UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Gabriel Pereira de Oliveira

On the dynamics of music virality and its relation with mainstream success

Belo Horizonte
2025

Gabriel Pereira de Oliveira

On the dynamics of music virality and its relation with mainstream success

Final Version

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Mirella Moura Moro
Co-Advisor: Ana Paula Couto da Silva

Belo Horizonte
2025

Oliveira, Gabriel Pereira de.

O48o On the dynamics of music virality and its relation with mainstream success [recurso eletrônico] / Gabriel Pereira de Oliveira– 2025.
1 recurso online (147 f. il., color.) : pdf.

Orientadora: Mirella Moura Moro.
Coorientadora: Ana Paula Couto da Silva.

Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f.130-145

1. Computação – Teses. 2. Computação social – Teses. 3. Redes sociais on-line – Teses. 4. Ciência de dados – Teses. 5.Música e internet – Teses I. Moro, Mirella Moura. II. Silva, Ana Paula Couto da. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. IV. Título.

CDU 519.6*04 (043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

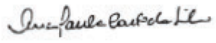
FOLHA DE APROVAÇÃO

On the dynamics of music virality and its relation with mainstream success

GABRIEL PEREIRA DE OLIVEIRA

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores(a):

PROFA. MIRELLA MOURA MORO - Orientadora
Departamento de Ciência da Computação - UFMG



PROFA. ANA PAULA COUTO DA SILVA - Coorientadora
Departamento de Ciência da Computação - UFMG

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES
Departamento de Ciência da Computação - UFMG

Documento assinado digitalmente
gov.br FLAVIO LUIZ SCHIAVONI
Data: 27/11/2025 14:37:13-0300
Verifique em https://validar.iti.gov.br

PROF. FLÁVIO LUIZ SCHIAVONI
Departamento de Computação - UFSJ

PROF. LUCA VASSIO
Dipartimento di Automatica e Informatica - Politecnico di Torino

PROF. LUCAS NASCIMENTO FERREIRA
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 31 de outubro de 2025.

To everyone who walked beside me, in life and in science, making this path lighter and more meaningful.

Acknowledgments

Doing a PhD is anything but simple, yet it has been one of the most rewarding and happiest experiences of my life. With all my gratitude, I would like to thank the many incredible people who accompanied me throughout this journey. Having great people together made this a truly calm and fulfilling period.

First and foremost, my family, my safe harbor. To my mother, Marileia, who inspires me every day with her faith and unconditional love, and who was the first to recognize the transformative power of education in my life. To my father, Gerson, who taught me to always do things well and who was never tired of searching for answers to my endless questions. To my sister, Nikolle, who is my example of hard work and resilience. To my grandparents, relatives, and to all those who came before me. *Obrigado!*

To all my dear friends, my family by choice, who helped me to breathe during the toughest moments of this journey. Special thanks to (in alphabetical order) Felippe Ribas, Gabriela Scarabelli, Guilherme Dutra, Hallini Jardim, Julia Camila, Larissa Malagoli, Lucas Lima, Luísa Reis, Luíza Trindade, Matheus Marinho, and Matheus Marques, who were with me every day, sharing the joys and challenges of being a researcher.

To my incredible advisors, for being not only scientific role models but also examples of generosity and empathy, qualities so necessary in academia. Thank you for creating a positive and healthy environment that allowed me to grow as a scientist and professional. To my advisor, Professor Mirella Moro, who made me fall in love with Computer Science and research, thank you for your encouragement, kindness, valuable conversations, advice, and unwavering support. To my co-advisor, Professor Ana Paula Couto, always so kind and humane, thank you for your contributions and constant support, and for reminding me that life extends beyond research. I also thank all my teachers and professors, since childhood, who encouraged me to follow the path of learning.

I am also grateful to all the wonderful people I met in Italy during my research stay at Politecnico di Torino. To Professors Marco Mellia and Luca Vassio, for their support and collaboration, which greatly enriched my experience and offered valuable learning opportunities. To everyone at SmartData, who warmly welcomed me from day one. And to my volleyball friends, who made the city feel like home. *Grazie mille!*

To my friends from the CS+X Lab, especially my dearest friends and academic siblings, Mariana Silva and Danilo Seufitelli. Doing research and sharing life with you made everything lighter and way more fun. Thank you for all the shared moments, the conferences, the gossip, and the amazing teamwork. Also, to Michele Brandão, Michele

Brito, and Natércia Aguilar, for their friendship and support during this time.

My gratitude to UFMG and DCC, their professors and staff, for providing me with both a high-quality education and a research environment that opened doors I never thought possible. Having access to free, public, and high-quality education was essential to my development. Without it, I would not have had the opportunities, lessons, and experiences that shaped who I am today.

I am also thankful to the other research projects that I have worked on: Analytical Capabilities Program and SyntheticGIST, from which I learned so much, as well as to CAPES for the scholarship during my research stay in Italy through the *Programa Doutorado Sanduíche no Exterior* (PDSE).

To everyone who directly or indirectly contributed to this work.

Finally, to music, which has accompanied me through every moment of life and holds the power to heal, to inspire dreams, and to renew faith in life.

"*Não há saber mais ou saber menos; há saberes diferentes.*"
(Paulo Freire)

Resumo

Músicas que se tornam virais não são um fenômeno novo na indústria musical. Ainda assim, tal fenômeno atingiu novos patamares com a popularização da Web e de plataformas sociais, permitindo que músicas alcancem o status de sucesso mundial quase instantaneamente. Enquanto a viralidade musical e o sucesso comercial estão intimamente relacionados, eles continuam sendo conceitos distintos, e plataformas como o TikTok desempenharam um papel crucial na amplificação da viralidade musical e na transformação de músicas em sucessos. O objetivo deste trabalho é analisar o fenômeno da viralidade musical e sua relação com o sucesso tradicional sob a perspectiva da ciência da computação. Especificamente, tal objetivo é avaliado através de três etapas distintas: (i) caracterização, na qual é realizada uma análise quantitativa do que difere as músicas virais das de sucesso em relação às suas características intrínsecas e extrínsecas; (ii) análise de causalidade, na qual é investigada a relação temporal entre viralidade musical e sucesso usando técnicas estatísticas e de causalidade; e (iii) modelagem, na qual foi empregada uma abordagem epidemiológica para representar a viralização musical como um processo de contágio social. Os resultados indicam que as características relacionadas ao artista e temporais estão entre os indicadores mais relevantes para distinguir músicas de sucesso e virais. Na verdade, há potencial para utilizar a viralidade musical para prever o sucesso futuro e vice-versa, embora isso não se aplique a todas as músicas. Além disso, a abordagem epidemiológica captura de forma eficaz a dinâmica da viralidade, mas se mostra menos adequada para outros aspectos da popularidade, como o sucesso em longo prazo. No geral, este trabalho reforça a distinção entre os conceitos de viralidade e sucesso, ao mesmo tempo em que enfatiza sua relação simbiótica impulsionada pelas plataformas sociais. Ele contribui não apenas para as pesquisas na interseção entre Computação e Música, mas também para outras áreas do conhecimento e para a própria indústria musical.

Palavras-chave: viralidade *online*; sucesso musical; computação social; ciência de dados.

Abstract

Songs going viral are not a new phenomenon in the music industry. Still, it has reached new heights with the popularization of the Web and social platforms, which allow songs to achieve worldwide hit status almost instantly. Whereas musical virality and commercial success are closely related, they remain distinct concepts, and platforms such as TikTok have played a crucial role in amplifying music virality and turning songs into successful hits. The goal of this work is to analyze the phenomenon of music virality and its relation to mainstream success from a computer science perspective. Specifically, we assess this goal through three distinct steps: (i) characterization, in which we perform a quantitative analysis of what differs viral from hit songs regarding their intrinsic and extrinsic characteristics; (ii) causality analysis, in which we investigate the temporal connection between musical virality and success by using statistical and causality techniques; and (iii) modeling, in which we use an epidemiological approach for representing music viralization as a social contagion process. The results indicate that artist-related and temporal features are among the most relevant for distinguishing between viral and hit songs. In fact, there is potential for using music virality to forecast future success and vice versa, although this does not apply to all songs. Moreover, the epidemiological approach effectively captures viral dynamics, whereas it is less suited to other aspects of popularity, such as long-term success. Overall, this work reinforces the distinction between virality and success while emphasizing their symbiotic relationship driven by social platforms. It contributes not only to the research in the intersection of computer science and music but also to other knowledge fields and the music industry itself.

Keywords: online virality; musical success; social computing; data science.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Every day, people have access to a massive volume of content on the Web, especially on Social Networks. In such platforms, users can share and repost content (i.e., a blog post, a video, or a song) from others at any moment, and some posts get a lot of shares in a short amount of time, reaching several other users. In 2023, approximately 4.9 billion people around the world used social media, and this number can reach 5.85 billion users by 2027.[1] In fact, the popularization and interconnected nature of these platforms transcend geographical boundaries and can define trends and influence behaviors on a global scale.

In social media, "going viral" means that specific content spreads quickly across social platforms, being shared by thousands or even millions of users in a very short time span [31]. Indeed, such processes are inherently social, meaning they heavily depend on people's actions – whether sharing content online or talking about it [48]. The viral phenomenon is present in several contexts, but most notably within online social networks, be it TikTok short videos [52] or Twitter posts [18]. Understanding viral spreading in social media may be useful for a myriad of purposes, including marketing [16] and dealing with fake news and other relevant issues [34, 45, 56].

Being more social every day, music is no exception to the effect of viral spreading. Viral songs are widely shared in a short amount of time, and they may (but not necessarily) become successful by reaching the top of the charts with millions of streams and digital sales. In fact, streaming services are now the most used form of music consumption, and platforms such as Spotify and YouTube allow users to share what they are listening to with their contacts directly. According to the International Federation of the Phonographic Industry (IFPI), audio and video streaming collectively account for 62% of the time individuals dedicate to interacting with music in 2023.[2] Such relevance is also reflected in the economy. For instance, in Brazil, streaming is responsible for 86.2% of the total revenue of the national phonographic market in 2022.[3]

Indeed, the streaming ecosystem enables the viral spread of songs, but this process

[1] Forbes: https://www.forbes.com/advisor/in/business/social-media-statistics/

[2] IFPI Engaging with Music 2023: https://www.ifpi.org/wp-content/uploads/2023/12/IFPI-Engaging-With-Music-2023_full-report.pdf

[3] Pró-Música Brasil - Mercado Fonográfico Brasileiro 2022: https://pro-musicabr.org.br/wp-content/uploads/2023/03/2023-03-20-Mercado-Brasileiros-em-2023.pdf

is also affected by other social platforms. For music, a viral song may gain extensive attention and popularity due to factors such as catchy melodies and engaging visuals. While music has long been subject to viral spread, platforms such as X (previously known as Twitter), YouTube, and TikTok have significantly magnified this trend by providing reach and accessibility to millions of users worldwide [42, 52].

Specifically, platforms such as TikTok have reshaped the way music is discovered, shared, and consumed, with viral trends on the platform often translating into mainstream success for artists. For example, in December 2023, the version of the Brazilian song "Escrito Nas Estrelas" by Lauana Prado (originally performed by Tetê Espíndola in 1985) topped the charts of the most streamed songs in Brazil after going viral on social media following a contestant's performance on a reality show.[4] On a global scale, the 1985 hit "Running Up That Hill (A Deal with God)" by Kate Bush reached a new peak on the music charts in 2022 after being used as a soundtrack in the Netflix series *Stranger Things*. The song went viral on TikTok soon after, reaching a younger audience who had not been born when the song was first released.[5]

1.1 Virality and Success Definition

For the purpose of this thesis, musical virality and success are not synonymous. Despite being closely connected, we consider them as distinct facets of a song's popularity, a broader concept frequently associated with getting noticed by many people [119]. We make such a separation following the music industry trend. Streaming services and specialized magazines started to differentiate viral from hit (i.e., successful) songs in their popularity charts. Therefore, distinguishing the two concepts is fundamental to understanding the dynamics of music consumption in the digital age. Here, we summarize the definitions of musical success and virality considered in this thesis.

The concept of **virality** is related to quickly spreading and disseminating content across various platforms and social networks [31, 32]. For music, a viral song gains widespread attention when it is shared by thousands or even millions of users in a very short time span. Indeed, this concept is strictly related to social platforms, and it does not have a unique metric for it. For example, Guerini et al. [31] define virality as the number of people who accessed a specific content in a given time interval, being associated with other measurable phenomena, including appreciation, buzz, and controversiality. In the

[4]Vagalume: https://www.vagalume.com.br/news/2024/01/02/lauana-prado-tem-o-primeiro-numero-1-de-2024-com-escrito-nas-estrelas2.html

[5]Forbes: https://www.forbes.com/sites/petersuciu/2022/11/21/running-up-that-hilltiktok-and-youtube-are-giving-old-songs-new-life/

music industry, specialized magazines and streaming services released virality charts that consider metrics such as the number of accesses and engagements (e.g., TikTok Billboard Top 50,[6] Spotify Viral 50,[7] YouTube Trending Videos[8]).

In contrast, musical **success** is associated with the music consumption itself, but it also does not have a unique metric or definition. Still, according to Seufitelli et al. [98], the most used success definitions in the field of Hit Song Science (HSS) can be divided into three main perspectives: (i) top-charts, which are rankings of songs based mainly on sales, media airplay, and/or online streaming (e.g., Billboard Hot 100,[9] Official UK Singles Chart,[10] Crowley Top 100 Brasil[11]); (ii) economy indicators, such as the number of single/album sales; and (iii) engagement, which considers the social dimension of success, including the number of views or likes on social platforms.

In this thesis, to analyze the dynamics of music popularity in current times (driven by digital consumption), we consider the top-chart perspective for measuring both virality and success. Nowadays, all major streaming platforms (e.g., Spotify, YouTube, and Deezer) produce rankings of viral and hit songs. Therefore, we consider viral all songs that have entered a viral chart, whereas hits are those songs that have made it into a distinct success chart (e.g., the most listened-to songs). We do so regardless of their position in the charts, that is, songs ranked first and last are equally considered viral/hit.

1.2 Motivation and Relevance

One goal of Computer Science is to develop computational models to understand the dynamics of society and to create applications to promote aspects of life. In this sense, interdisciplinarity plays a key role, leveraging the emergence of several new knowledge fields. In the context of this thesis, two key computing-related disciplines are Music Information Retrieval (MIR) and Web Science. The first one emerges as a research field based on musicology, psychology, and computer science to extract meaningful information from musical content [50, 120], whereas the second is related to the study of the Web, the information generated by it, and its impacts on society [115].

The phenomenon of content virality on the Web, particularly within social media platforms, has become an important aspect of modern digital culture. The virality of

[6] https://www.billboard.com/charts/tiktok-billboard-top-50/
[7] https://charts.spotify.com/
[8] https://charts.youtube.com/
[9] https://www.billboard.com/charts/hot-100/
[10] https://www.officialcharts.com/charts/singles-chart/
[11] https://charts.crowley.com.br/index.html

a content can also be used to measure its popularity. A notable example of this trend is observed on YouTube, where research predominantly centers around predicting the popularity of videos [37, 42]. Later, TikTok has emerged as a primary platform for witnessing content virality. Studies on such a platform delve into understanding personal motivations and behaviors on TikTok [48], along with investigating the content features that potentially underlie virality. Specifically, Ling et al. [52] evaluate three research hypotheses about what makes a viral content: content elements, the recommendation system, and the creator's profile. Their findings reveal that features such as the creator's popularity and the videos' point of view help distinguish between short videos that will go viral and those that will not.

Songs going viral is not something new in the music industry, as songs have become popular for decades due to various factors, such as memes, viral dances, or even by simply resonating with the public in an unexpected way. In fact, this phenomenon transcends eras and formats, but it has reached a new level with the popularization of the Web and social platforms such as YouTube and TikTok. Now, a song can go viral instantly, driven by dance challenges, hashtags, or massive sharing. The instantaneous and global nature of the Web allows a song to reach audiences worldwide in a matter of hours, resulting in a cycle of sharing and reproduction that may lead it to smash success.

In the digital age, the relationship between virality and success is not necessarily a two-way street, i.e., the virality of a song does not always translate into commercial success. While virality can temporarily drive a song to stardom, sustained success depends on many factors beyond initial viral reach. For example, the song "What is Love?" by TWICE is a K-pop song that reached viral status as of its release in April 2018 but has never become a hit in Global charts. The inverse scenario is also true; there are hit songs that have never become viral, such as "Formation" by Beyoncé (a big hit in 2016 that has never made it to the viral charts, so far).

Still, there are songs that start as viral and then become hits. For example, Figure 1.1 illustrates an example of the relationship between virality and success of a song on Spotify, one of the main streaming platforms worldwide. After its release in mid-2019, the track "Dance Monkey" by the Australian singer Tones and I quickly went viral on social media and reached the top of the viral charts. Streaming success followed soon after, gradually increasing until reaching a peak of approximately 8.9 million daily streams. The main difference between the two behaviors lies in their duration, as the song's virality is much more ephemeral than its success.

In this thesis, we are interested in analyzing the phenomenon of online music viralization and how it relates to the traditional definition of success. As mentioned in Section 1.1, our assumption is that virality and success represent distinct facets of musical popularity, and we analyze them from a top-chart perspective. In short, the question that guides this thesis is: *given that songs reach a certain degree of popularity,*

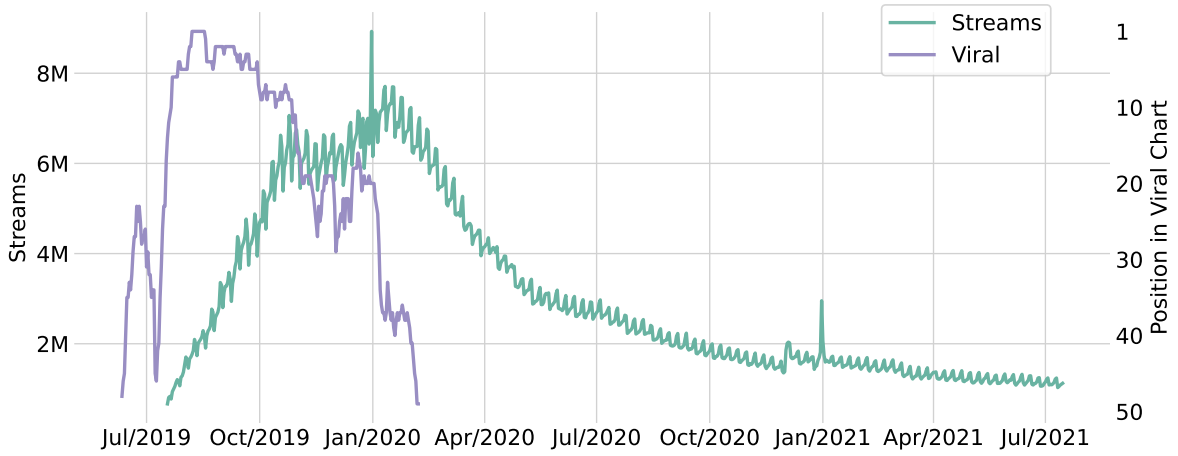


Figure 1.1: Performance of the song "Dance Monkey" by Tones and I on Spotify Global Charts (2019–2021).

what are the factors – intrinsic or extrinsic – that make some of them viral and what are the mechanisms behind their online spread? This question is formalized into three specific research goals, as detailed next.

1.3 Research Goals

The fast and widespread dissemination of music through social platforms not only amplifies the exposure of artists and their songs, but also transforms the way people discover and interact with music. However, there are not many studies addressing music virality and how it impacts success. Existing research focuses mainly on viral marketing to promote music [38, 39] and listeners' behavior [24]. In addition, Araujo et al. [3] consider only acoustic features to analyze how the presence of a song in viral charts affects its posterior popularity (i.e., the presence in the most streamed-to charts) and vice versa.

Therefore, this thesis aims to **understand the phenomenon of music virialization on social platforms and its relationship with mainstream success**. We do so by using a data-driven methodology, combining disciplines such as machine learning, time series analysis, and epidemics. Figure 1.2 presents the overview of this thesis, from the musical data to the final modeling step. Specifically, our analyses are divided into three specific research goals (RGs):

RG1. Characterize viral and hit songs in order to identify the factors that distinguish them;

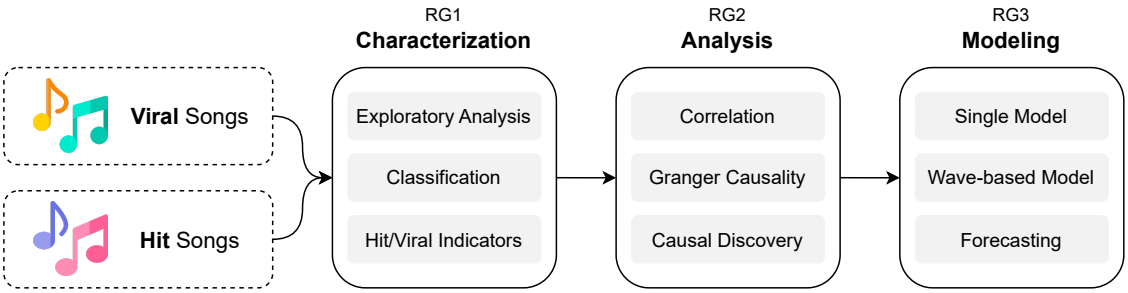


Figure 1.2: Thesis overview according to its research goals.

