1 content creator has 50.6% of all followers K-S: D: 0.3505; *p*:0.000 D-A: A: 1320; *p*: 2.56e-38	User accounts: 50 Followers: 17,830,310 Top 6 content creators have 60% of all followers K-S: D: 0.2421; *p*:0.003 D-A: A: 837; *p*: 0.98e-47	User accounts: 50 Followers: 62,952,690 Top 17 content creators have 90.1% of all followers K-S: D: 0.2516; *p*:0.002 D-A: A: 814; *p*: 0.94e-56

NurseTok	ParentTok	PetTok	PlantTok
User accounts: 50 Followers: 71,034,140 Top 12 content creators have 92.4% of all followers K-S: D: 0.3423; *p*:0.000 D-A: A: 1149; *p*: 1.28e-69	User accounts: 50 Followers: 32,529,016 Top 9 content creators have 82.5% of all followers K-S: D: 0.3029; *p*:0.000 D-A: A: 940; *p:* 1.11e-41	User accounts: 50 Followers: 48,512,176 Top 8 content creators have 88.9% of all followers K-S: D: 0.3707; *p*:0.000 D-A: A: 1166; *p*: 1.30e-71	User accounts: 50 Followers: 14,822,846 Top 3 content creators have 70.3% of all followers K-S: D: 0.3862; *p*:0.000 D-A: A: 1267; *p*: 2.24e-33
PrankTok	**QueerTok**	**SelfCareTok**	**SportsTok**
User accounts: 50 Followers: 201,846,063 Top 1 content creator has 80.2% of all followers K-S: D: 0.4789; *p*:0.000 D-A: A: 1542; *p*: 2.89e-72	User accounts: 50 Followers: 12,004,037 Top 5 content creators have 56% of all followers K-S: D: 0.3286; *p*:0.000 D-A: A: 903; *p:* 0.97e-48	User accounts: 50 Followers: 38,004,831 Top 8 content creators have 70.5% of all followers K-S: D: 0.2608; *p*:0.001 D-A: A: 944; *p*: 1.11e-45	User accounts: 50 Followers: 74,986,076 Top 2 content creators have 71.2% of all followers K-S: D: 0.3923; *p*:0.000 D-A: A: 1399; *p*: 2.76e-65
TVTok			
User accounts: 50 Followers: 85,934,019 Top 1 content creator has 93% of all followers K-S: D: 0.5054; *p*:0.000 D-A: A: 1614; *p*: 3.04e-26			

Appendix 4: Hypothesis 2

Table A4: Side by side comparison of fit for each distribution shape, using the K-S goodness-of-fit test. The higher the p-value the higher the likelihood of fit compatibility.

Power Law	Lognormal	Normal	Exponential	PL with Exp C	Weibull	Poisson
D-value: 0.332 p-value: 2.327e-103	D-value: 0.065 p-value: 0.000	D-value: 0.421 p-value:2.147e-169	D-value: 0.423 p-value:1.287e-170	D-value: 0.094 p-value: 2.316e-103	D-value: 0.036 p-value:0.1204	D-value: 0.420 p-value: 2.139e-169

Appendix 5: Hypothesis 3

Table A5: Residual procedure through variance partitioning conducted for each domain. Note: the missing values from the residuals indicate negative values generated from the variance partitioning procedure and were not included in the logtransformed residuals.

Domain	R^2	R^2 log residuals	Number of missing values from residuals	Scale parameter μ	Shape parameter σ	Spearman's log resid and KI
Advice	0.53	0.03	25	12.9	1.4	0.0346
Art	0.10	0.01	39	13.6	1.01	-0.2455
Beauty	0.18	0.00	32	14.5	1.15	0.3849
Book	0.26	0.42	28	11.8	1.02	0.4912
Cook	0.16	0.11	33	14.1	1.76	0.5417
Dance	0.26	0.26	27	14.3	1.41	0.4733

DIY	0.53	0.01	25	13.2	1.38	0.0338
Fit	0.054	0.08	38	13.7	1.61	-0.2168
Kid	0.16	0.22	28	13.5	1	0.1756
Music	0.27	0.56	32	13.5	1.5	0.7792
Nurse	0.42	0.05	30	14.2	1.09	0.1368
Parent	0.19	0.05	30	12.8	1.43	0.3263
Pet	0.05	0.51	38	13.8	1.6	0.7133
Plant	0.11	0.18	36	12.3	2.08	0.2220
Prank	0.15	0.07	23	14.4	1.64	-0.1111
Queer	0.23	0.26	24	11.2	1.60	0.4981
SelfCare	0.36	0.00	25	13	1.12	-0.0869
Sports	0.96	0.038	24	12.9	1.28	0.11
TV	0.07	0.60	40	14.6	1.44	0.6154

Appendix 6: Hypothesis 4

Table A6: Somer's Delta calculated for each domain, and Spearman's correlation between both predictor variables.

Domain	**Somer's D** **Initial Value, Outcome** ***95% CI***	**Somer's D** **Accu. Rate, Outcome** ***95% CI***	**Spearman Rank Correlation** **Initial Value, Accu. Rate** ***95% CI***
AdviceTok	0.99 [0.98, 1.00]	0.44 [-0.11, 0.20]	0.22 [-0.233, 0.278]
ArtTok	1 [0.97, 1.00]	0.053 [-0.139, 0.245]	-0.0361 [-0.318, 0.245]
BeautyTok	1 [1, 1]	0.095 [-0.097, 0.289]	0.0968 [-0.201, 0.395]
BookTok	0.994 [0.983, 1.004]	0.105 [-0.107, 0.318]	0.158 [-0.146, 0.463]
CookTok	0.994 [0.985, 1.003]	0.0238 [-0.160, 0.208]	0.0203 [-0.279, 0.320]
DanceTok	0.982 [0.965, 0.998]	-0.0318 [-0.177, 0.113]	-0.082 [-0.326, 0.160]

DYITok	0.981 [0.954, 1.004]	-0.009 [-0.0174, 0.155]	-0.103 [-0.344, 0.136]
FashionTok	0.998 [0.993, 1.002]	0.067 [-0.088, 0.224]	0.074 [-0.173, 0.321]
FitTok	0.993 [0.984, 1.002]	0.039 [-0.152, 0.231]	0.025 [-.0256, 0.306]
KidTok	0.993 [0.981, 1.005]	0.201 [0.024, 0.379]	0.255 [-0.012, 0.524]
MovieTok	0.995 [0.989, 1.002]	0.169 [-0.038, 0.376]	0.188 [-0.097, 0.474]
MusicTok	0.990 [0.978, 1.003]	0.118 [-0.090, 0.327]	0.086 [-0.219, 0.392]
NurseTok	0.991 [0.979, 1.004]	-0.036 [-0.194, 0.120]	-0.142 [-0.394, 0.109]
ParentTok	0.991 [0.979, 1.002]	0.089 [-0.092, 0.272]	0.0958 [-0.174, 0.365]

PetTok	0.983 [0.967, 0.999]	0.235 [0.026, 0.444]	0.2457 [-0.042, 0.534]
PlantTok	0.986 [0.970, 1.003]	0.190 [-0.016, 0.397]	0.2175 [-0.087, 0.522]
PrankTok	0.955 [0.877, 1.032]	0.083 [-0.115, 0.282]	0.0578 [-0.232, 0.347]
QueerTok	0.996 [0.990, 1.003]	0.291 [0.070, 0.511]	0.3380 [0.033, 0.642]
SelfCareTok	.0994 [0.985, 1.003]	0.088 [-0.102, 0.278]	0.110 [-0.171, 0.393]
SportsTok	0.995 [0.986, 1.004]	0.166 [-0.009, 0.342]	0.228 [-0.027, 0.484]
TVTok	0.990 [0.977, 1.003]	0.277 [0.585, 0.495]	0.3420 [0.051, 0.632]

Appendix 7: Robustness

Hypothesis 1

Table A7: Tests of correlation and non-normality using “Likes” as an alternate measure of performance

Statistic	Value	Meaning
Spearman’s correlation	0.926	Strong correlation between likes and followers
Pearson correlation	0.6692	Strong correlation between likes and followers
Skew	17.8	Significant right skew present
Kurtosis	403.9	Long tail present
K-S	D-value: 1.000, p-value: 0.000	Null hypothesis of normality can be rejected
A-D	A-value: 320.24, p-value: 1.456e-65	Null hypothesis of normality can be rejected