RG2. Analyze music virality as a stepping stone for mainstream success, i.e., the temporal relationship between both concepts; and

RG3. Model music virality as a process of social contagion for a deeper understanding of the dynamics of music viralization on such platforms.

1.4 Summary of Contributions

This thesis has interdisciplinary nature, and its main contributions are related to the social computing field. In particular, we contribute to a more in-depth understanding of a relevant phenomenon in modern society (i.e., music viralization) by modeling this phenomenon as a contagion process considering multiple dimensions. We also apply techniques such as machine learning and causal analysis to a large volume of data from multiple sources. However, because it is on the frontier between computing and other knowledge fields, the results of this thesis open up a world of possibilities for sciences such as music, communication, and sociology to advance in their respective areas.

Specifically, this thesis' first contribution is a literature review that summarizes the existing research on music virality and its implications. Then, the main practical contributions, which rely on analyzing the music virality phenomenon in both Global and Brazilian markets, are described as follows. The topics are organized according to the Research Goal (RG) they are related to.

RG1. Quantitative Characterization of Viral and Hit Songs (Chapter 3).

1. We perform a quantitative and data-driven comparison of hit and viral songs on the Global and Brazilian markets, as extracted from their respective charts, to unveil similar and distinct patterns regarding their features;

2. We analyze the impact of music-related features from multiple sources on the classification of popular songs as viral and hits. We do so by evaluating three incremental research hypotheses considering acoustic, intrinsic, and extrinsic features;

3. We model the distinction between viral and hit songs as a binary classification task. Our goal is not to predict virality but to understand what differs viral from hit songs. We evaluate a set of classifiers over the features from our research hypotheses, and the best result is achieved by the model containing both intrinsic and extrinsic features, with F1-Scores above 0.7; and

4. We perform a feature importance analysis to uncover those that most contribute to the performance of the chosen classifiers, i.e., what makes a song viral once it achieves popularity. Our results reveal that artist-related and temporal features are among the most important in such a task, highlighting the importance of considering features external to the songs' structure when analyzing popularity.

RG2. Causal Relationship Between Virality and Success (Chapter 4).

1. We model the evolution of musical virality and success by using time series that represent the performance of songs on charts from streaming platforms;

2. We also perform a correlation analysis to reveal synchrony patterns between virality and success. The results reveal that it is not possible to affirm that there is *always* synchrony (i.e., a high correlation) between virality and success. Although this is true for a set of songs, others present a low correlation between such trajectories;

3. We use Granger Causality in such time series to verify whether musical virality can be used to forecast success and vice versa. Our findings infer that for some songs, virality can forecast future success and vice versa, but this cannot be generalized to all songs in our dataset; and

4. We finally address the causal discovery task in this specific context, which aims to qualitatively describe the causal relationship between virality and success. Again, our findings confirm the previous results, revealing that virality may cause success (and vice versa) for some songs, but not all.

RG3. Music Virality as a Contagion Process (Chapters 5 and 6).

1. We apply epidemic models to songs' streaming data from Spotify to capture music popularity dynamics. The results show that such models are more suitable for representing virality than long-term success;

2. We introduce a wave-based modeling approach that better reflects the nature of viral diffusion on streaming. Such an approach successfully captures the multiple bursts of virality that a song may present;

3. We evaluate the forecasting performance of our method against traditional time-series methods, and the results are shown to be comparable to such conventional approaches; and

4. We perform a case study using data from TikTok to verify the suitability of our methodology in other platforms. Overall, the results show that while virality dynamics differ across platforms, the epidemiological approach successfully models the contagion process on TikTok, an important step towards its generalization.

1.5 Text Organization

The remainder of this thesis is organized as follows. We first present a literature review on music virality to summarize the existing research on this subject in Chapter 2. Then, in Chapter 3, we characterize viral and hit songs to identify the factors that differ them. Next, in Chapter 4 we analyze the temporal relationship between virality and success to verify the causality between the two phenomena. Chapters 5 and 6 contain the modeling of the music virality phenomenon as a contagion process and a case study on TikTok data, respectively. Finally, Chapter 7 concludes this thesis and discusses research future directions.

Chapter 2

A Literature Review on Music Virality

With the popularization of the Internet, virality has reached another level, allowing ideas, news, videos, and music to spread exponentially in a matter of hours or days. Online social platforms allow individuals to share content anytime and anywhere, and the inherent sharing nature of these platforms significantly amplifies the viral phenomenon associated with their content. Understanding viral spreading in social media may be useful for a myriad of purposes, including marketing [16] and dealing with fake news and other social issues [34, 45, 56]. For instance, Chou et al. [18] propose a method to early identify the viral spreading social issues (e.g., "Arab Spring" and "Ocean plastic pollution") on Twitter to solve them quickly. Employing such a method can assist in addressing these issues proactively, enabling policymakers to implement preventive measures for a range of social issues before they attract news coverage.

The virality of a content can also be used as a measure of its popularity. A notable example of this trend is observed on YouTube, where research predominantly centers around predicting the popularity of videos [37, 42]. Recently, TikTok has emerged as a primary platform for witnessing content virality. Studies on such a platform delve into understanding personal motivations and behaviors on TikTok [48], along with investigating the content features that potentially underlie virality. Specifically, Ling et al. [52] evaluate three research hypotheses about what makes a viral content: content elements, the recommendation system, and the creator's profile. Their findings reveal that features such as the creator's popularity and the videos' point of view help in distinguishing between short videos that will go viral and the ones that will not.

More recently, song virality has also become a trending research topic. Although interconnected, success and virality represent distinct facets of music popularity [73], and thus, it is important to analyze each of them separately. Investigating the factors leading to musical success has been extensively in a field called Hit Song Science (HSS), and Seufitelli et al. [98] present a comprehensive survey on the subject. In such a work, they review the main studies and present a generic workflow for HSS. Other surveys about the music industry analyze the crowdsourcing phenomena [29] and music connections with

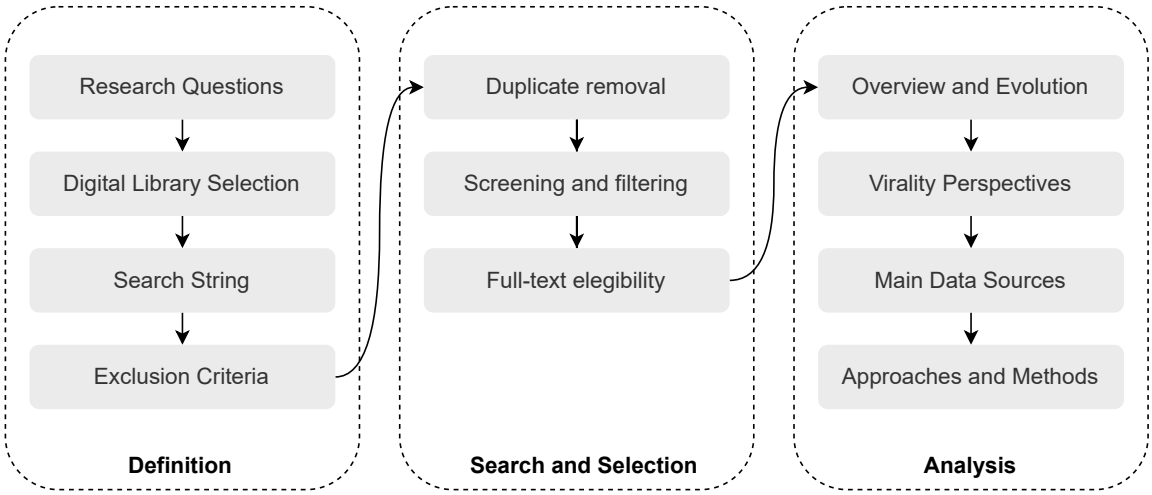


Figure 2.1: Summary of the literature review protocol.

business studies [84]. However, regarding music virality, besides the increase in studies analyzing this subject, there is still no work consolidating such knowledge.

Therefore, this chapter aims to understand and summarize existing research on music virality on social media from a computer science perspective. More specifically, we aim to discover the main virality perspectives and techniques used to analyze and model such a phenomenon. We do so through a literature review, which presents an overview of the most relevant studies on this subject. Performing it offers several advantages for research works, such as comprehensiveness and methodological rigor, ensuring a structured and impartial search for the most relevant studies. In addition, it allows for synthesizing the works by identifying patterns, gaps, and contradictions in them, as well as indicating open research problems for future work.

This chapter is organized as follows. Section 2.1 presents the three-phase protocol for performing the review. Then, Sections 2.2, 2.3, 2.4, and 2.5 present the four analyses of this review: (i) overview and temporal evolution; (ii) virality perspectives and relation with success; (iii) main platforms and data sources; and (iv) approaches and methodologies. Finally, Section 2.6 presents a discussion on the review and describes how this thesis assesses open research problems on music virality.

2.1 Review Protocol

Inspired by the work of Pizzolitto [84], we employ a three-phase protocol in this literature review. This protocol is illustrated by Figure 2.1 and allows a rigorous and structured approach to synthesize the existing knowledge on music virality and its impli-

cations (i.e., the topic of interest of this thesis). Next, we describe in detail each of the review phases: definition (Section 2.1.1), search and selection (Section 2.1.2), and analysis (Section 2.1.3).

2.1.1 Phase 1: Definition

The definition phase of the review comprises all the steps prior to the search and selection of papers themselves. Specifically, we define the research questions, the databases, the search string, and the criteria for exclusion of papers.

Research Questions. The first step of the review involves defining the research questions that will guide the research and analysis. Based on the main goal of the review (i.e., to understand the existing research on music viralization on social media), we aim to answer the following research questions (RQs):

RQ1. How has research on music virality evolved?

RQ2. How do music virality processes on social media relate to the concept of mainstream success?

RQ3. What are the main domains, i.e., media platforms, in which viralization processes occur?

RQ4. What are the main approaches used to understand and model viral processes?

Digital Library Selection. Next, we define the databases from which the articles will be retrieved. Despite being located in the field of social computing, this thesis has a strongly interdisciplinary character, being on the field with other areas of knowledge such as music itself, communication, sociology, among others. Therefore, it is important to analyze not only works from computer science, but also from these other areas, as such works can provide valuable knowledge about the definition and the factors behind the viralization processes. Here, we use four specific digital libraries, comprising works from several areas of knowledge: SCOPUS,[1] Web of Science,[2] EBSCOHost,[3] and DBLP.[4]

[1] https://www.scopus.com/
[2] https://www.webofknowledge.com/
[3] https://www.ebsco.com/
[4] https://dblp.org/

Search String. To retrieve papers on music virality, we use the following search string to the selected databases: `viral* AND (music OR song)`. We use the notation `viral*` to represent all derived terms of viral, including *virality*, *viralization*, and so on. We apply this string to the title, abstract and keywords fields of the works.

Exclusion Criteria. After defining the search string, we define article exclusion criteria (ECs) used in the selection phase (specifically, in the screening and filtering step). Such criteria help to filter the works that are actually related to the objective of the review, and each of them is described below.

EC1. The work is not written in English;

EC2. The work has not been published in conference proceedings or journals;

EC3. The work is not peer-reviewed;

EC4. The work is not about music virality processes in social media (e.g., health, viral diseases, bioinformatics);

EC5. The work does not analyze the virality of the songs themselves, but focuses in other aspects (e.g., music-related textual or image memes).

2.1.2 Phase 2: Search and Selection

The second phase of the review protocol involves the search and selection of works to be analyzed. In this thesis, we consider studies published up to June 30, 2024 in peer-reviewed journals or conference proceedings. In total, 1,096 records were identified in the four databases considered (SCOPUS: 266, Web of Science: 125, EBSCOHost: 693, DBLP: 12). From this initial set, we perform four steps to create the final set of studies.

Duplicate Removal. Because we consider more than one database to search papers, there are cases in which the same work is indexed in more than one database. Therefore, we perform a duplicate record removal step through titles and authors.

Screening and Filtering. In this step, we filter the records by reading the title and abstract. While evaluating the studies, we also apply the exclusion criteria defined in the previous phase to exclude works that are not of interest to this research.

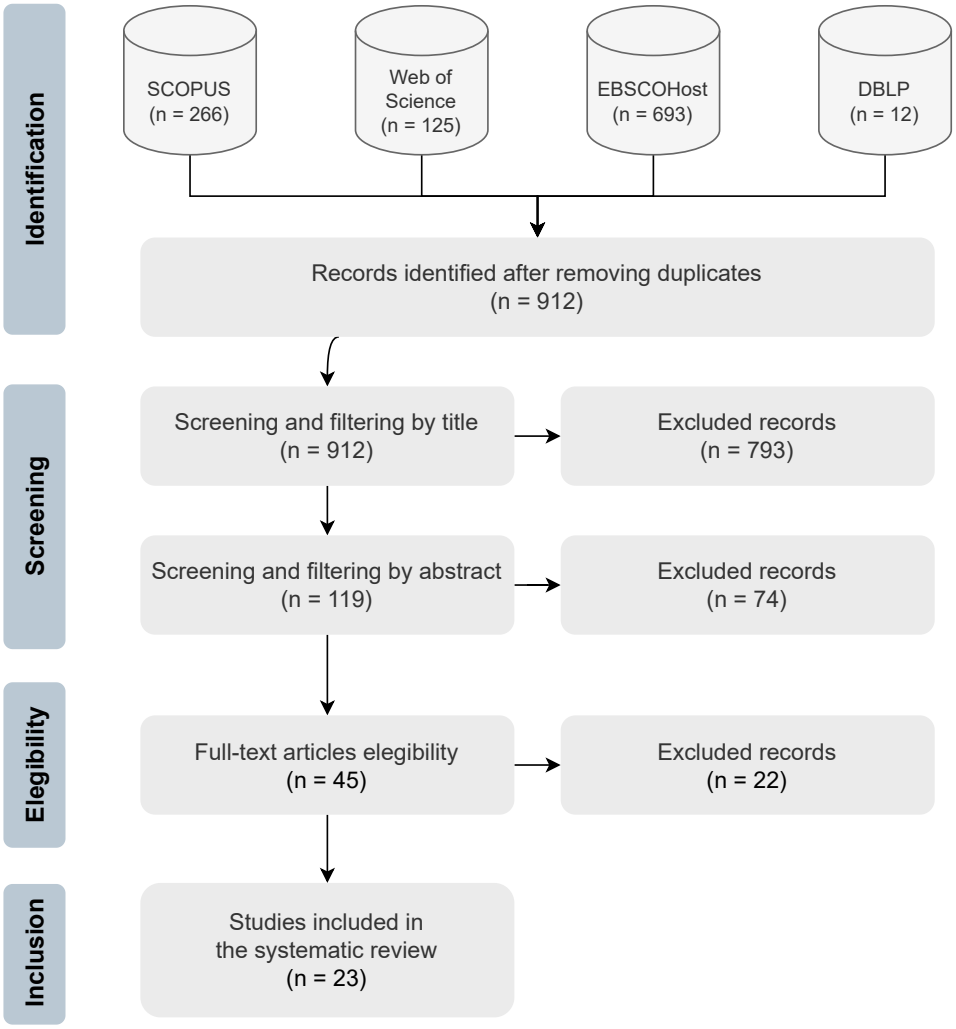


Figure 2.2: Literature review search and selection results based on the PRISMA framework.

Full-text Eligibility. The last stage of the selection is the evaluation of the articles according to the full text, excluding records that do not fit the research goal.

Figure 2.2 illustrates the execution flow of the search and selection phase based on the PRISMA framework [79],[5] including the number of studies considered in each step. After performing all the steps, our final set includes 23 studies within the review's scope and will be considered in the analysis phase. The complete list of papers is described further in this chapter, in Section 2.6.

[5]PRISMA stands for Preferred Reporting Items for Systematic Reviews and Meta-Analyses. It provides a guideline for reporting different types or aspects of systematic reviews.

2.1.3 Phase 3: Analysis

Finally, after selecting the final set of studies, the last phase of the review comprises the analyses themselves. Each of the analyses performed in this chapter aims to answer one of the research questions mapped in the definition phase of the review.

Overview and Evolution (RQ1). In this first analysis, we examine the historical development of research on music virality by identifying patterns and trends over time.

Virality Perspectives (RQ2). Next, we detail the definitions used for virality in the music context, in addition to investigating the relationship with music success and analyzing the metrics and criteria that connect these two concepts.

Main Data Sources (RQ3). Here, we discuss the main data sources used in studies on music virality, such as social networks, streaming platforms, and consumer data.

Approaches and Methods (RQ4). Finally, we investigate the approaches and techniques employed to study virality, from quantitative methods, such as statistical modeling and machine learning, to qualitative analyses, such as case studies and interviews.

2.2 Overview and Temporal Evolution

We start our analyses by assessing RQ1 (*"How has research on music virality evolved?"*) through an overview of research on music virality. Specifically, we analyze the temporal evolution of research on this subject, tracing how interest in the topic has grown over time and identifying key milestones in the field. We also examine the areas of knowledge where such research is conducted, highlighting interdisciplinary contributions from fields such as communication, musicology, marketing, and computer science. In addition, we consider the impact of technological advancements on the studies, such as the rise of specific social media platforms and streaming services.

Content viralization is not a recent phenomenon, nor is the dissemination of music to the general public. For example, in the 1990s and early 2000s, music videos exhibited on television were one of the main ways for people to get music known [20]. However, with the popularization and democratization of access to the Web, social media platforms have played a major role in this process. Therefore, the study by Tan [111] is the first to use the expression "going viral" to refer to the dissemination of musical content on social platforms. Specifically, the object of the research is a video containing a rap song

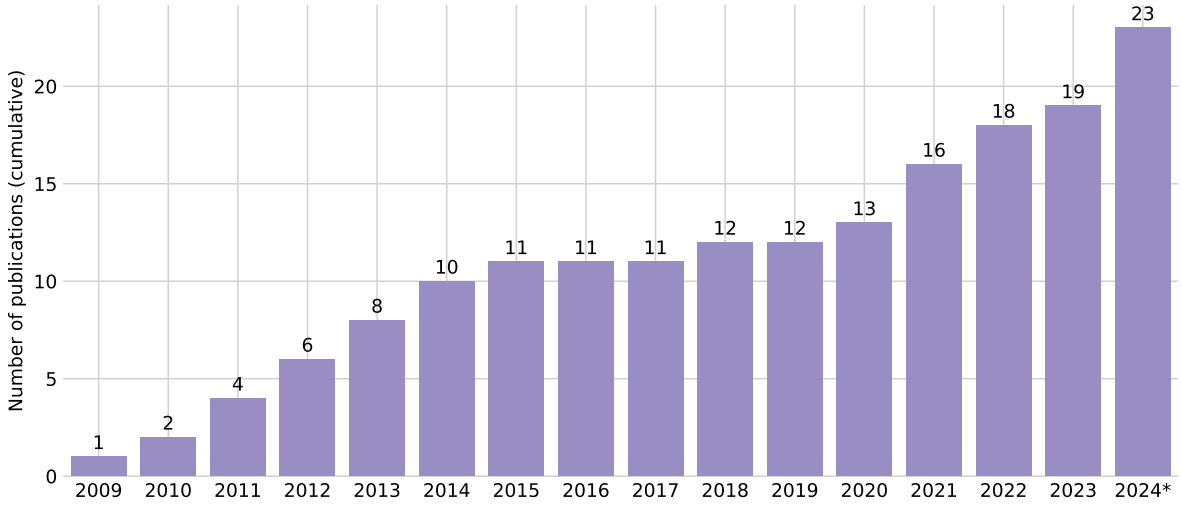


Figure 2.3: Cumulative publications on music virality, 2009 – 2024* (search performed in June 2024; i.e., before the year ended).

produced by the Media Development Authority (MDA) of Singapore. In this context, the author relates the popularity of the video on YouTube to the spread of a computer virus.

Since then, music viralization on social platforms has been a subject of constant research. Figure 2.3 illustrates the temporal evolution of studies on this phenomenon. As of June 2024, we identified 23 peer-reviewed studies in English that deal exclusively with musical virality, highlighting the relevance of this emerging research topic. Several other works deal with other aspects of music viralization or its role on social platforms, including memes, semiotics, and language analyses. However, such studies are outside the scope of this review, as they do not address the process of diffusion of these songs on networks, which is the primary goal of this research.