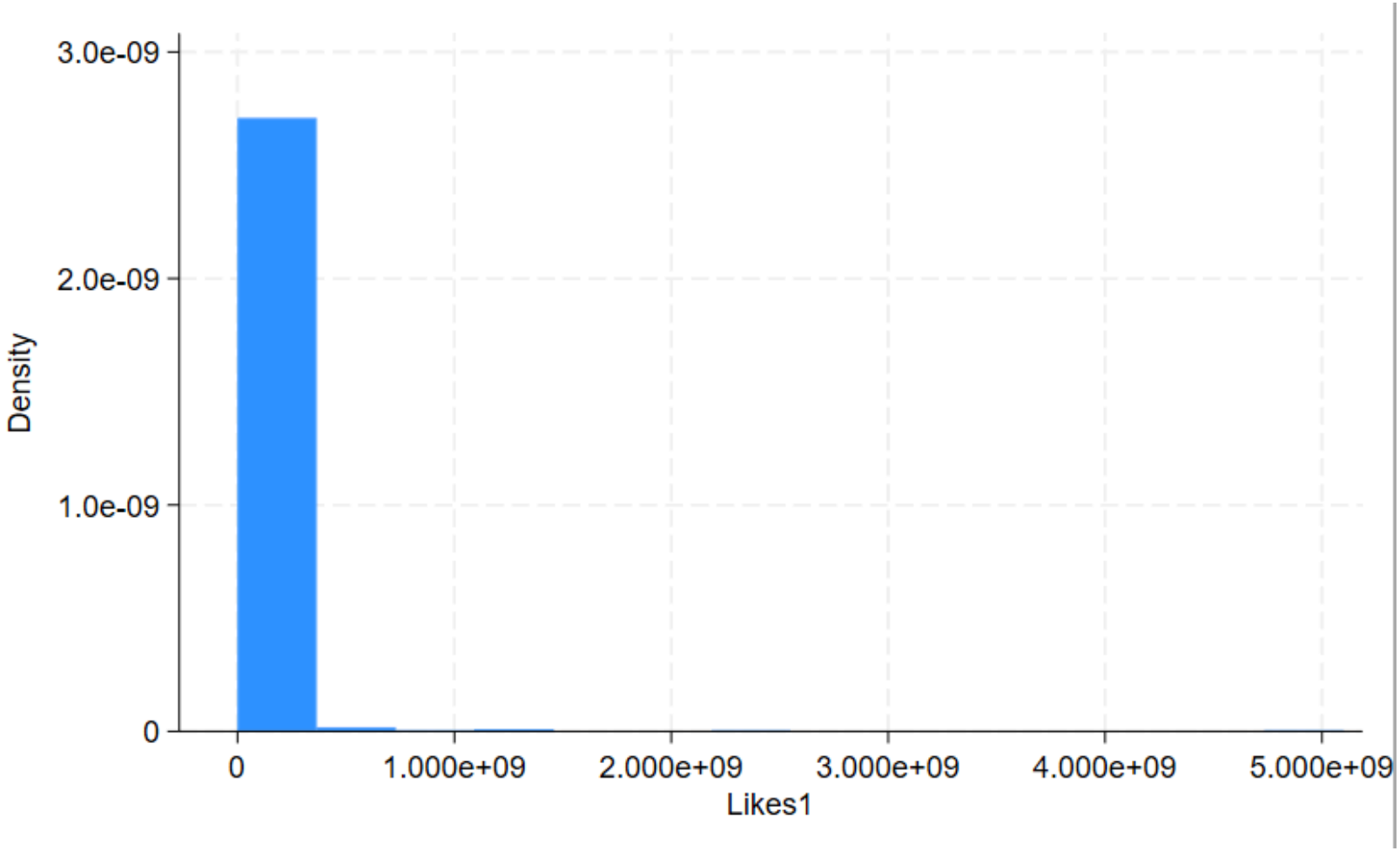


Figure A7: Histogram of number of "likes" showing right skewed heavy-tailed, non-normal distribution.

Table A7.1: Table showing Spearman and Pearson statistics, skew and kurtosis, and K-S and A-D values for each domain.