When analyzing the evolution of studies on music virality, there are two specific periods in which there was a significant increase in the number of scientific works. The initial boom occurred from 2009 to 2014, which in the music industry corresponds to the transition period from physical sales models to streaming and digital downloads. The consolidation of social networks reshaped how people interact with cultural products, and platforms such as Facebook, Twitter, and especially YouTube were spaces where music could be shared, discussed, and consumed directly. In addition, the meme culture and the growing trend of user-generated content became central to how music was consumed and disseminated. The song "Gangnam Style" by South Korean singer Psy is a good example of the power of platforms to boost music popularity. The song is a cultural phenomenon, having over 5.2 billion views on YouTube[6] and being considered by many to be one of the main reasons for the popularization of K-pop worldwide.[7]

[6] As of September 2024: https://www.youtube.com/trends/records/

[7] https://www.nme.com/features/opinion/psy-gangnam-style-10-years-anniversary-k-pop-impact-3269841

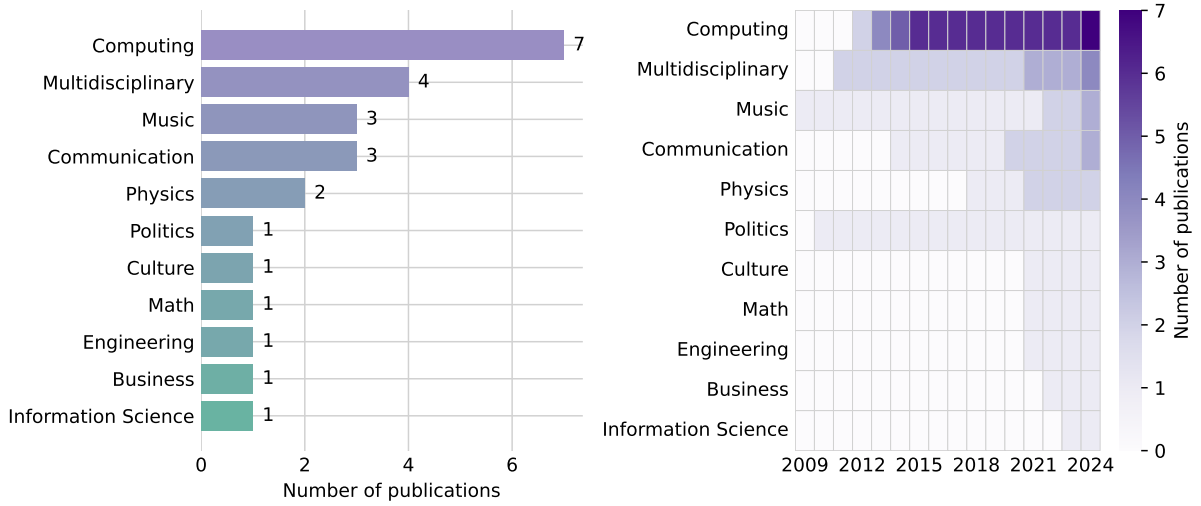


Figure 2.4: Number of publications (left) and temporal evolution (right) by venue discipline. Note that there are venues with more than one knowledge field.

The second period of significant growth in research on music virality began in 2020 and continues to this day. This period is marked by a new wave of social and technological transformations that have directly influenced music consumption. First, during the COVID-19 pandemic, digital media consumption skyrocketed, and music became a form of escape and social interaction. This period also marked the consolidation of short video platforms, such as TikTok, as the most influential for music discovery and viralization. Their ability to create and share music content easily has driven several tracks to popularity. For example, 13 of the 14 No. 1 songs on the Billboard Hot 100 (i.e., the main music chart in the United States) in 2022 were driven by trends on TikTok.[8]

All these social, economic, and technological aspects make the viralization of music on social networks the subject of studies in the most diverse areas of knowledge. Figure 2.4 (left) shows the distribution of studies by field of knowledge of the media (i.e., journals and conferences) where they were published. The areas with the largest number of studies are computer science (7), music (3), and communication (3). In addition, there are studies published in inherently multidisciplinary venues, reinforcing the relevance and complexity of this topic. Nevertheless, other aspects of the viralization of music are also the subject of research in areas such as physics, mathematics, politics, business, and information science. The temporal evolution of studies in these areas (Figure 2.4, right) shows that most studies in computing took place in the first years of research (2009–2015), while studies in multidisciplinary and music venues are more recent.

[8] https://www.musicbusinessworldwide.com/13-out-of-the-14-no-1-songs-in-the-us-in-2022-were-driven-by-viral-trends-on-tiktok/

2.3 Virality Perspectives and Connections with Success

After having an overview on the research evolution about music virality, we now address RQ2 (*"How do music virality processes on social media relate to the concept of mainstream success?"*) by analyzing how the concept of virality is portrayed in the studies considered in this review and how it relates to the concept of success defined in Chapter 1. Despite being a relatively simple concept related to the fast dissemination of content on social platforms [31, 32], the concept of virality is described from different perspectives in the studies analyzed. Therefore, we categorize such works into five groups according to their perspective of virality: epidemiological, sociocultural, technological, and marketing. Each of these perspectives is described next.

The first group analyzes musical virality from an **epidemiological** perspective, treating the phenomenon as a contagion process, in which songs spread rapidly, similar to the spread of a virus [60, 111]. Specifically, the study by Lehman [49] considers "going viral" as a digital contagion, whose origin in Latin means "touching". Furthermore, Rosati et al. [89] argue that a viral song could "infect" people through its dissemination across multiple media and platforms. Based on this analogy with epidemics, at the end of this period of contagion, a large part of the susceptible population would have been "infected" by this song, i.e., this population would recognize the song.

Other studies portray virality as a **sociocultural** phenomenon, that is, as a process deeply rooted in social and cultural interactions. Audiovisual content is referred to as the one that best represents the viral phenomenon [17], and people's engagement in the content disseminated on the platforms is an essential part of such a phenomenon [9, 23]. For example, when analyzing the virality of a music video in the 2008 US presidential campaign, Wallsten [118] relates virality to the result of a complex and multidirectional interplay between the actions of Internet users, bloggers, campaign members, and journalists. The culture of memes is also mentioned by other works [10, 114], and Freitas [26] states that "virality, trolling, sharing, and creating are common practices of contemporary media that are not exclusive to memes, and rather symbiotic to all these forms of media".

The third group of works uses a **technological** perspective, considering virality as a phenomenon shaped and facilitated by digital platforms and social media. For these studies, the platforms' structure and users' dynamics play a fundamental role in disseminating online content [51, 73]. For example, in the context of music videos, Edmond [22] cite actions such as tagging, sharing, liking, and using music videos as part of broader social media activities. In this sense, platforms such as YouTube [55, 93] and TikTok [20, 61] are frequently cited when assessing the viral phenomenon.

Finally, some studies analyze musical virality from a **marketing** perspective, using concepts such as viral marketing and electronic word-of-mouth. The former is defined as the set of techniques that exploit preexisting social networks and interpersonal influences to increase awareness and achieve commercial objectives (e.g., sales, streams, etc.) [6]. In contrast, the latter refers to the sharing of opinions on digital platforms to generate awareness about a subject or product among users [99, 100]. The works of Kahl [38] and Kahl and Albers [39] relate the two concepts by stating that music is inherently viral and that viral marketing strategies aim to stimulate word-of-mouth among consumers to actively propagate musical content, helping to expand its reach.

Virality Mechanism. It is also possible to classify the works according to their propagation mechanisms and underlying intentions. That is, whether the studies work with organic or induced viralization. Most of the works identified in this review consider the first type, which occurs spontaneously, driven by genuine interest and engagement from the public. In this process, the content spreads naturally as people share it because they find it relevant, interesting, or emotionally resonant [9, 10, 17, 22, 23, 26, 49, 55, 60, 61, 73, 89, 93, 114]. However, some works consider what we call induced virality, in which the dissemination of content results from different marketing actions (e.g., use of influencers, hashtags, or advertising campaigns) that aim to stimulate sharing and increase the visibility of the content [6, 38, 39, 51, 99, 100]. Moreover, there are also works that consider both mechanisms in their analyses by investigating the relationship between political campaigns and the viralization of a music video [118] and also the relationship between TikTok, music creation, and the industry [20].

Virality versus Success. As mentioned in Section 1.1, in this thesis, we consider virality and success as two distinct facets of a song's popularity. In this regard, not all studies differentiate musical virality from the traditional definition of success. However, those that do, recognize both as inherent processes of songs [26] and associate success with their frequency of consumption [6]. In fact, musical success is traditionally measured by tangible metrics such as album sales [99, 100], streams [10, 73], chart positions [20], and awards at large-scale events such as the Grammys or the MTV Video Music Awards [22].

Furthermore, virality can also be seen as a stepping stone to mainstream success, but not as a guarantee of it [55]. The symbiotic relationship between virality and success is addressed by Coulter [20] and Biasioli [10], who argue that there is a high level of virality spillover between social media and streaming. In other words, viral trends can serve as a way to bring a song to the mainstream. Therefore, the complex and multifaceted relationship between both concepts reflects the changing dynamics of the music industry as a whole, highlighting the need for a deeper understanding of music consumption relationships and the role of social platforms in this process.

2.4 Main Data Sources

In this section, we delve into the main platforms analyzed in studies on music virality to answer RQ3 (*"What are the main domains, i.e., media platforms, in which viralization processes occur?"*). As discussed in the previous section, there are studies that consider music virality from a technological perspective, that is, as a phenomenon shaped by platforms and social media. Such platforms, which range from video-sharing sites to social networks and streaming services, offer different mechanisms and characteristics that influence how a song goes viral. Here, we explore some of the main social platforms frequently addressed in the studies considered by this review, highlighting their specificities and impacts on music consumption and the music industry itself.

YouTube. Among the studies included in this review, YouTube is the most studied platform in the context of music viralization. Created in 2005, the platform was a pioneer in democratizing access to video production and consumption by allowing any user to upload content and reach a global audience. One of YouTube's main features is that it allows you to add a visual dimension to music, something that often influences the virality of content [111]. Regarding metrics for virality, the number of views on the platform is the most used among the studies considered in this review [9, 17, 51, 93, 118], being used as one of the ways to measure the reach of videos in different contexts.

In addition, YouTube allows people to express opinions and thoughts about videos through comments. In this context, Sharma and Pandey [99] and Sharma et al. [100] analyze such comments alongside views to identify the impact of the electronic word-of-mouth (eWOM) phenomenon on music sales. Lyrics and other external elements are used by March [55] to analyze how the music video for "Friday (Remix)" by Rebecca Black rebrands the singer's image and career after her original viral song in 2011. Other qualitative studies perform case studies and analysis of other aspects of music videos on the platform to understand the viral phenomenon better [23, 26].

Blogs. Before the popularization of social networks, blogs were the main spaces where people could share content and opinions. Indeed, the first studies identified in this review considered blogs important sources for discovering new music and artists and, consequently, for the viralization of such content. For example, Wallsten [118] reinforces the role of bloggers in political campaigns by convincing people to watch a music video. Similarly, Sharma and Pandey [99] and Sharma et al. [100] consider blogs as primary sources of eWOM, thus impacting music sales. Finally, Cha et al. [17] build a network of blog posts for extracting the social relationship between blogs and its relation to music spreading.

TikTok. In recent years (specifically after the COVID-19 pandemic), TikTok emerged as a social network that has significantly boosted the phenomenon of music virality. With its short, highly shareable video format, the platform allows users to create and share content in a quick, easy, and creative way, facilitating the emergence of challenges or viral memes that boost the popularity of songs. As a result, unknown artists can achieve success almost instantly if their songs become the soundtrack to popular trends on the platform. In this sense, Coulter [20] analyzes in depth how the platform differentiates itself from others in promoting new music releases. In addition, Biasioli [10] uses a case study to analyze phenomena adjacent to the viralization of a song, such as the removal of meanings from the song by users in favor of their self-expression, the memefication of a certain musical style, and the participation of the author in this process.

Music consumption platforms. In addition to the social platforms where music is disseminated, considering how music is distributed and consumed is also a relevant factor when studying the phenomenon of musical virality. Before the popularization of streaming services, downloads were the primary way of obtaining music for consumption. In this regard, Nika et al. [60] and Rosati et al. [89] use the number of music downloads on platforms such as BitTorrent and MixRadio to model musical virality (sometimes described as popularity) as an epidemiological phenomenon. As they became more popular, streaming platforms, such as Spotify, started to deal with music viralization, creating specific rankings of the most viral songs. Such rankings are used by Oliveira et al. [73], who use a quantitative methodology to identify the factors differing viral from hit songs.

Other platforms. Other social networks have also played important roles in music viralization and have been used in several studies to understand this phenomenon. Among these platforms, Myspace was one of the pioneers in promoting music directly to fans, and the studies by Sharma and Pandey [99] and Sharma et al. [100] use posts and listening records on the platform as relevant factors to understand the eWOM phenomenon. Similarly, listening data from Last.fm is also considered to model social influence and product adoption by Barbieri and Bonchi [6]. In addition, Tinati et al. [114] uses Wikipedia as a primary data source, based on the premise that access to its articles could reflect human activity and can therefore be used to measure trends in the music scene. Finally, publications on other popular social networks, such as Twitter and Facebook, are also considered in some studies on disseminating musical content online [26, 49].

Figure 2.5 presents an overview (left) and the temporal evolution (right) of the platforms considered in the studies of this review. Throughout the period analyzed, YouTube was the most used platform in studies of music virality, and it has been constantly used to this day. This reflects that, despite losing space as the main platform for content viralization, it is still highly relevant for online music consumption and is crucial in maintaining

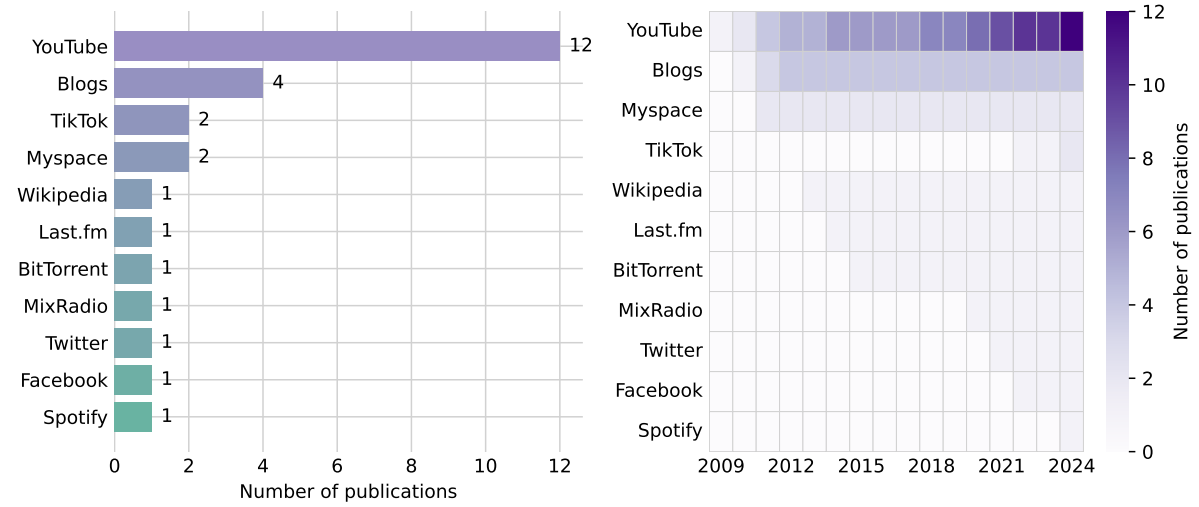


Figure 2.5: Number of publications (left) and temporal evolution (right) by data source. Note that there are publications that consider more than one data source.

its virality. In addition, the temporal evolution of the sources also reflects the adoption and decline of social networks over time. For example, while Myspace was used in the early years of research on music virality, TikTok emerged only recently, coinciding with its popularization. Thus, using different social platforms reflects the complexity of music virality, enabling the application of different quantitative and qualitative techniques.

2.5 Approaches and Methods

We now move into the final research question, RQ3 (*"What are the main approaches used to understand and model viral processes?"*), in which we analyze the methodology of the works assessing music virality. In this review, we evaluate the works in two dimensions: the approach and the methodology used. The first assesses whether the work uses a more theoretical and conceptual approach or follows a practical and experimental line. In contrast, the latter evaluates explicitly the methods used, classifying the works as either quantitative or qualitative methodology. Figure 2.6 and Table 2.1 present the works considered in this review according to the two dimensions mentioned. The joint analysis of such dimensions allows the identification of four categories used in the works' classification. Each of them is detailed next.

Conceptual Explorations. The first group comprises works that analyze theories and concepts related to musical virality through case studies or content analysis without re-

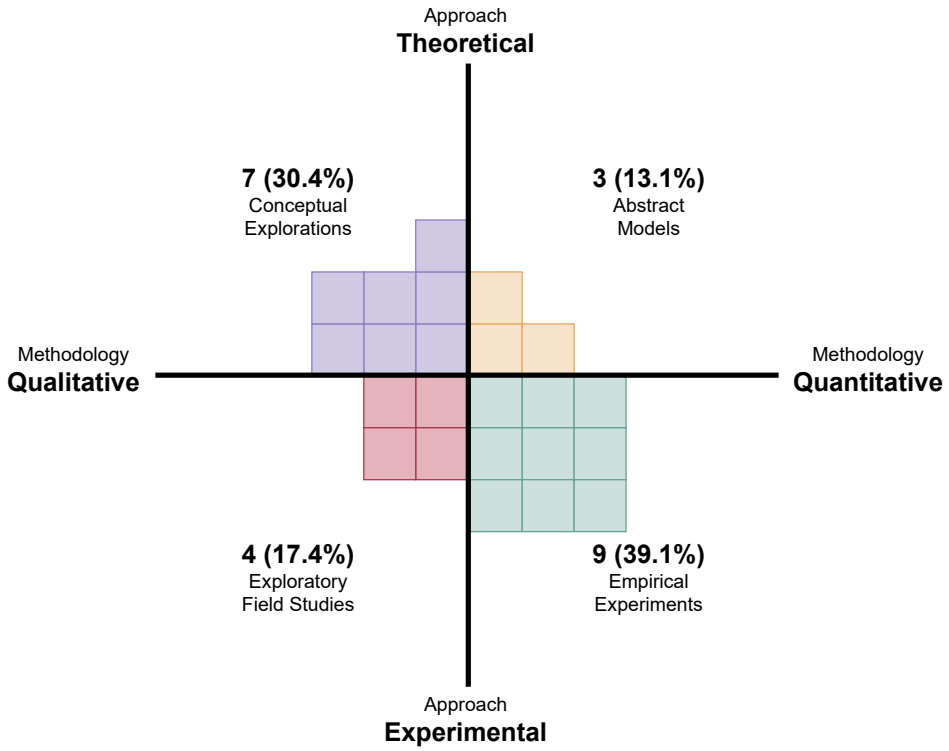


Figure 2.6: Classification of the works considered in this review regarding their approaches (i.e., theoretical or experimental) and methodology (i.e., quantitative or qualitative).

Table 2.1: Studies considered in this review according to their approach and methodology.

Conceptual Explorations *Theoretical + Qualitative*	**Abstract Models** *Theoretical + Quantitative*	**Exploratory Field Studies** *Experimental + Qualitative*	**Empirical Experiments** *Experimental + Quantitative*
Biasioli [10]	Li and Shao [51]	Kahl and Albers [39]	Barbieri and Bonchi [6]
Coulter [20]	Nika et al. [60]	March [55]	Baños-Gonzalez et al. [9]
Edmond [22]	Sachak-Patwa et al. [93]	Nwagwu and Akintoye [61]	Cha et al. [17]
Eromosele [23]		Tan [111]	Oliveira et al. [73]
Freitas [26]			Rosati et al. [89]
Kahl [38]			Sharma and Pandey [99]
Lehman [49]			Sharma et al. [100]
			Tinati et al. [114]
			Wallsten [118]

lying on numerical data. In addition, they can develop theories about qualitative factors regarding this phenomenon, including emotions and other cultural aspects. For example, Kahl [38] conducts literature research to analyze how the characteristics of digital music relate to viral marketing strategies. In contrast, the works of Edmond [22], Eromosele [23] and Biasioli [10] use case studies to perform conceptual analyses on changes in the music industry, user responses, and the processes behind the circulation of viral content on digital platforms, respectively. Other studies in this category analyze references in popular songs [49], the structure of specific platforms such as TikTok [20], and the production of memes [26] to discuss concepts about the phenomenon of music viralization.

Abstract Models. In the context of musical virality, theoretical studies with a quantitative methodology propose theoretical models or conceptual explanations using metrics

to study the phenomenon in question. In particular, these studies develop epidemiological models for music viralization based on the premise that such a phenomenon can be interpreted as a contagion process. For example, Nika et al. [60] propose a model for the spread of online content as a set of multiple epidemics, combining classic models such as SIR (Susceptible-Infected-Recovered) and IR (Infected-Recovered). Moreover, Sachak-Patwa et al. [93] develop a SEIRS (Susceptible-Exposed-Infected-Recovered-Susceptible) model to represent the evolution of the popularity of viral videos. Finally, Li and Shao [51] modify the classic SIR model to incorporate the influences of sharing and advertising in videos that go viral. Although all such studies also present validation with real data, their focus relies on proposing a model, thus justifying their classification as theoretical.

Exploratory Field Studies. This category refers to practical and experimental studies that explore the phenomenon of music virality through qualitative data, such as interviews and observations. For example, Kahl and Albers [39] use interviews with several people in the music industry to assess the critical factors for the viral marketing of digital music. Furthermore, Nwagwu and Akintoye [61] interview Nigerian artists to understand how social platforms are used to disseminate their productions. In contrast, the works of Tan [111] and March [55] use content and discourse analysis to analyze how the viral phenomenon relates to government bodies and artists' public image, respectively.

Empirical Experiments. The last group comprises research with a practical approach that uses numerical data and clear metrics to test or validate concepts and hypotheses in a real scenario. Studies in this category use a wide variety of techniques, but statistical approaches are the most common. For example, Wallsten [118] uses a vector autoregression model and Granger Causality to examine the relationship between variables such as number of views, blog mentions, political campaigns, and media coverage on the viralization of a music video. In contrast, the works of Sharma and Pandey [99] and Sharma et al. [100] use the so-called Object-Oriented Theoretical Framework to model the effect of electronic word-of-mouth on music sales. Specifically, both apply multivariate linear regression to correlate consumer actions across platforms with sales rankings.

Furthermore, other studies explore several other methods according to their objective. For instance, Cha et al. [17] model a network to represent the relationship between blogs through HTML links and evaluate the diffusion of musical content through them. The work of Tinati et al. [114] uses time series analysis and correlation to measure viral trends from user activity on Wikipedia. In turn, Barbieri and Bonchi [6] treat the viral adoption process of music as an optimization problem in viral marketing by addressing social influence and product design characteristics. Then, Baños-Gonzalez et al. [9] perform a content analysis to explain the viral capacity of music videos, while Rosati et al. [89] perform a comparative analysis between epidemiological and phenomenological models to

Table 2.2: Overview of the papers considered in this review.

Year	Ref.	Venue Field	Perspective	Mechanism	VxS	Platform	Approach	Methodology
2009	[111]	Music	Epidemiological	Organic		YouTube	Experimental	Qualitative
2010	[118]	Politics	Sociocultural	Both		YouTube, Blogs	Experimental	Quantitative
2011	[99]	Multidisciplinary	Marketing	Induced	✓	YouTube, Blogs, Myspace	Experimental	Quantitative
2011	[100]	Multidisciplinary	Marketing	Induced	✓	YouTube, Blogs, Myspace	Experimental	Quantitative
2012	[17]	Computing	Sociocultural	Organic		Blogs, YouTube	Experimental	Quantitative
2012	[38]	Computing	Marketing	Induced		N/A	Theoretical	Qualitative
2013	[39]	Computing	Marketing	Induced		N/A	Experimental	Qualitative
2013	[114]	Computing	Sociocultural	Organic		Wikipedia	Experimental	Quantitative
2014	[6]	Computing	Marketing	Induced	✓	Last.fm	Experimental	Quantitative
2014	[22]	Communication	Technological	Organic	✓	YouTube	Theoretical	Qualitative
2015	[60]	Computing	Epidemiological	Organic		BitTorrent	Theoretical	Quantitative
2018	[93]	Physics	Technological	Organic		YouTube	Theoretical	Quantitative
2020	[9]	Communication	Sociocultural	Organic	✓	YouTube	Experimental	Quantitative
2021	[23]	Culture	Sociocultural	Organic		YouTube	Theoretical	Qualitative
2021	[89]	Math, Physics, Engineering	Epidemiological	Organic		MixRadio	Experimental	Quantitative
2021	[49]	Multidisciplinary	Epidemiological	Organic		Twitter	Theoretical	Qualitative
2022	[20]	Business	Technological	Both	✓	TikTok	Theoretical	Qualitative
2022	[26]	Music	Sociocultural	Organic	✓	YouTube, Facebook	Theoretical	Qualitative
2023	[61]	Info. Science	Technological	Organic		N/A	Experimental	Qualitative
2024	[51]	Multidisciplinary	Technological	Induced		YouTube	Theoretical	Quantitative
2024	[55]	Communication	Technological	Organic	✓	YouTube	Experimental	Qualitative
2024	[10]	Music	Technological	Organic	✓	TikTok	Theoretical	Qualitative
2024	**[73]***	**Computing**	**Sociocultural**	**Organic**	✓	**Spotify**	**Experimental**	**Quantitative**

VxS: The paper differs the concepts of virality and success. * This thesis.

explain the viralization of music through downloads, demonstrating the better efficiency of the former. Finally, Oliveira et al. [73] quantitatively investigate which characteristics are important to differentiate virality and musical success.

2.6 Discussion and Overall Considerations

This section discusses the results of the analyses performed in this literature review on music virality. Following a research protocol that defines the research questions,

search repositories, and exclusion criteria, we considered 23 peer-reviewed articles in English published in journals and conferences from different areas of knowledge. The research on music virality was then analyzed under four aspects: temporal evolution, perspectives and relationship with success, platforms considered, and methodology. Table 2.2 summarizes the works considered in this review and their main characteristics according to the aforementioned analyses.

Research on music virality has been recurrent since the first decade of the 21st century. With the popularization of the Web and social networks, content dissemination has become one of the main phenomena observed on such platforms. As it involves concepts from areas such as music, computing, and communication, analyses on music virality have a strong multidisciplinary character. Despite being relevant to Computing, this is reflected in the diversity of knowledge fields of the venues of the publications in this review. When analyzing the temporal evolution of the research, we also identify two periods of high increase in publications. Such periods coincide with the rise and consolidation of platforms such as YouTube and TikTok as protagonists in music virality. Both platforms have in common the fact that they allow users to share videos quickly and easily.

Indeed, such platforms appear on the list of the most used in this review. YouTube is by far the most used, as it is the main place where music videos are posted. The viral effect makes these videos widely disseminated online, reaching millions or even billions of views. In recent years, TikTok has revolutionized video sharing and reshaped the phenomenon of music viralization. Such a platform allows excerpts of songs to go viral as part of challenges and viral trends. However, other platforms and data sources are also relevant to understanding this phenomenon. In particular, blogs were widely used in the early years of research on music virality because they were important places for sharing content and opinions. Download and streaming platforms are also relevant in this context because they represent music consumption itself.

Music consumption through streaming, downloads, and sales can be one of the ways to measure musical success Seufitelli et al. [98]. Despite being closely related concepts, in this work, we consider virality and success as distinct yet interconnected facets of musical popularity. Regarding this review, although not all studies make this distinction, a significant part of them contrasts the two concepts. Whereas all of them relate virality to rapid sharing on online platforms, success is more solid and linked to music consumption itself. The relationship between both concepts becomes clearer when we consider that the viralization of a song can also be an initial step towards its subsequent success, with viral trends potentially causing a considerable increase in streams.

Finally, regarding the main methodologies used, the existing studies are divided between qualitative and quantitative analyses, thus performing conceptual and content analyses, as well as experimental studies. In the field of Computing and other exact

sciences, a group of studies stands out for proposing and using contagion models to represent the phenomenon of viralization on social networks. Such studies are based on the definition of virality and the fact that the spread of online content can be understood as an epidemic. The main advantage of these models is the interpretability of their parameters, which is fundamental for understanding this phenomenon in depth. However, to the best of our knowledge, none of these studies explicitly consider the new reality of music consumption and sharing, i.e., the use of streaming platforms and short video sharing.

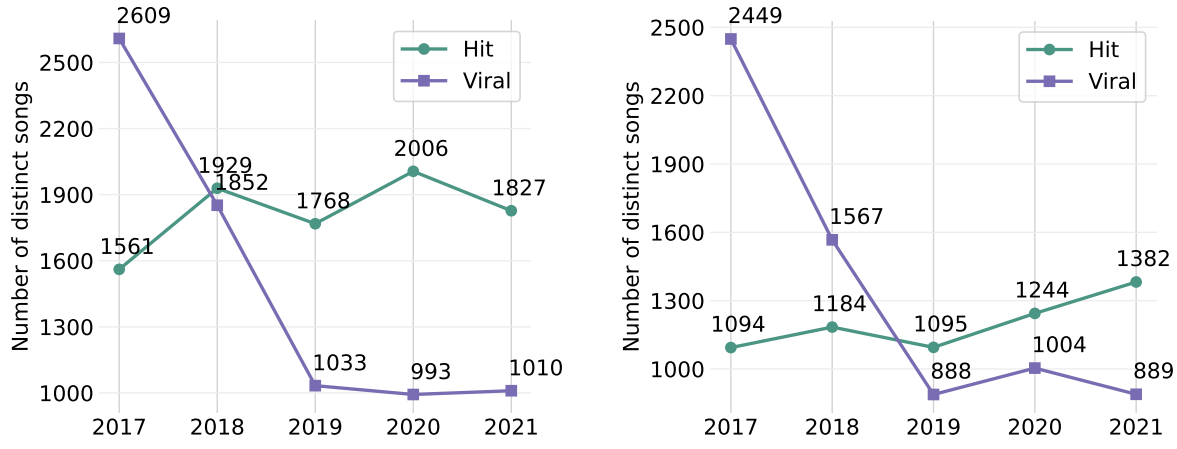
This thesis' contributions. Although the existing studies on musical virality have already elucidated several points about this phenomenon, there are still many open topics on this subject. For example, what makes a viral song? What differentiates a viral from a hit song? What is the temporal relation between virality and success? Do viral songs always become hits? How do viral songs spread in a post-pandemic world? In this thesis, the main goal is to answer this and other questions about musical virality from a computer science perspective. Thus, we analyze virality as a sociocultural phenomenon, using relevant but still little-explored social platforms, such as Spotify and TikTok. Unlike previous works, this thesis innovates by using a quantitative-experimental approach rooted in streaming data to differentiate viral from hit songs and to establish their temporal and causal relation. Furthermore, we also introduce a novel wave-based model for music virality on social platforms. By integrating machine learning, time series analysis, and epidemiological techniques, we provide a unique data-driven framework to unveil the dynamics of virality and mainstream success.

Chapter 3

A Quantitative Characterization of Viral and Hit Songs

Analyzing music popularity is not a trivial task since there is not a singular, universally accepted definition for popularity. A popular song may have a high number of streams but it may also be subject of extensive discussions (online or offline) [98]. Further, not all popular songs (or hits) go viral, and going viral does not necessarily mean becoming a hit. Still, with new social platforms changing how people consume music, virality has taken a key role in music popularity. Following such a transformation, streaming platforms have started producing distinct rankings for viral and hit (i.e., successful) songs. Specifically, charts from Brazilian and worldwide Spotify reveal different trends: the number of distinct viral songs decreases over time, while the number of hits increases (Figure 3.1).

In this chapter, our goal is to answer the following questions: *"Are viral and hit songs two sides of the same coin? What differs them?"* Hence, we perform a comparative analysis of hit and viral songs, as extracted from their respective charts on Spotify, to unveil similar and distinct patterns in both Global and Brazilian markets. It is important to notice that we do not seek to assess how Spotify builds up its ranks, nor do we intend to verify their accuracy. Still, by the time we started to work on such research questions



(a) Global (b) Brazil

Figure 3.1: Number of distinct songs in Spotify Charts per year (2017–2021).

(mid 2023), Spotify was the most popular music streaming platform that also published music charts online (i.e., easily collectible). We further understand that there are other platforms available (as already mentioned). Still, none of them had so many options of charts (global and local, hits and virals) that we could explore.

We achieve such a goal through a quantitative and data-driven methodology, in which we analyze an enhanced set of song features extracted from Spotify and Genius, two of the most relevant online music-related platforms (Section 3.1). Besides an exploratory characterization of intrinsic and extrinsic features (Section 3.2), we also aim to automatically classify whether a song is a hit or a viral given it has already reached a popularity status. To do so, we evaluate three research hypotheses related to which musical features have more influence in such distinction (Section 3.3). Distinguishing viral from hit songs may help to uncover the underlying factors behind the viral phenomenon and its specific diffusion process, in addition to understanding users' behavior when it comes to consuming and sharing music with others. After presenting a discussion on the results (Section 3.4), we present our overall considerations (Section 3.5).

3.1 Data Acquisition

In this section, we first describe how we built the dataset of hit and viral songs (Section 3.1.1). Then, we present the set of features applied to investigate potential differences among songs concerning success and virality (Section 3.1.2).[1]

3.1.1 Dataset

The definition of musical success may vary according to the research goal and the data sources [98]. In this thesis, we consider musical success and virality from a chart-based perspective; i.e., songs that manage to enter the success charts are considered popular. We evaluate hit and viral song charts from Spotify, one of the most used streaming platforms in the world, comprising more than 551 million users in 184 markets.[2] Spotify has charts for each market it is present in, as well as the aggregated global

[1]Note that our intention is not to reverse engineer any existing ranking mechanism, nor do we aim to predict which songs will become hits or go viral. In fact, our goal is to characterize such songs to better understand the differences between hits and viral songs.

[2]Spotify (Oct. 2023): https://investors.spotify.com/about/

charts. We consider Global and Brazilian charts in our analyses to contrast what is being listened to globally and regionally. Additionally, we use Genius[3] to augment our dataset with metadata and song lyrics. We describe each of the sources as follows.

Hit Songs. Our primary source for hit song data is the Music Genre Dataset (MGD+) [97], an open structured dataset that gathers and enhances data obtained from Spotify. From MGD+, we obtain the Top 200 daily charts (i.e., the most streamed songs) from January 2017 to March 2022. Therefore, we consider as hits all songs that reached the Top 200 charts. MGD+ also provides important features related to such songs and their artists, including acoustic features, release dates, and artists' musical genres.[4]

Viral Songs. Besides MGD+, we aggregate the Viral 50 daily charts (i.e., the most viral songs) for the same period. Similarly to hit songs, we label all songs that entered the Viral 50 charts as viral. According to Spotify, such charts capture the songs gaining the most buzz on the platform, as they are calculated by considering the rise in the plays, the number of shares, and the number of people who have recently discovered the song.[5] We then use the Spotify Web API[6] to collect the same features and metadata that MGD+ contains about hit songs. All data was last updated in March 2022, when the Spotify Charts platform closed their free access for download.

Lyrics and Metadata. We use Genius to collect the lyrics and metadata for the viral and hit songs. In short, Genius is a collaborative platform to disseminate music knowledge in which users share facts and insights about songs and artists. It is mainly known as a platform for obtaining song lyrics and annotations (i.e., explanations of the context and meaning of such lyrics). Users add all content, but editors use moderation to ensure the quality of the information and to certify specific content.

Since there is no connection between Spotify and Genius data, we perform a data integration step. Specifically, we use the Genius Search with the song title and artists from Spotify to find the corresponding record ID. Then, we run a web crawler to obtain the lyrics. Additionally, we use the Genius API[7] to gather further metadata on the songs, including their language and whether a song samples, remixes, or covers another song.

Genius does not provide information for all songs in the dataset, as some songs have not been added to the platform. Hence, we include only the songs with lyrics and metadata from Genius in the final dataset, from now on called MGD++ to better distinguish our enhanced version from the original MGD+. Table 3.1 presents the general statistics for

[3]Genius: https://genius.com/

[4]The Spotify API (and, by extension, MGD+) offers a list of music genres for each artist rather than for individual songs.

[5]Spotify: https://support.spotify.com/us/artists/article/charts/

[6]Spotify API: https://developer.spotify.com/

[7]Genius API: https://docs.genius.com/

Table 3.1: Dataset overview statistics.

	Global			Brazil		
	Overall	**Hit**	**Viral**	**Overall**	**Hit**	**Viral**
Songs	11,648	7,156 (61.4%)	6,578 (56.4%)	7,171	3,783 (52.7%)	4,963 (69.2%)
Artists	5,005	1,977 (39.5%)	4,505 (90.0%)	3,387	1,322 (39.0%)	3,139 (92.7%)
Languages	47	27 (57.4%)	45 (95.7%)	26	12 (46.2%)	26 (100%)

our dataset splited into hits and virals. Note that the sum of the percentages of hits and virals is not 100%, since there is an intersection between both sets. Specifically, 2,086 songs reach both hit and viral charts in the Global market, representing 29.1% and 31.7% of hit and viral songs, respectively. In Brazil, there are 1,575 songs in the intersection, i.e., 41.6% of the hits and 31.7% of the virals. Hence, they are counted for both sets.

3.1.2 Set of Features

To investigate whether there are specific features that distinguish viral songs from hit songs, we focus on two groups of features based on their relationship with the song, following the taxonomy proposed in [98]. The first group, called ***intrinsic features***, considers the song's metadata, acoustic and lyrics-related features. The second group, called ***extrinsic features***, focuses on artists and chart-related data. Next, we briefly describe such features.

Metadata. Here, we select two types of information. The first one regards the relationships of song with respect of previously released songs, i.e., whether a song contains a *sample*, is a *remix* or a *cover*.[8] Such versions are becoming widely used in viral videos on TikTok and Instagram.[9] The second information we are interested in is the music genre. Since Spotify's genres are associated with artists rather than individual songs, we associate each song with all genres attributed to the artists who sing it.

Acoustic Features. Provided by the Spotify API, these features rely on musical data extracted from the audio properties (e.g., pitch, rhythm, dynamics, and timbre). Table 3.2 lists all such features used in our analyses.

[8]In short, sampling involves taking a portion of an existing song and using it as part of a new one; remixing is altering a song to create a new version of it; and a cover is a new performance of an existing song. Such aspects are not mutually exclusive, as a song may present one or more of them at the same time (i.e., being a remix and containing a sample from another song).

[9]NBC News: https://nbcnews.to/47Q2YO1

Table 3.2: Acoustic features considered in this thesis.

Feature	Type	Description
acousticness	Float	Probability of a song being acoustic or not
danceability	Float	Probability of a song being suitable for dancing
duration_ms	Integer	Duration of a song in milliseconds
energy	Float	Intensity and activity of a song in terms of perceived loudness, timbre, and general entropy
instrumentalness	Float	Probability of a song being instrumental, i.e., without vocals
key	Integer	Estimated overall key of a song, mapped as an integer number (e.g., $C = 0$, $C\# = 1$, and so on)
liveness	Float	Probability of a song being performed live, i.e., the presence of an audience in a song
loudness	Float	General loudness measured in decibels (dB)
speechiness	Float	Probability of a given song having spoken words in it
tempo	Float	Speed of the song, measured in beats per minute (BPM)
time_signature	Integer	Amount of beats in each bar (measure)
valence	Float	Positiveness of a song (in which high valence values represent happier songs, whereas low values, the opposite)

Lyrics-related Features. Lyrics are frequently used in music-related research for evaluating rhyme and text. We analyze four types of lyrics-related features: (i) *General characterization*, with the number of words, lines, and verses; (ii) *Language*, a categorical feature for the language of a song; (iii) *Main topics*, extracted by using the Latent Dirichlet Allocation (LDA) algorithm [14]; and (iv) *Psycholinguistic features*, extracted by using Linguistic Inquiry and Word Count (LIWC) [112], which assigns words within a given text to linguistic and psychological dimensions (e.g., emotions, word categories, slangs, and so on).

Artist-related Features. Analyzing the artists who sing a given song may reveal important dimensions of a song's virality and popularity. We focus on artist collaboration (i.e., when two or more artists are involved in a song), a specific dimension which previous studies directly relate to musical success [11, 103].

Temporal Features. Regarding song popularity, research studies usually consider information such as position in charts to grasp the level of success. We consider Spotify's charts to derive two temporal features: the time from the songs' release until they reach the charts for the first time, and the time they spend on the charts. Both features are measured in days. Note that the period of the last feature is not necessarily continuous. For example, if a song stays ten days on the charts, leaves, and then re-enters for five days, the time it spends on the charts is 15 days.

Table 3.3: Proportion of song relationships with other pre-existing songs.