AdviceTok	**ArtTok**	**BeautyTok**	**BookTok**
Spearman's rho: 0.95 Pearson: 0.97 Skew: 4.7 Kurtosis: 26.8 K-S: D: 1.000; *p*:0.000 D-A: 1145; *p*: 1.85e-23	Spearman's rho: 0.96 Pearson: 0.95 Skew: 3.95 Kurtosis: 20.5 K-S: D: 1.000; *p*:0.000 D-A: 1100; *p*: 1.68e-20	Spearman's rho: 0.91 Pearson: 0.85 Skew: 3.92 Kurtosis: 17.9 K-S: D: 1.000; *p*:0.000 D-A: 1182; *p*: 1.99e-21	Spearman's rho: 0.84 Pearson: 0.88 Skew: 5.77 Kurtosis: 37.1 K-S: D: 1.000; *p*:0.000 D-A: 1231; *p*: 2.19e-27
CookTok	**DanceTok**	**DIYTok**	**FashionTok**
Spearman's rho: 0.94 Pearson: 0.82 Skew: 4.1 Kurtosis: 20.8 K-S: D: 1.000; *p*:0.000 D-A: 1164; *p*: 1.92e-22	Spearman's rho: 0.94 Pearson: 0.97 Skew: 4.48 Kurtosis: 24.0 K-S: D: 1.000; *p*:0.000 D-A: 1229; *p*: 2.18e-26	Spearman's rho: 0.94 Pearson: 0.90 Skew: 5.93 Kurtosis: 38.6 K-S: D: 1.000; *p*:0.000 D-A: 1366; *p*: 2.78e-54	Spearman's rho: 0.89 Pearson: 0.93 Skew: 4.21 Kurtosis: 22.3 K-S: D: 1.000; *p*:0.000 D-A: 1038; *p*: 1.46e-17
FitTok	**KidTok**	**MovieTok**	**MusicTok**
Spearman's rho: 0.92 Pearson: 0.92 Skew: 3.94 Kurtosis: 18.0 K-S: D: 1.000; *p*:0.000 D-A: 1089; *p*: 1.64e-20	Spearman's rho: 0.95 Pearson: 0.97 Skew: 6.31 Kurtosis: 42.8 K-S: D: 1.000; *p*:0.000 D-A: 1361; *p*: 2.76e-53	Spearman's rho: 0.96 Pearson: 0.81 Skew: 4.68 Kurtosis: 27.9 K-S: D: 1.000; *p*:0.000 D-A: 1042; *p*: 1.48e-18	Spearman's rho: 0.96 Pearson: 0.92 Skew: 3.15 Kurtosis: 13.5 K-S: D: 1.000; *p*:0.000 D-A: 907; *p*: 1.05e-11

NurseTok	**ParentTok**	**PetTok**	**PlantTok**
Spearman's rho: 0.89 Pearson: 0.84 Skew: 3.73 Kurtosis: 17.6 K-S: D: 1.000; *p*:0.000 D-A: 1126; *p*: 1.78e-20	Spearman's rho: 0.88 Pearson: 0.93 Skew: 2.6 Kurtosis: 8.87 K-S: D: 1.000; *p*:0.000 D-A: 1021; *p*: 1.41e-19	Spearman's rho: 0.94 Pearson: 0.94 Skew: 3.0 Kurtosis: 11.7 K-S: D: 1.000; *p*:0.000 D-A: 1141; *p*: 1.84e-21	Spearman's rho: 0.95 Pearson: 0.96 Skew: 6.02 Kurtosis: 40.1 K-S: D: 1.000; *p*:0.000 D-A: 1254; *p*: 2.29e-31
PrankTok	**QueerTok**	**SelfCareTok**	**SportsTok**
Spearman's rho: 0.87 Pearson: 0.99 Skew: 6.81 Kurtosis: 47.6 K-S: D: 1.000; *p*:0.000 D-A: 1521; *p*: 3.55e-12	Spearman's rho: 0.84 Pearson: 0.69 Skew: 2.62 Kurtosis: 8.93 K-S: D: 1.000; *p*:0.000 D-A: 1074; *p*: 1.59e-17	Spearman's rho: 0.90 Pearson: 0.92 Skew: 6.36 Kurtosis: 43.5 K-S: D: 1.000; *p*:0.000 D-A: 1240; *p*: 2.23e-30	Spearman's rho: 0.89 Pearson: 0.98 Skew: 6.82 Kurtosis: 47.7 K-S: D: 1.000; *p*:0.000 D-A: 1569; *p*: 3.81e-16
TVTok			
Spearman's rho: 0.91 Pearson: 0.97 Skew: 6.05 Kurtosis: 40.0 K-S: D: 0.5054; *p*:0.000 D-A: 1317; *p*: 2.65e-51			

Hypothesis 2

Table A7.2: Distribution fitting for each domain comparing goodness-of-fit values of all seven distribution shapes, parameters included.