	Global				Brazil			
	Hit		Viral		Hit		Viral	
	Songs	%	Songs	%	Songs	%	Songs	%
Sampling	1,106	15.46%	697	10.60%	342	9.04%	460	9.27%
Remix	391	5.46%	376	5.72%	190	5.02%	280	5.64%
Cover	152	2.12%	189	2.87%	81	2.14%	155	3.12%

Table 3.4: Top 10 most frequent music genres of artists whose songs are hit or viral on Spotify Global.

Hit			Viral		
Genre	**Songs**	**%**	**Genre**	**Songs**	**%**
pop	2,729	38.14%	pop	1,127	17.13%
rap	2,293	32.04%	rap	1,064	16.18%
hip hop	1,397	19.52%	hip hop	701	10.66%
trap	1,261	17.62%	trap	612	9.30%
melodic rap	799	11.17%	pop rap	407	6.19%
atl hip hop	688	9.61%	urbano latino	358	5.44%
pop rap	678	9.47%	melodic rap	328	4.99%
urbano latino	629	8.79%	trap latino	324	4.93%
trap latino	611	8.54%	reggaeton	316	4.80%
reggaeton	588	8.22%	r&b	300	4.56%

3.2 Feature Characterization

We now analyze the intrinsic features of the songs: metadata (Section 3.2.1), acoustic (Section 3.2.2), and lyrics-related (Section 3.2.3). We also characterize the songs' extrinsic features: artist-related (Section 3.2.4) and chart-related ones (Section 3.2.5).

3.2.1 Metadata

We first verify which songs contain samples from previously released songs, and the ones that are remixes or covers. Table 3.3 presents the percentage of hit and viral songs associated with pre-existing songs. Overall, the majority of hit or viral songs in the Global market are original, with only 15.46% and 10.6% of them, respectively, incorporating elements from pre-existing songs. Brazil is similar, with 9.04% and 9.27% of the hit and viral songs sampling other previous songs. Moreover, there are no substantial differences between hit and viral songs regarding the percentage of remixes and covers as well.

We now look into the music genres of hit and viral songs. Tables 3.4 and 3.5

Table 3.5: Top 10 most frequent music genres of artists whose songs are hit or viral in Spotify Brazil.

Hit			Viral		
Genre	**Songs**	**%**	**Genre**	**Songs**	**%**
pop	1,451	38.36%	pop	1,116	22.49%
sertanejo universitário	739	19.53%	rap	486	9.79%
arrocha	670	17.71%	funk carioca	450	9.07%
funk carioca	579	15.31%	pop nacional	374	7.54%
pop nacional	542	14.33%	hip hop	340	6.85%
sertanejo	503	13.3%	brazilian hip hop	335	6.75%
agronejo	337	8.91%	sertanejo universitário	302	6.09%
dance pop	325	8.59%	dance pop	292	5.88%
rap	319	8.43%	k-pop	266	5.36%
funk rj	255	6.74%	arrocha	251	5.06%

present the top 10 most frequent genres of artists who sing both hit and viral songs in the Global and Brazilian markets, respectively. We can outline some conclusions. On the global scenario, there is a significant similarity between the most frequent genres in hit and viral songs. For example, pop, rap, hip hop, and trap (as well as their subgenres) dominate both rankings. There is also a strong presence of Latin genres in the songs, such as *urbano latino*, *trap latino* and *reggaeton*.

In Brazil, in terms of similarities, pop and its variants (*pop nacional* and dance pop) appear in both rankings. This follows a global trend [63], establishing *pop* artists as the prevailing ones among hit and viral songs. Regional genres also play a key role in the Brazilian market, with a strong presence of *sertanejo* (*sertanejo universitário*, *agronejo*), Brazilian funk (*funk carioca*, *funk rj*) and *arrocha*.[10] Indeed, it is common that the Top 10 hit songs in Spotify Brazil are composed entirely (or almost) of songs of artists belonging to these three main genres.

When comparing hits versus viral songs, hip-hop (and its subgenre Brazilian hip-hop) is among the most prominent genres in virals, indicating widespread sharing of songs from artists belonging to this genre. For instance, Emicida (one of Brazil's most influential hip-hop artists) boasts 13 viral chart entries but only three hits. Similar patterns happen in K-pop, with 95 artists in viral charts, contrasting with 16 in hit charts.

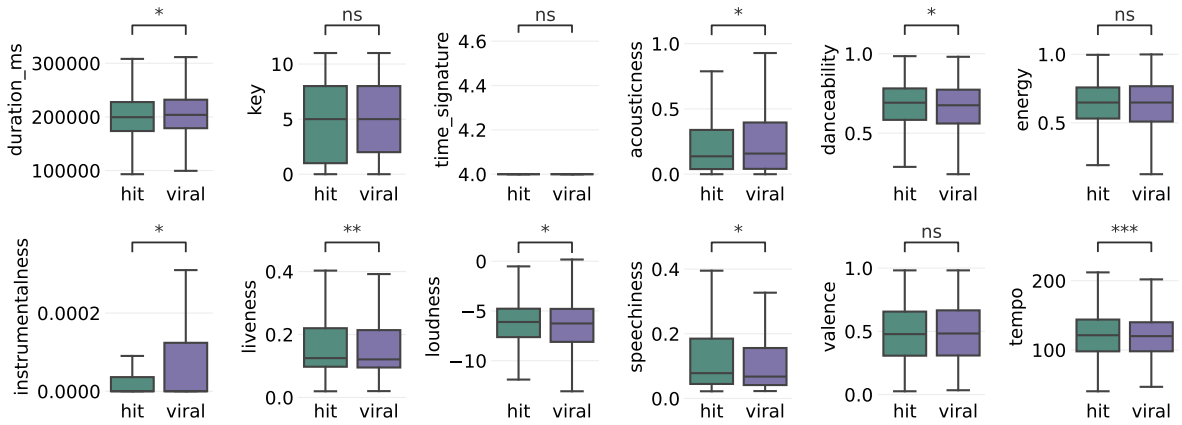


Figure 3.2: Boxplots with the distribution of acoustic features' values for hit and viral songs in the Global market. Significance levels of the Mann-Whitney U test: * for $p < 0.001$; ** for $p < 0.01$; *** for $p < 0.05$; and 'ns' otherwise.

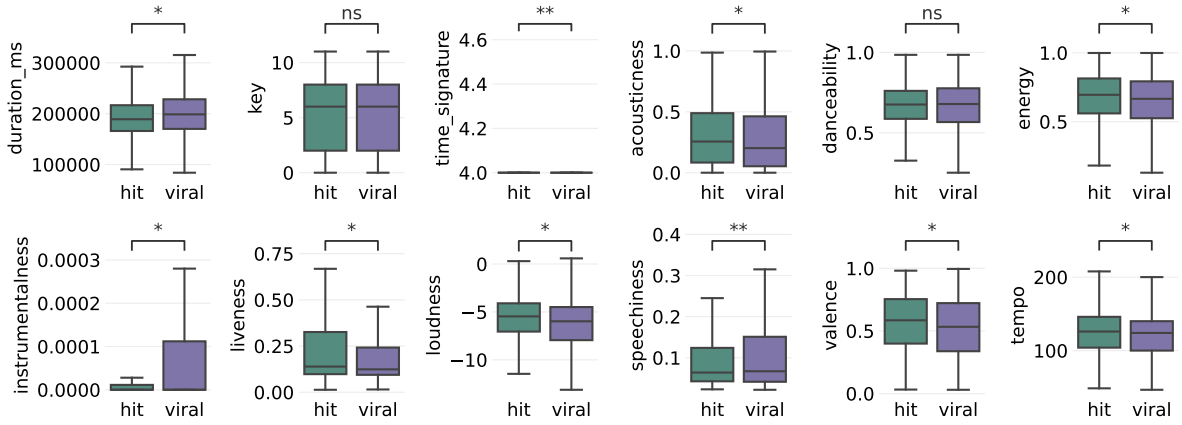


Figure 3.3: Boxplots with the distribution of acoustic features' values for hit and viral songs in the Brazilian market. Significance levels of the Mann-Whitney U test: * for $p < 0.001$; ** for $p < 0.01$; *** for $p < 0.05$; and 'ns' otherwise.

3.2.2 Acoustic Features

Since the acoustic features provided by Spotify are all represented by numeric values (see Table 3.2), we analyze the distribution of each one to compare hit and viral songs. Figures 3.2 and 3.3 illustrate this comparison in the Global and Brazilian markets through boxplots.[11] In addition, we also perform a two-sided Mann-Whitney U test [54] to verify the statistical significance of the difference between the distributions for hit and

[10] *Arrocha* is a music genre originated in the Northeast region of Brazil that is closely related to genres such as *forró* and *axé*.

[11] Boxplots visually represent numerical data distribution, skewness, and key summary statistics: minimum score (bottom whisker), lower quartile (25% below the filled area), median (mid-point line), upper quartile (25% above the filled area), and maximum score (upper whisker).

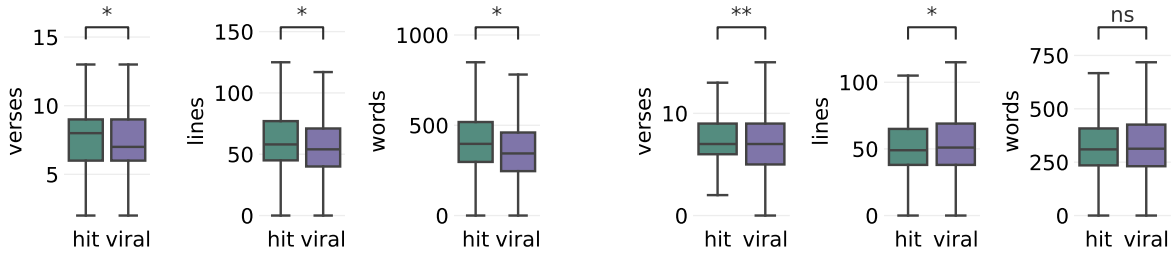
viral songs. In such a test, our null hypothesis is that the distribution underlying the two sets of values is the same. If the p-value of the test is less than a predefined threshold, the null hypothesis is rejected, suggesting a significant difference between the groups.

There are no large differences between hits and virals, but most features present some statistical differences. Such data variance provides valuable insights into the disparities that exist between hit and viral songs, shedding light on the factors that set them apart. For example, hit songs tend to have a shorter duration on average compared to viral tracks. Indeed, songs have been getting shorter in the last years. Songs with less than three minutes are not uncommon.[12] For the sake of illustration, "As It Was" by Harry Styles and "Mal Feito - Ao Vivo" by Hugo & Guilherme and Marília Mendonça (i.e., the most listened songs on Spotify globally and in Brazil in 2022) have 2 minutes and 47 seconds, and 2 minutes and 57 seconds, respectively.

On the other hand, hit songs have higher values for acousticness and liveness than viral songs in Brazil, meaning that there are more hits with higher probabilities of being acoustic and being performed live. Such a result may seem counterintuitive at first, as only (or mostly) studio versions are expected to reach the hit status. However, some popular genres in Brazil (such as *sertanejo* and *pagode*) have a specific behavior in which the live versions of songs are the most consumed by listeners. Indeed, the most streamed song in 2022 (i.e., "Mal Feito - Ao Vivo" by Hugo & Guilherme and Marília Mendonça) is a live version. This may be a particular characteristic of the country, which may not be reflected worldwide. It also corroborates previous work that emphasizes the importance of analyzing regional markets individually, as each one has its own patterns [63].

Last, features such as energy, loudness, valance, and tempo reveal that hits are more energetic, more positive, louder, and faster than virals in Brazil. The same happens for loudness and tempo in the Global market. Such characteristics are often tailored to have broad mass appeal and are more likely to resonate with a large and diverse audience. Viral songs may span a broader range of styles, some of which may not emphasize these qualities as much. Moreover, speechiness values reveal that viral songs have, in general, a higher probability of having spoken words than hits in Brazil. For instance, chill songs that become a trend in TikTok videos and Instagram Reels are widely shared to become viral, but they are not massively streamed to reach the hit charts. An example is the song "Boho Days" of the soundtrack from the movie *tick, tick... BOOM!* (2021), which is currently the third most streamed song from the OST in Spotify.

[12] Vice (Oct. 2023): https://bit.ly/3SakvMc



(a) Global (b) Brazil

Figure 3.4: Distribution of the number of verses, lines, and words for hit and viral songs. Significance levels of the Mann-Whitney U test: * for $p < 0.001$; ** for $p < 0.01$; and 'ns' otherwise.

Table 3.6: Most frequent languages in hit and viral songs in the Global market. The category "Other" includes songs with unknown language.

Hit			Viral		
Language	**Songs**	**%**	**Language**	**Songs**	**%**
English	5,703	79.70%	English	4,649	70.67%
Spanish	611	8.54%	Spanish	542	8.24%
German	258	3.61%	Portuguese	253	3.85%
Portuguese	161	2.25%	Korean	220	3.34%
Other	423	5.91%	Other	914	13.89%

3.2.3 Lyrics-related Features

Here, we consider four dimensions of song lyrics: general features, languages, main topics of the lyrics, and their psycholinguistics. We perform such analyses over the subset of songs with lyrics extracted from Genius (i.e., our reduced dataset, see Section 3.1.1).

General Characterization. Figure 3.4 presents the distribution of the number of verses, lines, and words for hit and viral songs. We perform a two-sided Mann-Whitney U test to check whether the difference between the distributions is statistically significant. The results reveal that in general, hit songs have more verses, lines, and words than virals in the Global market. In Brazil, although the distribution of verses and words is statistically different, its visual analysis does not show significant differences, i.e., the distributions are similar. Nonetheless, it is noteworthy that viral songs have a broader range in such values, yet their median is very close to that of hit songs. In other words, hits are more similar as a set than viral songs.

Language. Tables 3.6 and 3.7 reveal the main languages of hit and viral songs in the Global and Brazilian markets. English is the most prevalent language in the global scenario, accounting for 79.7% of the hits and 70.67% of the virals, respectively. In addition,

Table 3.7: Most frequent languages in hit and viral songs in Brazil. The category "Other" includes songs with unknown language.

Hit			Viral		
Language	**Songs**	**%**	**Language**	**Songs**	**%**
Portuguese	1,808	47.79%	English	2,673	53.86%
English	1,745	46.13%	Portuguese	1,780	35.87%
Spanish	102	2.70%	Spanish	208	4.19%
Korean	101	2.67%	Korean	201	4.05%
Other	27	0.71%	Other	101	2.04%

other frequent languages include Spanish, German, Portuguese, and Korean, highlighting the relevance of Latin genres and K-pop in the Global music market.

Moreover, Portuguese is the most popular language for music hits in Brazil, corroborating the results that show Brazilians tend to consume more local music [63, 96]. As expected, Brazil is also influenced by external markets. Specifically, English comes as the second most popular language in hit songs. The number of hits in such language is almost the same as the songs in Portuguese, which reveals a very strong influence of other Western countries, mainly the United States. Spanish and Korean are the third and fourth most popular languages, reflecting the growing popularity of Latin genres (mostly due to the proximity to other Latin American countries and their popularization in the US) and K-pop, respectively.

Regarding viral songs, the most frequent languages are the same four languages. However, English is, by far, the prevailing language in viral songs in Brazil. This may reflect the influence of short video platforms (e.g., TikTok and Instagram Reels) on the listening habits in the country [8]. A large number of videos posted on such platforms have songs in their background. The more such videos go viral, the more the songs are shared, and people go to Spotify to search and listen to the whole song. A good example of a song that went viral in Brazil following such mechanism is "death bed (coffee for your head)" by Powfu feat. beabadoobee.

Lyrics' Main Topics. Next, we check the lyrics' topics by applying the Latent Dirichlet Allocation (LDA) algorithm,[13] which automatically infers the topics in a set of documents (i.e., the song lyrics). In short, LDA is a probabilistic model that operates by iteratively assigning topics to words in documents and adjusting those assignments based on the observed word-topic and document-topic relationships. Besides the lyrics text, the algorithm also receives a predefined number k of topics as input.

For this analysis, we consider only songs in English and Portuguese, as they represent approximately over 90% of hit and viral songs in Brazil and over 70% in the Global

[13]We use the implementation of the Gensim Python library: https://radimrehurek.com/gensim/models/ldamodel.html

Table 3.8: Most representative terms (sorted by significance) in the topics inferred by LDA for the Global market. Offensive terms and swear words are edited. The translation of Portuguese terms is presented in Appendix A.

	English			Portuguese		
	Topic	**Songs**	**Terms**	**Topic**	**Songs**	**Terms**
Hit	H1	29.51%	yeah, like, b*tch, f*ck, ni**a, sh*t, ni**as, know, back, money	H4	27.33%	senta, toma, pega, pode, novinha, casa, pero, chama, então, bota
	H2	7.29%	christmas, love, smoke, july, thunder, baby, version, february, june, september	H5	36.02%	yeah, aqui, quer, chão, fazer, hoje, bunda, agora, pode, tava
	H3	63.20%	know, love, yeah, like, baby, never, want, time, feel, need	H6	36.65%	amor, gente, coração, cara, sabe, saudade, vida, tudo, porque, beijo
Viral	V1	69.11%	know, love, like, yeah, baby, want, never, time, need, feel	V4	31.62%	senta, toma, amor, tudo, quero, aqui, casa, p*ta, hoje, quatro
	V2	3.97%	young, black, shake, high, love, version, september, june, thunder, life	V5	39.53%	joga, amor, então, gente, quero, rebola, vida, bunda, assim, sabe
	V3	26.92%	yeah, like, b*tch, f*ck, sh*t, ni**a, know, ni**as, back, money	V6	28.85%	yeah, quer, amor, pode, pega, arrasta, cara, tapa, pero, gente

market. To extract the topics, we first remove the songs' stop words[14] and annotations (i.e., indications of who sings each part of the song). We then perform the topic coherence metric [88] to find the best number k of topics. Topic coherence measures how well-defined and semantically meaningful the identified topics are within a given text (i.e., our lyrics). The higher its value, the higher the coherence and interpretability of topics. Hence, we choose $k = 3$ since it produces the higher values for this metric.

Tables 3.8 and 3.9 summarize the LDA output for hit and viral songs in Global and Brazilian markets, with the most representative terms for each topic. In general, both hit and viral English songs present topics related to different facets of romance and love. Terms such as *love* and *baby* appear together in four out of six identified topics in both markets. LDA unveils very specific topics (Topics H1 and V3 in Global and V7 in Brazil) that contain mostly swear words and other explicit terms. Songs within such topics are mostly rap and hip-hop, which are genres already known for having more explicit lyrics.

Topics that characterize hit and viral in Portuguese also share similarities, related to romance and love, containing terms such as *amor*, *coração*, *saudade*, *vida* (Topics H6 in Global and H11, H12, and V11 in Brazil). Such topics prevail in Brazil, corresponding to 79.35% and 63.25% of hit and viral songs, respectively. However, there is also a relevant set of topics that are related to the sexual context (Topics H4 and V4 in Global and H10, V10, and V12 in Brazil). Similarly to rap and hip-hop in English, such themes are very

[14]Stop words are commonly used words in a language that are filtered out before executing Natural Language Processing (NLP) tasks because they are considered to be of little value for such tasks. Examples include articles, prepositions, and conjunctions.

Table 3.9: Most representative terms (sorted by significance) in the topics inferred by LDA for Brazil. Offensive terms and swear words are edited. The translation of Portuguese terms is presented in Appendix A.

	English			Portuguese		
	Topic	**Songs**	**Terms**	**Topic**	**Songs**	**Terms**
Hit	H7	64.61%	know, yeah, love, like, never, baby, time, feel, need, take	H10	20.65%	senta, toma, quer, joga, hoje, chão, bumbum, bunda, então, desce
	H8	27.50%	like, yeah, love, want, baby, come, know, look, back, little	H11	24.42%	yeah, tudo, vida, hoje, então, aqui, deus, quer, amor, sempre
	H9	7.88%	yeah, baby, love, version, beautiful, life, girl, know, next, blue	H12	54.93%	amor, gente, quero, vida, coração, tudo, saudade, nada, tempo, boca
Viral	V7	20.39%	yeah, like, b*tch, f*ck, sh*t, know, ni**a, want, ni**as, make	V10	22.73%	senta, toma, quer, joga, quero, desce, yeah, hoje, pode, chama
	V8	24.16%	yeah, know, love, baby, like, need, back, make, feel, tell	V11	63.25%	amor, tudo, vida, gente, quero, hoje, aqui, tempo, porque, nada
	V9	55.45%	know, love, never, like, time, away, could, life, want, come	V12	14.01%	bumbum, então, quer, soca, tudo, nego (v.), boca, chão, quatro, bota

present in Brazilian funk and, more recently, in *sertanejo universitário* lyrics.