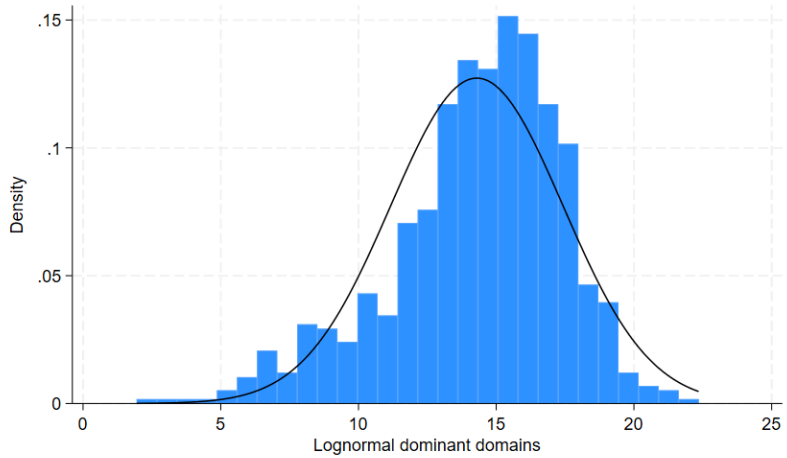
Domain	Power Law	Lognormal	Normal	Exponential	PL with Exp C	Weibull	Poisson
AdviceTok	D-value: 0.099 p-value: 0.006 α= 1.12	D-value: 0.15 p-value: 0.20 μ= 14.14 σ= 3.38	D-value: 0.35 p-value:4.41e-06 μ= 1.99e+07 σ= 5.33e+07	D-value: 0.3813 p-value:4.79e-07 λ= 5.03e-08	D-value:0.089 p-value:0.006 α=1.12 λ=5.03e-08	D-value:0.31 p-value:0.000 β= 1.232 λ=10e+03	D-value:0.29 p-value:0.000 μ=1.99e+07
ArtTok	D-value:0.106 p-value:0.005 α=1.71	D-value: 0.28 p-value: 0.39 μ=13.6 σ=2.96	D-value: 0.3 p-value:0.000 μ=1.04e+07 σ= 2.51e+07	D-value:0.26 p-value:1.44e-08 λ=1.6e+08	D-value:0.102 p-value:0.005 α=1.71 λ=1.6e+08	D-value:0.29 p-value:0.000 β=1.745 λ=10e+03	D-value:0.29 p-value:0.000 μ=1.04e+07
BeautyTok	D-value:0.087 p-value:0.013 α=3.0	D-value: 0.25 p-value: 0.38 μ=15.7 σ=2.98	D-value:0.23 p-value:0.0002 μ=9.72e+07 σ=2.61e+08	D-value:0.45 p-value:2.03e-07 λ=2.65e+08	D-value:0.074 p-value:0.01 α=3.0 λ=2.65e+08	D-value:0.18 p-value:1.20e-07 β=2.35 λ=10e+03	D-value:0.15 p-value:0.0002 μ=9.72e+07
BookTok	D-value:0.089 p-value: 0.00146 α=2.21	D-value: 0.135 p-value: 0.324 μ= 13.98 σ= 2.62	D-value:0.30 p-value:0.000 μ=8748451 σ=2.55e+07	D-value:0.23 p-value: 0.000 λ=1.84e-07	D-value:0.080 p-value:0.00124 α=2.21 λ=1.84e-07	D-value:0.27 p-value:0.013 β=1.46 λ=10e+04	D-value:0.24 p-value:0.000 μ=8748451
CookTok	D-value:0.0879 p-value:0.122 α=1.1	D-value: 0.145 p-value: 0.192 μ= 15.6 σ= 3.05	D-value:0.298 p-value: 0.183 μ=7.75e+07 σ=2.05e+08	D-value:0.002 p-value:0.000 λ=2.3e+09	D-value:0.081 p-value:0.012 α=1.1 λ=2.3e+09	D-value:0.25 p-value:0.000 β=1.87 λ=10e+07	D-value:0.19 p-value:0.000 μ=7.75e+07
DanceTok	D-value:0.0924 p-value:0.0098 α=1.0	D-value: 0.18 p-value: 0.324 μ=15 σ=3	D-value:0.25 p-value:0.00 μ=7.82e+07 σ=2.32e+08	D-value:0.125 p-value:0.000 λ=1.04e-06	D-value:0.091 p-value:0.000 α=1.0 λ=1.04e-06	D-value:0.22 p-value:0.000 β=4.65 λ=10e+04	D-value:0.19 p-value:0.000 μ=7.82e+07
DIYTok	D-value:0.136 p-value:0.0034	D-value: 0.179 p-value: 0.29	D-value:0.34 p-value:0.000	D-value:0.114 p-value:0.000	D-value:0.10 p-value:0.0032	D-value:0.30 p-value:0.000	D-value:0.21 p-value:0.000