Psycholinguistic analysis. We now apply the Linguistic Inquiry and Word Count (LIWC) to analyze the psycholinguistic properties of song lyrics to uncover patterns within hit and viral songs. LIWC uses a predefined dictionary of words and linguistic categories to group words and terms within in a given text (i.e., the lyrics) into several hierarchical attributes related to linguistic style, affective, and cognitive concepts. We again focus on songs in English and Portuguese.[15]

We then identify attributes that characterize both hit and viral songs. To that end, we search for statistical differences across them based on the average frequencies of their respective attributes. Having identified those attributes, we rank them according to their capacity to discriminate across different keywords, estimated by the Gini Coefficient [121]. In such a ranking, we do not consider exclusively linguistic attributes (e.g., linguistic dimensions such as pronouns, auxiliary verbs, and other grammar categories), as our primary interest lies in emotions and other psychological processes.

Figures 3.5 and 3.6 show heatmaps for the top-10 ranked attributes for (a) English and (b) Portuguese lyrics in the Global and Brazilian markets, respectively. The heatmap cells in a column indicate the relative deviation of each attribute for the given keyword from the other keywords. That is, each column (attribute) is normalized following the z-score – i.e., $z = (x - mean)/std$. Thus, each value gets subtracted from the average of the column, then divided by the standard deviation of the column. Therefore, red cells indicate that an attribute is more present in such a category than the average, whereas

[15]For English lyrics, we use LIWC-2015, whereas for Portuguese, we use the version of 2007.

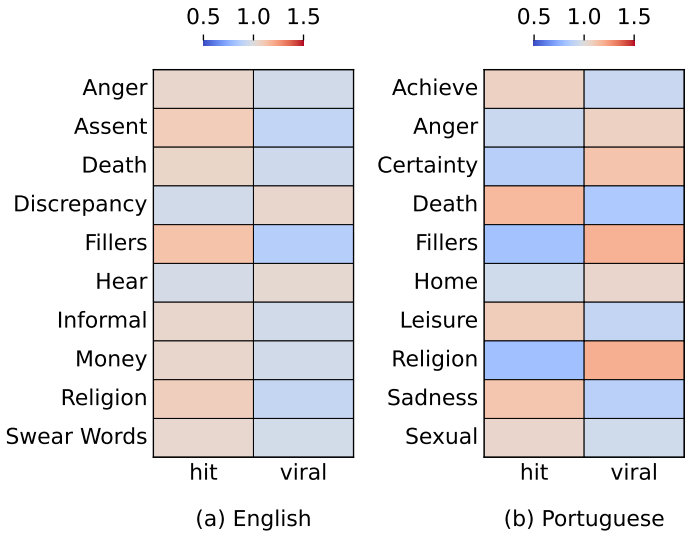


Figure 3.5: Top 10 most discriminative LIWC attributes in viral and hit songs in the Global market.

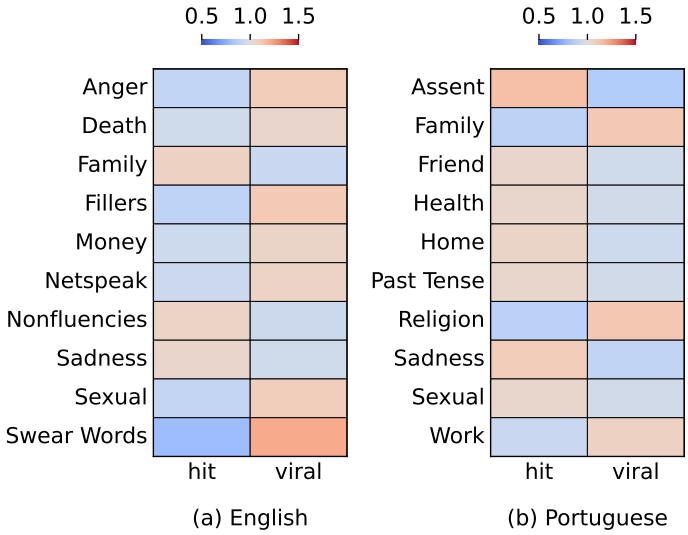
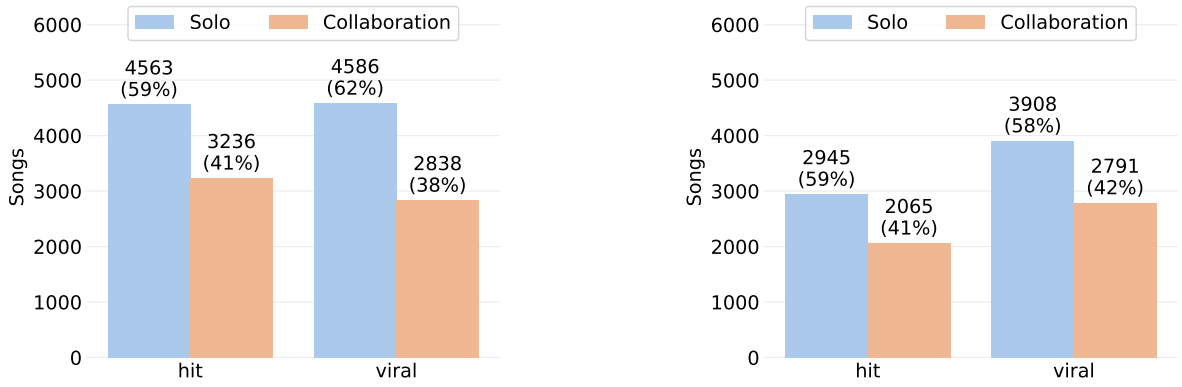


Figure 3.6: Top 10 most discriminative LIWC attributes in viral and hit songs in Brazil.

blue cells mean the opposite. For example, hit songs in English tend to have a higher frequency of swear words compared to the average in the Global market, whereas viral songs use less such language.

In the Global market (Figure 3.5), hit songs in English have a higher frequency for terms related to anger, more, money and religion, while viral songs have a higher frequency for terms related to discrepancy (e.g., should, would) and to hear. In contrast, songs in Portuguese that reach such charts have a different behavior, with viral songs having a higher frequency of terms related to religion and anger, for example.

In Brazil (Figure 3.6), the results show that terms related to family and sadness



(a) Global (b) Brazil

Figure 3.7: Proportion of solo and collaboration for hit and viral songs.

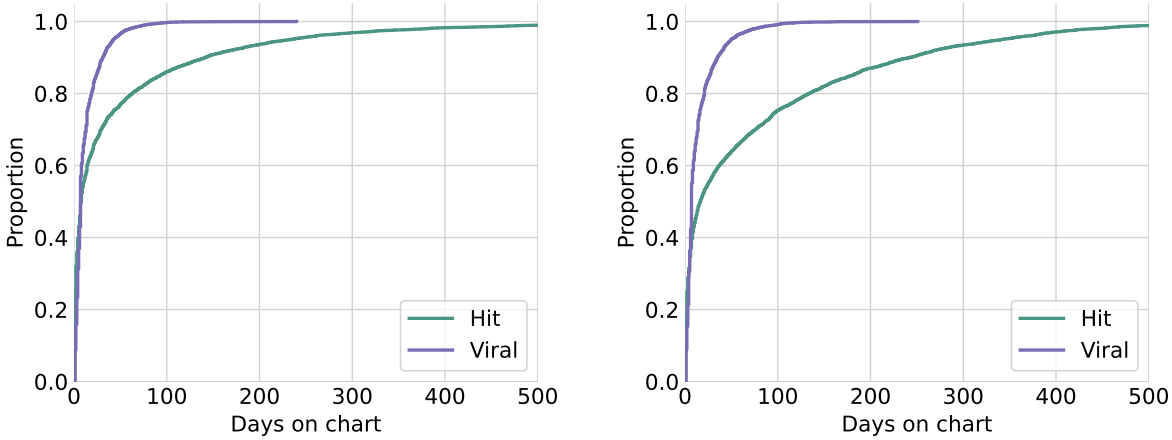
have a higher frequency in English hit songs compared to viral ones. In contrast, viral songs often include terms associated with anger, death, money, and swear words. These themes are commonly found in genres such as rap and hip-hop, which are among Brazil's top viral genres. Moreover, viral English songs have more terms related to sexual content, which again may be related to the influence of viral music genres such as rap and hip-hop.

For songs in Portuguese that reached Brazilian charts, terms associated with sexual contexts are more frequently found in hit songs than in virals, potentially reflecting the influence of popular music genres. For instance, Brazilian funk and pop songs often use more explicit lyrics in such a context when compared to other genres like *sertanejo* or *arrocha*. However, sexual themes are also present in the latter genres, albeit in a more implicit manner. Viral songs in Portuguese have more terms related to religion, work, and family, which are frequently reported in hip-hop songs.

3.2.4 Artist-related Features

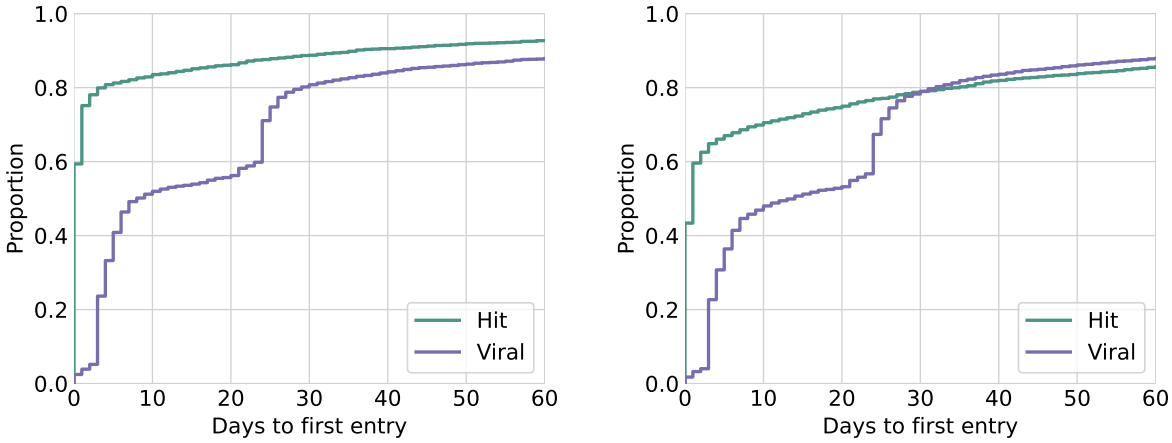
Collaboration has proven to be an important dimension behind musical success [63, 103]. We now move into the analysis of the type of songs regarding collaborations. A song is said to be a collaboration when two or more artists perform it, whether it is a *featuring* or a duet, for example. Conversely, solo songs are sung by only one artist. In addition, in this thesis, we consider groups and bands to be single artists.

In our dataset, the proportion of solos and collaborations within hit and viral songs is similar. Figure 3.7 shows that most of the hits and virals in the Global market are solo songs (59% and 62%, respectively). The same applies to Brazil, where the majority of songs are also solos, accounting for 59% for hits and 58% for virals. Collaboration occurs



(a) Global (b) Brazil

Figure 3.8: Cumulative Distribution Function of the number of days on charts.



(a) Global (b) Brazil

Figure 3.9: Cumulative Distribution Function of the number of days from a song's release to its first entry on the charts.

in 41% and 42% of hit and viral songs, respectively.

3.2.5 Temporal Features

Here, we analyze two specific temporal features to compare hit and viral songs: the number of days a song stays on the charts, and the number of days a song takes to reach the charts, also known as *song's debut on the charts*. Regarding the first one, hit songs last much more days on the charts than viral ones. On average, hit songs in the Global market stay around 49 days on the charts, whereas virals last only 13 days. The

median values are eight and seven days, respectively. In Brazil, hit songs stay 75 days on the charts, whereas viral songs stay around 14 days. The median values are 14 and seven days, respectively. In addition, Figure 3.8 shows the Cumulative Distribution Function (CDF)[16] of the number of days on charts. The results show that most viral songs last up to 100 days on charts, whereas for hit songs the proportion is around 85% in the Global market and 75% in Brazil. Such a result confirms the intuition behind what makes viral content: its ephemeral nature [44]. In fact, the popularity of a hit song is much more solid and lasting, whereas viral songs are not.

Regarding the days from a song's release to its debut on the charts, the median values for hit and viral songs are one and 14 days, respectively. That is, besides the average being similar, some viral songs take more time to reach the charts. We observe such a pattern in Figure 3.9, which presents the CDF of the feature over time. Since we are interested in the behavior in the first days after release, we truncate the plot at day 60. For instance, ten days after the release, around 80% of the hit songs in the Global market (70% in Brazil) had already reached the charts, whereas the value for viral songs is below 60%. In Brazil, the cumulative distributions meet around 30 days after the release.

The position and the duration of a specific song in the charts can be seen as a measure of its success or virality (but not the only one) [1, 3, 102]. In other words, the longer a song stays in the charts, the more successful/viral it is.

3.3 Distinguishing Viral and Hit Songs

The characterization of hit and viral songs in the previous section reveals the similarities and differences between them. Still, both can be considered distinct facets of popularity. Now, we go further in this quantitative analysis by asking: given that a song has achieved popularity status, being a viral or a hit song, can one automatically distinguish them? Specifically, given that a song (any song) has reached a chart for hits or viral songs (e.g., Spotify Top 200, Spotify Viral 50), can one automatically classify it as a viral or hit song? If so, which features are more relevant in the classification process?

This section answers such questions by evaluating three research hypotheses related to which musical features have more influence in such distinction. Distinguishing viral from hit songs may help to uncover the underlying factors behind the viral phenomenon and its specific diffusion process, in addition to understanding users' behavior when it

[16] A Cumulative Distribution Function (CDF) is a probability distribution function that describes the probability that a random variable takes on a value less than or equal to a given point. As you move along the x-axis from left to right, the CDF either stays the same or increases, reflecting the cumulative probability of observing a value less than or equal to a specific value of x.

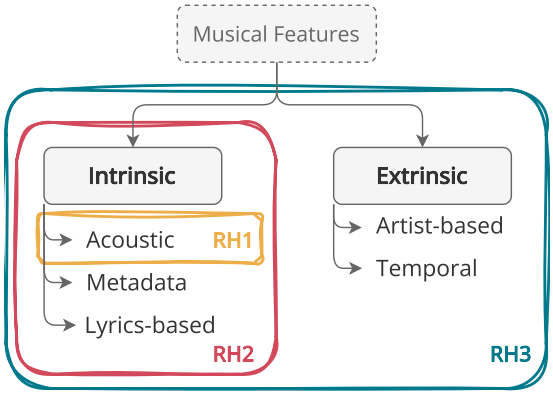


Figure 3.10: Our research hypotheses (RHs) according to the taxonomy proposed by Seufitelli et al. [98].

comes to consuming and sharing music with others. Note that we do not intend to reverse engineering any existing ranking mechanism; neither do we intend to predict which song will become a hit or viral. The goal is to unravel measurable, song-related factors that aid in the process of a song becoming viral.

In this section, we only use Spotify Global data, but the methodology can be easily applied to data from the Brazilian market (and other markets as well). After describing our research hypotheses (Section 3.3.1), we define and run the distinction between hits and virals as a classification task (Section 3.3.2). Finally, we perform a feature analysis to verify the most relevant factors in separating hits and virals (Section 3.3.3).

3.3.1 Research Hypotheses

Inspired by previous research on TikTok viral videos [52], in this thesis, we evaluate three research hypotheses (RHs) that comprise different sets of features that can be used to separate a viral song from a hit song. We follow the taxonomy proposed by Seufitelli et al. [98] in the context of *Hit Song Science* (i.e., to predict and analyze musical success) to argue that viral and hit songs may be related to acoustic features (RH1) but also to other intrinsic (RH2) and extrinsic features (RH3).[17] The features considered in this step are those defined and characterized in Section 3.2.

Figure 3.10 illustrates our research hypotheses. Note that they are incremental; that is, RH2 contains the features of RH1, and RH3 includes both RH1 and RH2 features. Such an approach allows systematic and sequential testing of hypotheses, in which each

[17]Rather than strictly adhering to traditional definitions, the taxonomy proposed by Seufitelli et al. [98] categorizes features based on their relationship with the song itself. Hence, we use such a taxonomy since it provides a useful framework for understanding the different factors contributing to musical success.

one represents a step forward by introducing new features to evaluate their individual and cumulative contributions to the analysis. In addition, starting with simpler hypotheses and progressively adding complexity helps manage the overall complexity of our analysis. We now describe each of our RHs with their rationales and features.

RH1: Acoustic Features. Despite its various facets, music is an artistic expression that uses sound as its medium. Therefore, several characteristics are related to a song's composition, structure, and style. In general, acoustic features include pitch, rhythm, dynamics, and so on [98]. Since the seminal work by Dhanaraj and Logan [21], they are the most widely used feature set in hit song prediction studies [3, 57, 103, 117]. Analyzing such features can provide valuable information about a song's emotional content and potential appeal to listeners [47].

Here, we argue that acoustic features may play a relevant role in distinguishing between viral and hit songs through their influence on a song's overall appeal, engagement, and reception. To illustrate, hit songs may be more energetic than virals, and viral songs may be more danceable due to TikTok challenges and other external factors.[18] Hence, in our analyses, we consider 13 acoustic features (present in our dataset) that summarize the structural properties of the songs: *acousticness*, *danceability*, *duration of the song (in milliseconds)*, *energy*, *instrumentalness*, *key*, *liveness*, *loudness*, *mode*, *speechiness*, *tempo*, *time signature*, and *valence*.

RH2: Intrinsic Features. In addition to the acoustic features, other features directly related to songs may help distinguish hits from viral ones. In the context of *Hit Song Science*, several works study the influence of factors such as metadata and song lyrics on their success [1, 40, 109, 123], named as *intrinsic features* by Seufitelli et al. [98]. Therefore, our second research hypothesis aims to evaluate the impact of such features when defining viral and hit songs. By definition, intrinsic features also include acoustic ones; therefore, they are all included in RH2. Hence, in addition to the RH1 features, we consider metadata and lyrics-related features: *sampling*, *remixing*, *covering*, *explicit*,[19] *language*, *number of verses*, *number of lines*, *number of words*, and *number of characters*.

RH3: Intrinsic and Extrinsic Features. Our third research hypothesis posits that the success and virality of a song may also be further directly or indirectly affected by factors external to it, called *extrinsic features* [98]. Such features have been recently added to research on hit song analysis to include factors such as social media, market data, and so on [4, 86, 116]. Here, besides all intrinsic features (RH1 and RH2), we consider extrinsic

[18] Viral Music versus Billboard Hits - What's The Difference, Really? https://blog.musiio.com/posts/viral-music-versus-billboard-hits-whats-the-difference-really

[19] A binary flag indicating whether a song contains explicit lyrics or not.

features related to artists and temporal information: *number of artists*, *song type* (solo or collaboration), *artist genres*, and *time to first entry*.

3.3.2 Classifying Hits and Virals

In this section, we aim to verify whether the features listed in our research hypotheses may distinguish hit songs from viral ones. Similar to Ling et al. [52], we evaluate the performance of several classifiers in such a task. Our goal is not to perform a prediction task but to rely on classification models to better understand what makes a song viral once it has already achieved popularity status. After presenting the problem definition (Section 3.3.2.1), we discuss data preprocessing (Section 3.3.2.2). Then, we describe our experimental setup (Section 3.3.2.3) and respective results (Section 3.3.2.4).

3.3.2.1 Problem Definition

We define the problem of distinguishing hit from viral songs as a binary classification task that, given a popular song, classifies whether it is a viral or a hit song. Formally, let $\mathcal{X}$ represent a set of songs and $\mathcal{Y} = \{0, 1\}$ the label space, where 1 corresponds to a viral song and 0 to a hit song. The goal of binary classification is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ using a training set of m instances $\{(x_i, y_i)|1 \leq i \leq m\}$. Here, each instance $x_i \in \mathcal{X}$ represents a song described by its features, and $y_i \in \mathcal{Y}$ denotes the songs' corresponding target value (i.e., hit or viral).

Such a task requires defining precisely what constitutes both hit and viral songs. Recalling Section 3.1.1 and following previous research on Hit Song Science [103, 123], we define as hits all the songs that reached the Spotify Top 200 Charts regardless of their position. In other words, the #1 and the #200 songs are treated equally as hits. Similarly, viral songs are those that reached the Spotify Viral 50 Charts. Songs present in both charts are treated as both hit and viral, i.e., we create two instances for such a song: one with the label for hit and the other for viral.

3.3.2.2 Data Preprocessing

Preprocessing data is an essential step in preparing our dataset for the classification task. It involves various techniques to enhance the quality and effectiveness of the input data. First, despite our dataset being nearly balanced, with 7,156 hits (52.1%) and 6,578 viral songs (47.9%), we perform an undersampling phase. This ensures that the model does not exhibit a bias toward the majority class, improving generalization and performance across both classes.