	α=1.0	μ=14.85 σ=2.98	μ=2.03e+07 σ=7.98e+07	λ= 2.45e-09	α=1.0 λ=2.45e-09	β=1.32 λ=10e+05	μ=2.03e+07
FashionTok	D-value:0.0923 p-value:0.009 α=1.25	D-value: 0.25 p-value: 0.7 μ=14.98 σ=2.55	D-value:0.35 p-value:0.000 μ=2.09e+07 σ=4.81e+07	D-value:0.023 p-value:0.000 λ=5.74e-08	D-value:0.089 p-value:0.000 α=1.25 λ=5.74e-08	D-value:0.28 p-value:0.003 β=4.87 λ=10e+05	D-value:0.26 p-value:0.000 μ=2.09e+07
FitTok	D-value:0.0822 p-value:0.016 α=1.8	D-value: 0.39 p-value: 0.315 μ=15.23 σ=1.87	D-value:0.36 p-value:0.000 μ=1.99e+07 σ=4.77e+07	D-value: p-value:3.33e-08 λ=10.2e-05	D-value:0.0765 p-value:0.015 α=1.8 λ=10.2e-05	D-value:0.32 p-value:0.000 β=3.56 λ=10e+03	D-value:0.27 p-value:0.000 μ=1.99e+07
KidTok	D-value:0.152 p-value:0.0031 α=1.5	D-value: 0.39 p-value: 0.841 μ=12.48 σ=0.0	D-value:1 p-value:0.841 μ=1.03e+07 σ=4.19e+07	D-value:0.42 p-value:0.000 λ=2.65e-05	D-value:0.122 p-value:0.000 α=1.5 λ=2.65e-05	D-value: 0.94 p-value:0.95 β=3.12 λ=10e+09	D-value:0.89 p-value:0.0003 μ=1.03e+07
MovieTok	D-value:0.092 p-value:0.2099 α=3.0	D-value: 0.025 p-value: 0.0204 μ=13.5 σ=4.03	D-value:0.42 p-value:0.000 μ=1.86e+07 σ=4.45e+07	D-value:0.11 p-value:0.000 λ=1.5e-08	D-value:0.085 p-value:0.0201 α=3.0 λ=1.5e-08	D-value:0.39 p-value:0.0655 β=1.24 λ=10e+04	D-value:0.34 p-value:0.000 μ=1.86e+07
MusicTok	D-value:0.094 p-value:0.20 α=2.0	D-value: 0.13 p-value: 0.236 μ=15.13 σ=3.19	D-value:0.31 p-value:0.000 μ=3.36e+07 σ=6.34e+07	D-value:0.012 p-value:4.76e-05 λ=5.12e+08	D-value:0.088 P-value: 0.02 α=2.0 λ=5.12e+08	D-value:0.28 p-value:0.056 β=2.78 λ=10e+06	D-value:0.26 p-value:0.0003 μ=3.36e+07
NurseTok	D-value:0.0967 p-value:0.0079 α=1.2	D-value: 0.32 p-value: 0.41 μ=14.54 σ=3.28	D-value:0.26 p-value:1.00e-05 μ=3.36e+07 σ=6.34e+07	D-value:0.06 p-value:2.25e-05 λ=7.4e-05	D-value:0.087 p-value:0.0045 α=1.2 λ=7.4e-05	D-value:0.24 p-value:0.000 β=0.98 λ=10e+07	D-value:0.17 p-value:0.0001 μ=3.36e+07
ParentTok	D-value: 0.0946 p-value:0.0087 α=1.14	D-value: 0.35 p-value: 0.37 μ=14.98 σ=3.24	D-value:0.27 p-value:1.00 μ=2.56e+07 σ=5.26e+07	D-value:0.002 p-value:1.24e-08 λ=6.7e+07	D-value:0.087 p-value:0.0124 α=1.14 λ=6.7e+07	D-value:0.20 p-value:0.000 β=2.22 λ=10e+03	D-value:0.18 p-value:0.000 μ=2.56e+07
PetTok	D-value:0.099	D-value: 0.36	D-value:0.37	D-value:0.012	D-value:0.089	D-value:0.245	D-value:0.24

	p-value:0.0068 α=1.6	p-value: 0.40 μ=14.10 σ=3.66	p-value:0.000 μ=1.99e+07 σ=4.77e+07	p-value:4.55e-07 λ=1.2e+08	p-value:0.056 α=1.6 λ=1.2e+08	p-value:2.3e-07 β=3.87 λ=10e+04	p-value:0.000 μ=1.99e+07
PlantTok	D-value: 0.18 p-value:0.0026 α=1.42	D-value: 0.136 p-value: 0.285 μ= 14.10 σ=3.48	D-value: 0.383 p-value:4.20e-07 μ=4330946 σ=1.47e+07	D-value: 0.422 p-value:1.37e-08 λ=6.2e-08	D-value:0.09 p-value:0.0024 α=1.42 λ=6.2e-08	D-value: 0.261 p-value:0.0017 β=4.12 λ=10e+05	D-value:0.3 p-value:0.000 μ=4330946
PrankTok	D-value: 0.0908 p-value:0.010 α=2.1	D-value: 0.102 p-value: 0.452 μ=15.3 σ=2.42	D-value: 0.372 p-value:7.11e-07 μ=6.24e+07 σ=3.38e+08	D-value: 0.412 p-value:2.22e-08 λ=10.5e+08	D-value:0.086 p-value:0.010 α=2.1 λ=10.5e+08	D-value: 0.242 p-value:0.0032 β=3.56 λ=10e+09	D-value:0.31 p-value:0.0001 μ=6.24e+07
QueerTok	D-value: 0.106 p-value:0.0045 α=1.32	D-value: 0.089 p-value: 0.792 μ=13.5 σ=3.03	D-value: 0.382 p-value:4.11e-07 μ=9743483 σ=2.11e+07	D-value: 0.415 p-value:2.31e-08 λ=9e-07	D-value:0.095 p-value:0.034 α=1.32 λ=9e-07	D-value: 0.010 p-value:0.0014 β=2.452 λ=10e+03	D-value:0.31 p-value:0.002 μ=9743483
SelfCareTok	D-value: 0.0919 p-value:0.0194 α=1.13	D-value: 0.114 p-value: 0.500 μ=15.0 σ=2.78	D-value: 0.380 p-value:4.10e-07 μ=2.95e+07 σ=9.92e+07	D-value:0.312 p-value:1.15e-07 λ=7.1e-07	D-value:0.082 p-value:0.0123 α=1.13 λ=7.1e-07	D-value: 0.32 p-value:0.0016 β=2.12 λ=10e+04	D-value:0.31 p-value:0.000 μ=2.95e+07
SportsTok	D-value: 0.102 p-value:0.0063 α=1.5	D-value: 0.16 p-value: 0.45 μ=14.14 σ=3.32	D-value:0.345 p-value:4.12e-07 μ=1.19e+08 σ=7.20e+08	D-value:0.012 p-value:3.12e-07 λ=11.2e-07	D-value:0.098 p-value:0.0045 α=1.5 λ=11.2e-07	D-value: 0.33 p-value:0.000 β=3.23 λ=10e+05	D-value:0.28 p-value:0.000 μ=1.19e+08
TVTok	D-value:0.104 p-value:0.0053 α=2.0	D-value: 0.24 p-value: 0.36 μ=13.6 σ=3.08	D-value:0.32 p-value:4.09e-07 μ=1.23e+07 σ=4.66e+07	D-value:0.02 p-value:0.00 λ=10e-06	D-value:0.096 p-value:0.0041 α=2.0 λ=10e-06	D-value:0.31 p-value:0.0001 β=1.45 λ=10e+03	D-value:0.24 p-value:0.000 μ=1.23e+07

Figure A7.1: Logtransformed domains compatible with a lognormal fit, logtransformed for better visualization. Note: the effect of logtransforming a non-normal distribution can skew the original right tail to the left, as seen in the images. This is an attempt at normalizing the distribution.