Then, we handle different ranges for both numeric and categorical features. For the former, we employ the MinMax Scaler, which transforms the numerical values into a range from 0 to 1 [28]. It prevents any particular feature from dominating due to its larger scale. For the latter, we perform the One-Hot Encoding transformation, which converts categorical variables into binary vectors, creating a binary feature for each unique category [28]. Such a method ensures that the model can interpret and process categorical data, preventing the misinterpretation of ordinal relationships and providing a comprehensive representation of the features.

3.3.2.3 Experimental Setup

In this section, we present the experimental setup for performing the classification task to distinguish between viral and hit songs. Specifically, we present the classification models, the set of features, and the performance metrics used.

Classifiers. To identify the most important features for distinguishing between hit and viral songs, we consider five classifier models: Logistic Regression (LR), Decision Tree (DT), Gaussian Bayesian (GaussianNB), Support Vector Machines (LinearSVC), and Random Forest (RF) [52]. As a baseline model, we use a *naive* model that predicts each class based on the distribution in our dataset. All classifiers were implemented using the *scikit-learn* Python package [83], and we maintain all classifier parameters at their default values for a fair comparison.

Feature Models. For the classification task, we compare three distinct models that test our research hypotheses, each with its corresponding features to describe the songs (see Section 3.3.1). The first model (RH1) contains only acoustic features, whereas the second

Table 3.10: Classification performance (F1-Score with 95% CI). Underlined values represent the best classifier for that model, and bold values denote the best overall result.

	RH1: Acoustic	**RH2: Intrinsic**	**RH3: Intrinsic and Extrinsic**
Baseline	0.508 ± 0.006	0.508 ± 0.006	0.508 ± 0.006
LR	0.533 ± 0.008	0.561 ± 0.014	**0.755 ± 0.009**
DT	0.491 ± 0.038	0.523 ± 0.033	0.677 ± 0.028
GaussianNB	0.257 ± 0.012	0.149 ± 0.009	0.294 ± 0.017
LinearSVC	0.530 ± 0.007	0.486 ± 0.028	**0.748 ± 0.011**
RF	0.552 ± 0.039	0.586 ± 0.040	**0.738 ± 0.025**

model (RH2) comprises songs' intrinsic features, including the acoustic ones. Finally, the third model (RH3) considers extrinsic features in addition to intrinsic characteristics.

Evaluation. To evaluate our models, we consider the F1-Score as performance metric. The dataset was randomly split into training (70%) and test (30%) sets. For each classifier and feature model, we perform a 10-fold cross-validation to obtain a robust estimate of the model's performance on unseen data. We also calculate a 95% Confidence Interval (CI) for the resulting F1-Scores based on the results from these ten folds to guarantee the reliability of the performance across different subsamples of the data.

3.3.2.4 Results

Table 3.10 summarizes the results of our classification task. In all models, the best results outperform the baseline, revealing that our features help distinguishing between hit and viral songs. Regarding the classifiers, all methods except GaussianNB present good performance for RH1. As for RH2, the top-performing classifiers are LR, DT, and RF. In contrast, the best classifiers for RH3 (i.e., considering both intrinsic and extrinsic features) are LR, LinearSVC, and RF, which have statistically similar values for F1-Score.

When comparing the three models that represent our research hypotheses, RH3 is the one with the best results (LR, LinearSVC, and RF). This means that considering both intrinsic and extrinsic features has a higher impact on differing viral from hit songs. Furthermore, comparing the results for RH1 and RH2 reveals that adding intrinsic features beyond the acoustic ones only improves the performance of RF. On the other hand, the results from RH2 to RH3 improve considerably for all methods except GaussianNB. Such results suggest that considering only acoustic (or only intrinsic) features may not be enough to distinguish viral and hit songs, i.e., other factors may influence song popularity.

The performance comparison indicates that considering intrinsic and extrinsic fea-

tures, in addition to the acoustic fingerprints, can improve the effectiveness of our models. Except for GaussianNB and DT, all classifiers achieve good results on RH3, with F1-Scores greater than 0.7 when considering the confidence interval. Moreover, LR, LinearSVC, and RF are the classifiers with the best overall performances. In the next section, we examine the factors with the most significant influence in distinguishing viral and hit songs.

Additional experiments. In addition to this main experiment, we also perform two further experiments using RH3, which considers both intrinsic and extrinsic features. The first one (E1) is a binary classification between hits and viral songs, but with the difference that it excludes songs that appear in both categories. The second experiment (E2) addresses a multiclass scenario, in which the songs can belong to one of three classes: hit, viral, or both. In both experiments, the classifiers achieve results superior to the baseline (best results with RF: F1-Score of 0.878 ± 0.020 for E1 and Weighted F1 of 0.851 ± 0.018 for E2). Such findings further confirm that our features are effective in differentiating between hits and viral songs, reinforcing the robustness of our approach.

3.3.3 Hit and Viral Indicators

In this section, we aim to unveil the features that impact the most on the performance of our classifying models. Following the methodology in Khatibi et al. [41], we perform three distinct feature importance analyses. We first evaluate the impact of specific groups of features on our models (Section 3.3.3.1). Then, we analyze the impact of removing individual sets of features (Section 3.3.3.2). Finally, we go further by analyzing the impact of individual features on our classification results (Section 3.3.3.3).

3.3.3.1 Feature Subset Analysis

Here, we investigate how subsets of features impact the performance of selected classification methods. We consider the groups of features that constitute our RH3 (intrinsic and extrinsic features), as it is the best-performing model. Specifically, we evaluate acoustic, lyrics-based, metadata, artist-based, and temporal features (see Section 3.3.1). We compare the classifiers' performance using the complete set of features (denoted by F) against the performance using a single group of features. Again, we use the F1-Score

Table 3.11: Classification results (F1-Score with 95% CI) for hit/viral songs: all features versus individual subset of features. For each classifier, the subset with the best prediction results is marked in bold.

	Baseline	**LR**	**DT**	**GaussianNB**	**LinearSVC**	**RF**
All Features (F)	0.508 ± 0.006	0.755 ± 0.009	0.677 ± 0.028	0.294 ± 0.017	0.748 ± 0.011	0.738 ± 0.025
Acoustic	0.508 ± 0.006	0.533 ± 0.008	0.491 ± 0.038	0.257 ± 0.012	0.530 ± 0.007	0.552 ± 0.039
Lyrics-based	0.508 ± 0.006	0.515 ± 0.017	0.529 ± 0.032	0.201 ± 0.036	0.529 ± 0.014	0.563 ± 0.040
Metadata	0.508 ± 0.006	0.652 ± 0.004	0.653 ± 0.004	0.114 ± 0.129	0.652 ± 0.004	0.653 ± 0.004
Artist-based	0.508 ± 0.006	**0.716 ± 0.013**	**0.677 ± 0.020**	0.287 ± 0.018	**0.704 ± 0.014**	**0.702 ± 0.018**
Temporal	0.508 ± 0.006	0.562 ± 0.011	0.562 ± 0.011	**0.562 ± 0.011**	0.562 ± 0.011	0.562 ± 0.011

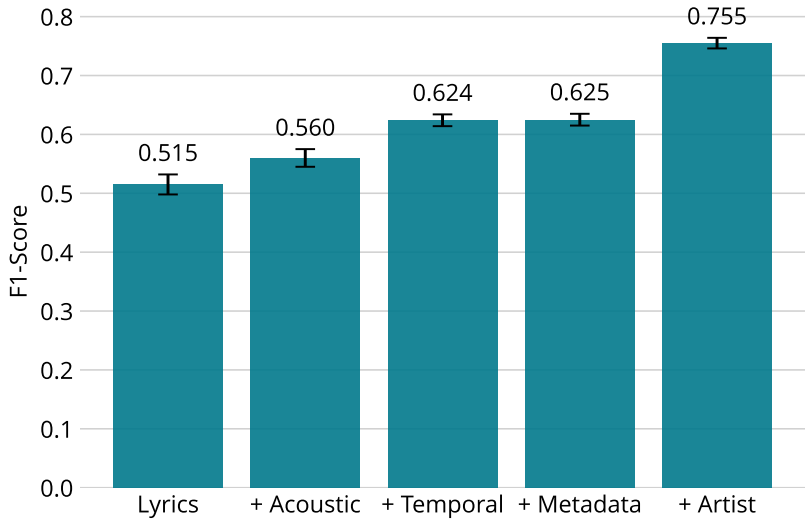


Figure 3.11: Performance of Logistic Regression (F1-Score, 95% CI) on viral/hit songs as we introduce new feature subsets.

with a 95% confidence interval (CI) to evaluate the performance.

Table 3.11 presents the performance results for each technique using all features and the individual groups of features. Considering only artist-based features produces the best results for all classifiers, except for GaussianNB, which performs best for temporal features (Naive Bayes is better suited for categorical features). Such a result suggests that features including collaboration, artist genre, and the time from release to chart debut may be among the most relevant in our classification method. However, the complete set of features F is the one with the best overall results, reinforcing the importance of simultaneously using both intrinsic and extrinsic song data to improve the classification between hit and viral songs.

We now go further in our analysis by evaluating how introducing new feature groups improves the classifier's performance, from the worst individual group (lyrics-related) to the best one (artist-related features). Among the methods with the best results, we choose LR since it is simpler and computationally less expensive. The results are illustrated by Figure 3.11. Considering lyrics alone does not produce good results compared to the baseline method. However, adding new sets of features gradually improves the results.

Table 3.12: Classification results (F1-Score with 95% CI) for hit/viral songs: all features versus all but one feature subset. Values in bold show the most contributing feature subset for each classifier.

	Baseline	**LR**	**DT**	**GaussianNB**	**LinearSVC**	**RF**
All Features (F)	0.508 ± 0.006	0.755 ± 0.009	0.677 ± 0.028	0.294 ± 0.017	0.748 ± 0.011	0.738 ± 0.025
F - Acoustic	0.508 ± 0.006	0.756 ± 0.009	0.696 ± 0.023	0.294 ± 0.017	0.746 ± 0.010	0.746 ± 0.022
F - Lyrics-based	0.508 ± 0.006	0.754 ± 0.010	0.679 ± 0.023	0.287 ± 0.018	0.748 ± 0.010	0.741 ± 0.025
F - Metadata	0.508 ± 0.006	0.756 ± 0.009	0.679 ± 0.025	0.294 ± 0.017	0.749 ± 0.011	0.738 ± 0.026
F - Artist-based	0.508 ± 0.006	**0.625 ± 0.010**	**0.608 ± 0.033**	**0.151 ± 0.009**	**0.627 ± 0.009**	**0.679 ± 0.031**
F - Temporal	0.508 ± 0.006	0.715 ± 0.012	**0.615 ± 0.027**	0.294 ± 0.017	0.706 ± 0.012	**0.679 ± 0.029**

For example, including temporal features significantly enhances the classification, but further improvements happen when adding the remaining sets of features.

3.3.3.2 Feature Subset Ablation

We continue to investigate the predictive power of our features by performing a feature ablation analysis. From the complete set of features F of RH3, we evaluate the performance of all classifiers if we remove a group of features f (i.e., $F - f$). For example, "F - Acoustic" indicates the model in which the acoustic features are removed from the input set. Table 3.12 presents the results of such analysis in terms of F1-Score. For each classifier, we emphasize (in bold) the most contributing feature group, i.e., the group whose removal produces the lowest F1-Score.

For all considered classifiers, the weakest features are acoustic, lyrics-based, and metadata (i.e., the intrinsic ones). Removing such features from the input set does not impact the classification outcome. In contrast, removing extrinsic characteristics such as artist-based and temporal features significantly affects the classifiers' performance. Specifically, artist-based features are the group that contributes the most to each learning method (alongside temporal features in DT and RF), as their removal reduces the classification performance by approximately 17%.

Such results are consistent with our previous analysis, and they suggest that extrinsic factors may be among the most significant ones for defining a song's virality. In other words, features such as the presence of artist collaboration and artists' musical genres may be fundamental to distinguishing whether a song is a hit or a viral given that it has reached popularity. Next, we delve into the individual features within such groups to uncover the specific factors associated with music virality.

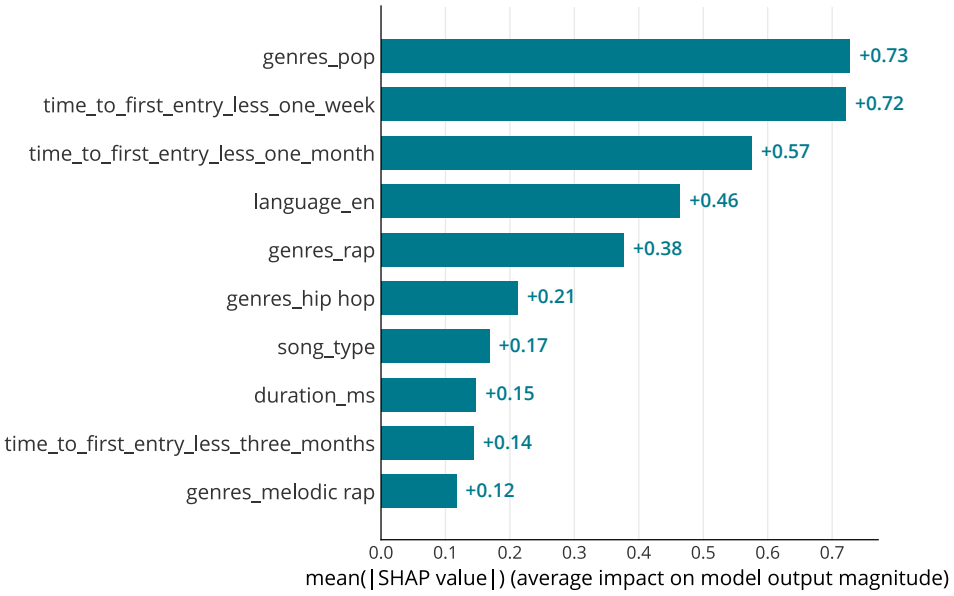


Figure 3.12: Top 10 features with the highest absolute mean SHAP values.

3.3.3.3 Individual Feature Importance

Understanding why and how a model makes a specific prediction can be as crucial as the outcome, shedding light on the "black box" within the learning algorithms. Here, we use SHAP (SHapley Additive exPlanations) [53] in our classification method to allow its interpretability. In contrast to the analyses from the previous section, we now analyze the features individually rather than in groups. In short, SHAP is a game-theoretic approach for explaining the output of a learning model, assigning an importance value for a particular prediction for each feature. From a global perspective, such values can be aggregated to show how much each predictor contributes to the target variable, either positively or negatively.

We first analyze the features with the highest absolute mean SHAP values. Figure 3.12 presents the top 10 features with the highest impact on the classification of hit and viral songs. Artist-related (i.e., artist genre, song type) and temporal features (i.e., time to first entry) are among the most relevant for the prediction, corresponding to 8 of the top 10, confirming the findings of the previous section. In particular, the features that inform whether a song reached the charts in less than a week or a month impact the classification output most. In addition, the descriptive genre features obtained from the discretization preprocessing are also relevant, with genres such as pop, rap, and hip-hop contributing significantly to differing hits from viral songs. Other relevant features include song duration and the presence of the English language in the lyrics.

Next, we evaluate the positive and negative relationships of the features with the target variable. Figure 3.13 goes further in the summary plot, using SHAP values to show

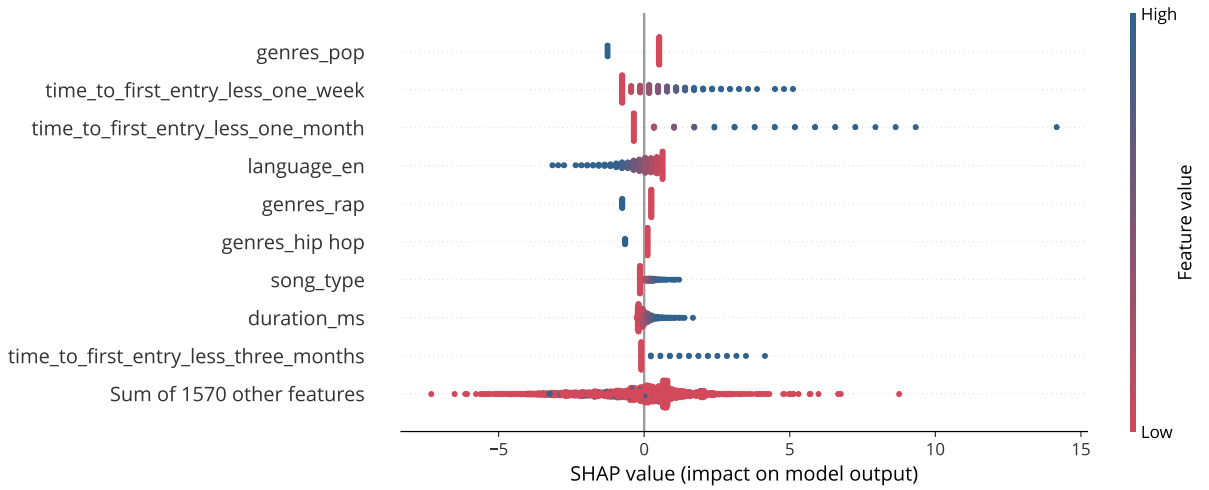


Figure 3.13: SHAP values. Features are sorted by the absolute average SHAP value calculated over all samples. The color represents the feature values (teal high, red low).

the distribution of the impact of each feature in the output. Features are sorted based on their mean absolute SHAP values, and each point on the x-axis indicates whether the effect of that value corresponds to a higher or lower prediction (recalling that $hit = 0$ and $viral = 1$). The color scale provides information on whether the feature value is high (teal) or low (red) for a particular instance. For example, songs with higher duration positively impact the output, i.e., they are more related to viral songs. Such a relation may be justified by the fact that hit songs have been getting shorter in the last few years.[20]

Moreover, high values for features that inform whether a song reached the charts in less than a week or a month (i.e., two of the Top 3 most relevant for the classifier) also positively impact the outcome. Thus, they are more related to viral songs. In other words, positive values for such features increase the probability of the classifier output being "viral". On the other hand, songs in English negatively impact the classification, which is related to hit songs. When it comes to genres, the presence of artists from genres such as pop, rap, and hip-hop negatively impacts the output, i.e., being more related to hit songs. Indeed, such genres are among the most popular globally. Overall, all the aforementioned features significantly impact the prediction. Therefore, they can be considered the main factors to distinguish viral from hit songs in our model.

[20]Vice (Oct. 2023): https://bit.ly/3SakvMc

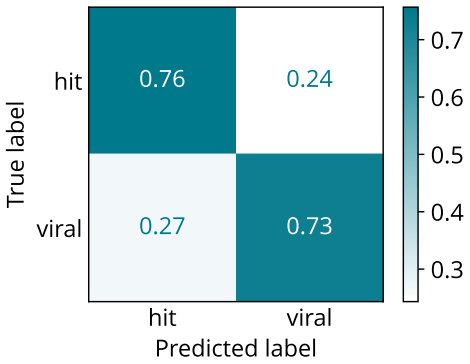


Figure 3.14: Confusion matrix of the classification (LR, RH3).

3.4 Discussion on the Results

Overall, our results reveal that despite sharing similarities in their relation with popularity, *hit and viral songs are not equivalent entities*. Beyond the acoustic features of music, this study emphasizes the importance of considering other intrinsic and extrinsic features for a more accurate definition of hits and virals. Intrinsic factors, such as metadata and lyrics-related characteristics, and extrinsic factors, such as artist-based and temporal features, play crucial roles in differentiating hits from viral songs. Such a comprehensive approach allows a deeper understanding of the dynamics behind musical popularity in online platforms, enriching our ability to classify these two distinct categories.

Regarding feature subsets, extrinsic features related to artists (i.e., their music genres and the presence of collaborations) stand out as the most important for telling apart hits from viral. However, when analyzing features individually, temporal features such as the time from release to chart entry emerge as the most relevant for our classification. Furthermore, artists' genres are also highly relevant, with the presence of genres like pop, rap, and hip-hop being crucial in defining hit songs.

Although this is not a Hit Song Science (HSS) study (i.e., our goal is not to predict the popularity of a song, but to characterize it), our findings support results from previous research on HSS. For instance, artist collaboration (song type) emerges as one of the most relevant features in our classification task, confirming the result from Silva et al. [103], which identifies it as a relevant factor for success, and from other studies which use collaboration in their prediction [101]. Besides, the presence of artists from popular genres underscores the importance of considering musical genres when analyzing success and artists' careers [63, 70].

Regarding the cases in which our method cannot differ hits from virals, Figure 3.14 presents the confusion matrix for the LR classifier using intrinsic and extrinsic features

(RH3). Although it correctly classifies most instances, there is a considerable set of songs for which it fails (i.e., 24% of hits and 27% of virals). We do not find significant differences in the value distribution for any feature, suggesting that there are other factors necessary to distinguish hit and viral songs that are not considered here.