Lognormal dominant domains, logtransformed



Lognormal co-dominant domains, logtransformed

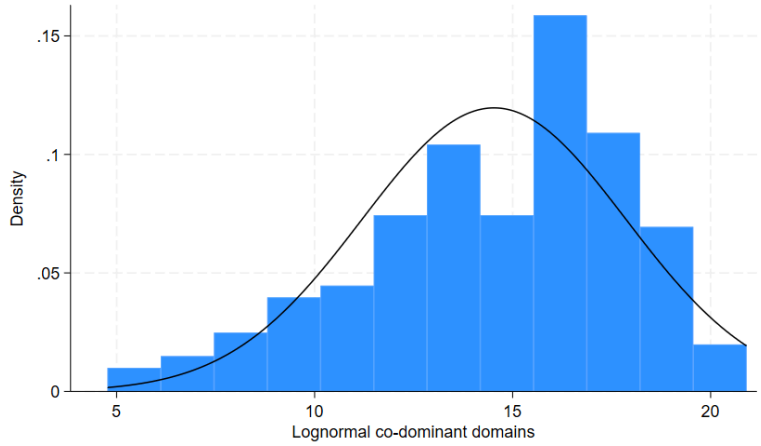


Table A7.3: Frequency of performance distribution per domain using “Likes” as the measure of performance.

		Domain count	Percentage
Dominant Distributions	Exponential Tail	0	0
	Lognormal	17	80.95%
	Normal	0	0
	Power Law	1	4.76%
	Power Law with cutoff	0	0
	Poisson	0	0
	Weibull	0	0
Co-dominant Distributions	Lognormal and Power Law	1	4.76%
	Lognormal, Normal, Weibull	1	4.76%
	Lognormal and Normal	1	4.76%
No dominant or co-dominant distribution	N/A	0	0
Total		21	100%

Hypothesis 3

Table A7.4: Spearman correlation of residuals (generated through variance partitioning using “Likes” as the measure of performance) and knowledge intensity (KI) for each the 11 lognormal dominant and co-dominant domains.

Domain	Spearman’s log resid and KI	Domain	Spearman’s log resid and KI
Advice	0.27	Kid	0.16
Art	0.10	Music	0.65
Beauty	0.08	Nurse	0.50
Book	0.44	Parent	0.056

Cook	0.41	Pet	0.81
Dance	0.53	Plant	0.29
DIY	-0.035	Prank	-0.14
Fashion	0.64	Queer	0.61
Fit	-0.053	SelfCare	0.11
Sports	0.28		
TV	0.42		

Table A7.5: OLS regression of kurtosis. Note: n=20; *p < 0.05, **p < 0.01

NP regression	**Resid. Df**	**Df**	**Coefficient**	**t**	**P > | t |**	**F**	**Pr(>F)**	**Adj. R^2**
Content.	18	1	-21.80283	-0.70	0.492	0.49	0.4916	-0.0274
Cont+ KI	17	2	Cont.: -32.9	Cont.:-1.04	Cont.: 0.311	1.15	0.3404	0.0154
			KI: 113.34	KI: 1.34	KI: 0.199			

Hypothesis 4

Table A7.6: Kendall's Tau-A measure of association across domains (performance= number of likes). Note: CI - 95% Confidence Interval.

Kendall Tau-A	**All domains (n=21)**	**Lognormal dominant domains (n=6)**	**Lognormal co-dominant domains (n=8)**
Initial value and Outcome *(avr)* CI 95% (lowest to highest)	0.95 [0.9505, 0.9852]	0.96 [0.9581, 1.006]	0.93 [0.9186, 0.9888]
Accumulation Rate and Outcome *(avr)* CI 95% (lowest to highest)	0.42 [0.5227, 0.6215]	0.42 [0.4786, 0.6650]	0.40 [0.4721, 0.6281]