Therefore, other features can be incorporated into future analyses to enhance the comprehension of success and virality in online platforms. For instance, including social media data, such as user playlists, may be relevant for studying the elements that drive a song to go viral or achieve hit status. Such playlists directly reflect individual preferences and choices, capturing real-time dynamics of music consumption. Moreover, virality is often associated with a song's ability to spread across online communities organically, and considering the discussion around songs in social networks may also serve as a direct indicator of these sharing patterns. In summary, considering social media data may expand the understanding beyond traditional metrics and provide a more comprehensive view of the social interactions driving musical popularity.

3.5 Overall Considerations

In this chapter, we compared hit and viral songs by analyzing data obtained from Spotify and enhanced with Genius metadata. While hits and virals share some characteristics, we identify specific differences in their intrinsic and extrinsic features. For example, the temporal features reveal important insights into the behavior of the songs in the charts, as they inform how music is consumed. For the sake of illustration, there is a clear difference in the number of days that viral and hit songs stay in the charts, and also in the period that such songs take to reach them. This suggests the main distinction between hit and viral songs lies mainly in the **diffusion process** itself (i.e., the songs' consumption and their viral spreading), and not necessarily in the intrinsic characteristics of the song. Although this idea may seem intuitive because of the definition of hit and viral, it is essential to support it with concrete data.

In addition, we perform a classification task to evaluate three research hypotheses comprising different feature sets. Our results reveal that using only acoustic or other intrinsic features is not enough to differentiate such songs, and extrinsic features are also relevant in such a task. The feature importance analysis reveals that artist-related and temporal features are the most important within the comparison; the time from the release to its first chart entry, the song duration, and artist collaboration are the ones that most positively impact the classification, increasing the probability of a song being viral.

Our findings reinforce the definition of hit and viral songs as two distinct facets

of music popularity. Essentially, they represent two interconnected yet different aspects. For example, **virality may be a stepping stone to achieving hit status**, but not all viral songs will become hits. Indeed, all this context highlights the complexity of the music industry, and unveiling factors that help to identify viral songs may serve as a basis for understanding the dynamics of music consumption. As the music industry adapts to technological advances and shifting audience preferences, this chapter provides valuable insights into the viral phenomenon in streaming platforms and its relation to success.

Limitations. The main limitation of the study presented in this chapter lies in the integration of Spotify data with Genius. Such integration may have affected the extent of our analyses, as some songs could not have their lyrics analyzed due to a lack of information. On the distinguishing phase, the random split of the dataset into training and test sets did not consider the temporal aspect of popularity, which could potentially impact the stability of the classification results when dealing with time-dependent phenomena. Furthermore, at the time this work was conducted and the data was obtained (up to March 2022), Spotify had not yet widely implemented video features, although we acknowledge that they may also be relevant for such analysis. We also do not consider other features, such as data from social media to represent the discussion about the songs.

Chapter 4

On the Causal Relationship Between Music Virality and Success

As mentioned in the previous chapter, the symbiotic relationship between musical virality and success supports the hypothesis that virality on social platforms can be a stepping stone for a song's commercial success. Platforms such as TikTok have demonstrated significant power in amplifying music virality, where dance challenges and memes can transform a relatively unknown or niche-consumed track into a global phenomenon. For instance, in 2022, almost all No. 1 hits in the US and the UK were driven by viral trends on TikTok.[1] The opposite may also happen, with successful hits generating viral content, preserving its relevance in popular culture over time.

The complex dynamic between virality and success highlights the significant role of social platforms in shaping contemporary musical trends, as well as explaining musical consumption worldwide. Within such a dynamic, relevant context, this chapter aims to **analyze the temporal relation between music virality and success**. By studying the trajectories of viral and hit songs in streaming platforms (more specifically, Spotify) over time, we aim to understand the dynamics underlying the interplay between these two facets of music popularity. Specifically, we aim to answer the following research questions: *Is there a synchrony between the virality of a song and its success? Can the virality of a song be used as an indicator of its future success or vice-versa? Is there a causal relationship between music virality and success?*

We follow the methodology illustrated in Figure 4.1 to assess the temporal relationship between songs' virality and success. For each song present in a global music chart dataset (Section 4.1), we build two distinct time series that represent its virality and success (Section 4.2). We then perform an initial correlation analysis to verify the synchrony between success and virality (Section 4.3). Next, we apply two techniques to analyze distinct aspects of the relationship between those two time series: Granger Causality to check whether the viral time series can be used to forecast the success series (Section 4.4); and causal discovery to infer the proper causal relationship between them

[1]Music Business Worldwide: https://www.musicbusinessworldwide.com/13-out-of-the-14-no-1-songs-in-the-us-in-2022-were-driven-by-viral-trends-on-tiktok/

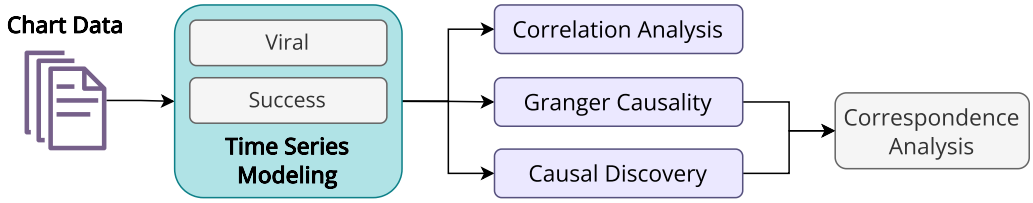


Figure 4.1: Methodology for analyzing the temporal relationship between musical virality and success.

Table 4.1: Dataset main statistics.

	Songs	Artists	Period
Viral	7,424	5,257	January/2017 to March/2022
Hit	7,799	2,159	January/2017 to March/2022
Both	2,298	1,352	January/2017 to March/2022

(Section 4.5). Finally, we perform a correspondence analysis between the results of both approaches (Section 4.6) and present our overall considerations (Section 4.7).

4.1 Data Acquisition

Here, to analyze temporal relationship between musical virality and success, we also consider data from the Music Genre Dataset (MGD+) [97], which has enriched data from Spotify music charts. Specifically, we consider the Top 200 and Viral 50 daily charts as the source of hit and viral songs on the platform, respectively.[2] The Top 200 is simply the list of the most listened-to songs on the platform sorted by the number of streams, whereas the Viral 50 considers several factors to rank the songs, including the increase in the number of plays and the amount and frequency in which people share a song.[3]

Spotify produces charts for every market where it is present, in addition to the so-called Global charts, which aggregate global music consumption trends. Since we are more interested in analyzing the temporal relationship between music virality and success rather than focusing on a specific country, Global charts are more suitable for this thesis. Hence, we consider daily charts available within MGD+, which contain data from January 2017 to March 2022. Using more recent data is not feasible due to core changes within the Spotify Charts platform, which closed public access to their charts on March 2022.

Table 4.1 presents the general characterization of the dataset. Overall, we consider

[2]Note that we do not seek to evaluate how Spotify creates or verifies its rankings; we chose this platform because, at the time of the research, it was the most popular and provided the most extensive publicly available set of global and local charts.

[3]Spotify: http://support.spotify.com/us/artists/article/understanding-spotify-charts

7,424 viral and 7,799 hit songs. However, as our goal is to analyze the causal relationship between music virality and success, we only use songs that are present on both charts, i.e., songs that are (at some point) both viral and hit. Therefore, our final set of songs contains 2,298 songs, representing 30.9% and 29.5% of viral and hit songs, respectively.

4.2 Time Series Modeling

Time series analysis has been extensively used in several domains, from metal production [85] to the oil industry [90]. Here, we build two distinct time series for each song in our dataset to represent the temporal evolution of their virality and success. In other words, such time series convey the song's performance in the Viral 50 and Top 200 charts. Specifically, each time series spans from the song's release date (or the first available chart, if the song was released before 2017) to the most recent chart available. Hence, each data point in the time series denotes the song's daily virality or success, as inferred from its chart performance.

To quantify the performance of songs, we employ a rank score metric derived only from their chart positions. The rank score $RS(i)$ of a song ranked at position i is given by $RS(i) = max\_rank - i + 1$, in which max_rank represents the highest achievable rank (i.e., 50 for the viral chart and 200 for the success charts), and i denotes the song's position on the chart. For instance, if a song is ranked #10 on the Viral 50 charts, its rank score is 41 since $RS(10) = 50 - 10 + 1 = 41$. We deduce from the formula that the rank score is always in the $[1, max\_rank]$ range. However, if a song is not in the charts on a particular day, we set its rank score to zero.

An example of the construction of time series is illustrated by Figure 4.2, which presents the viral and successful time series of the song "positions" by Ariana Grande. The track was released on October 23, 2020, and debuted at the top of Spotify's Top 200 (i.e., success rank score of 200), remaining in that position for 11 days. From its release to the most recent chart date available, the song has remained consistently present on the charts, although it has gradually lost positions over time. As for its viral performance, the song debuted on the Viral 50 at #16 (i.e., viral rank score of 35) but remained on the chart for only 14 days.

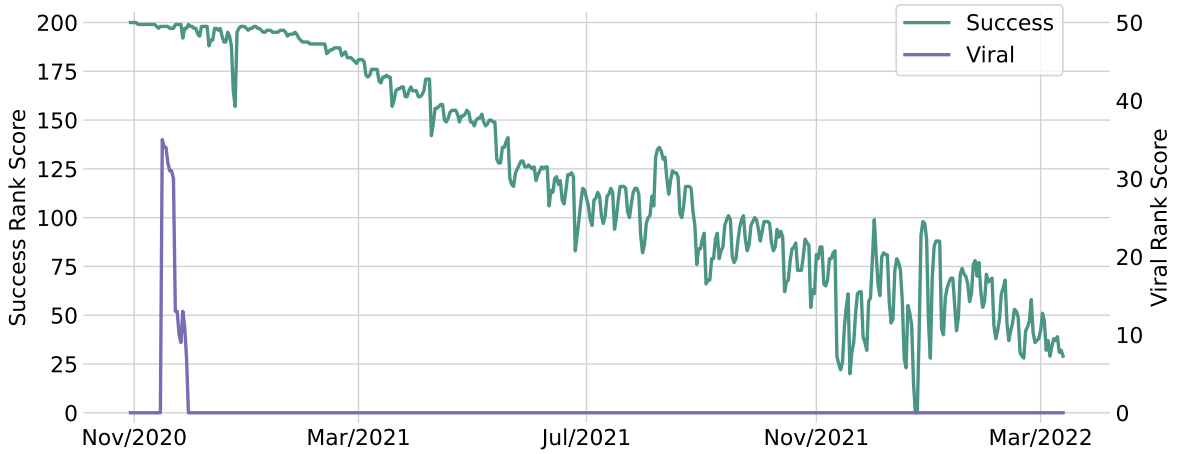


Figure 4.2: Time series for the song "positions" by Ariana Grande. Note the different y-axis scales for success and viral scores.

4.3 Correlation Analysis

In this section, we analyze the synchrony between the trajectories of song virality and success within the dataset. By analyzing the time series derived from both hit and viral charts, we aim to verify whether there is any correlation between them in order to better understand the relationship between virality and success in the Global music market. Although simple, correlation analysis is a powerful tool to unveil significant patterns in the temporal evolution of hit and viral songs, offering valuable insights into their dynamic relationships over time.

Here, we calculate two distinct correlation coefficients: Pearson (r) and Spearman (ρ). Whereas Pearson correlation assesses the linear relationship between two variables, Spearman correlation evaluates the monotonic (i.e., rank-based) association between them. Both are numbers ranging from -1 (negatively correlated) to 0 (not correlated) to 1 (perfectly correlated). Thus, we calculate the two coefficients for each song in the dataset by comparing their success and virality time series.

Figure 4.3 illustrates the distribution of Pearson and Spearman correlation values for the time series. In both cases, a considerable proportion of songs exhibit coefficients close to zero (approximately 18%), suggesting that there is no linear or monotonic correlation between success and virality for these songs. Such an observation highlights the complexity and variability in the relationship between these variables across the dataset.

However, most of the songs in the dataset present positive correlation coefficients (around 80.1% of the songs), revealing a tendency (weak or strong) of synchrony between the virality and their success. Specifically, the median correlation value is 0.324 for Pearson and 0.343 for Spearman, representing a weak to moderate correlation (according to

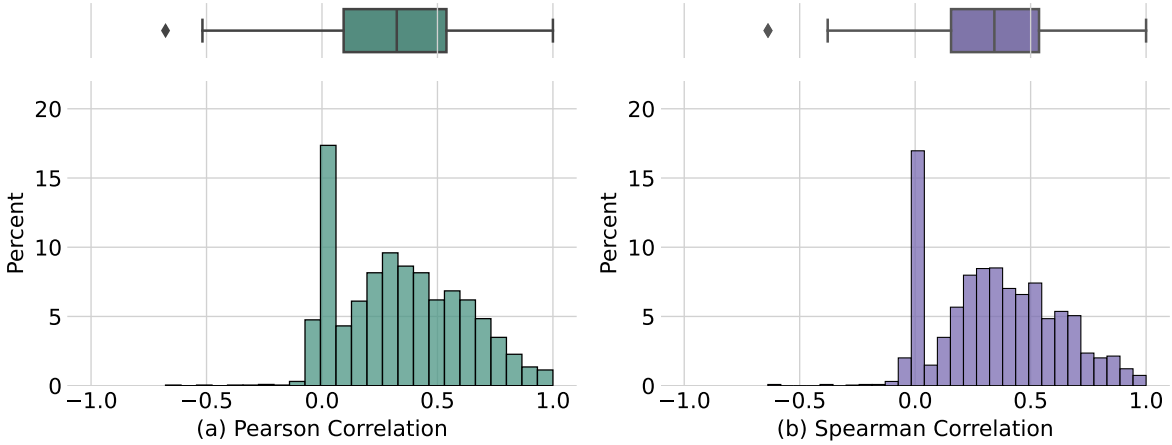


Figure 4.3: Distribution of the (a) Pearson and (b) Spearman correlation coefficients for the songs in our dataset.

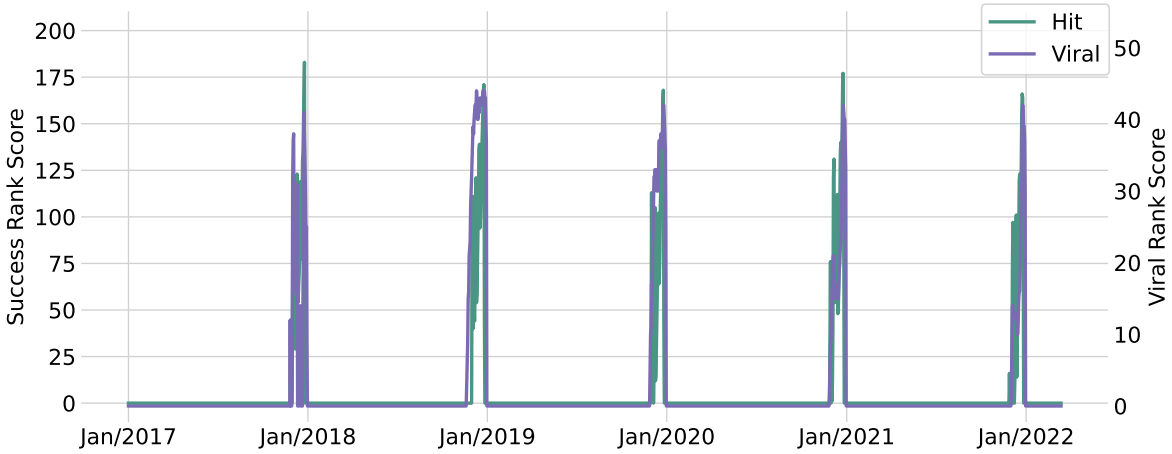


Figure 4.4: Time series for the song "Fairytale of New York" by The Pogues feat. Kirsty MacColl. There are distinct y-axis scales for success (left) and viral (right) charts.

Cohen [19]). Note that while Pearson's coefficient assesses the synchrony between the actual values of the virality/success metric (in our case, the rank score), Spearman's coefficient only considers the order of these values, making it robust against outliers and non-linear relationships. Since both correlation coefficients present similar results, we proceed to analyze only the Pearson Coefficient (r) from now on.

Songs with strong positive correlations. Although the median correlation value is weak to moderate, some songs show a strong correlation between their viral and success trajectories. Specifically, three songs present nearly a total positive correlation ($r \approx 1$), indicating perfect synchrony between such trajectories. A detailed manual analysis reveals that all such songs have a single entry (i.e., they are present on only one day) on both charts. Consequently, the shapes of the two curves mirror each other closely, explaining the extremely high correlation values.

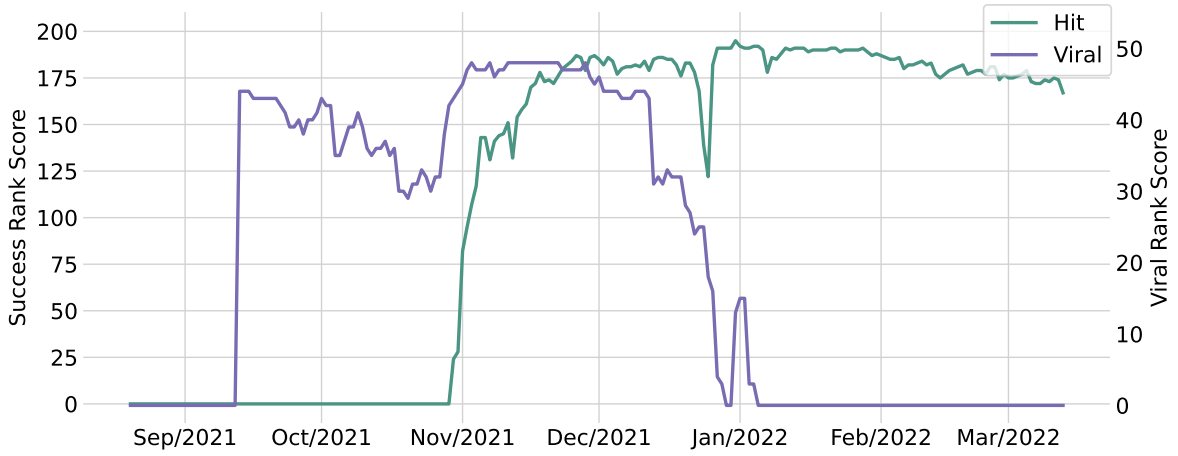


Figure 4.5: Time series for the song "Do It To It" by ACRAZE and Cherish. There are distinct y-axis scales for success (left) and viral (right) charts.

However, there are still other songs with a high correlation between their viral and success time series. As an example, we analyze the song "Fairytale of New York" by The Pogues featuring Kirsty MacColl (Figure 4.4), which has a correlation coefficient of $r = 0.757$. The song was originally released in 1987 as a single from the London-based band's album *If I Should Fall from Grace with God*. Over time, it has become one of the most popular Christmas songs in the United Kingdom, being the most listened to in the country in the 21st century.[4] This behavior is reflected in the song's performance on Spotify's Global charts, given the UK's large representation in its user base. In all years considered in our dataset, the song reaches both hit and viral charts almost simultaneously around Christmas. Thus, the similarity in the curves of both charts explains the high correlation existence for this song.

Songs with negative correlations. Unlike most of the songs in the dataset, a subset of ten songs shows a representative negative correlation ($r < -0.1$) between their viral and success trajectories. In other words, as a song's position on the viral chart increases, its position on the success chart decreases, and vice versa. An illustrative example is the song "Do It To It", released by the DJ ACRAZE featuring the American girl group Cherish. The song was released in August 2021, but it is a tech house rework by the DJ of the original version released by Cherish in 2006.

The song gained significant attention on social media platforms, resulting in a large wave of sharing and a great discussion about the song. As depicted in Figure 4.5, this viral process helped the song to enter Spotify's viral chart first, maintaining high positions in this ranking for a few months. After a few weeks, the song also started appearing on the success chart and grew there from then on. In fact, this song earned ACRAZE his

[4]The Telegraph: https://www.telegraph.co.uk/news/2023/11/30/fairytale-of-new-york-finally-reach-no-one-shane-macgowan/

first entry on Billboard Hot 100.[5] The negative correlation observed between these two trajectories shows the phenomenon in which the song's position on the viral charts begins to decline as it ascends within the success chart.

Overall, based on the correlation analysis, it is not possible to affirm that there is *always* synchrony (i.e., a high positive correlation) between virality and success. Although this is true for a set of songs, other songs present a low correlation between such trajectories. In addition, correlation is only a snapshot of overall synchrony. Therefore, it does not inform the directionality between the two curves, i.e., which signal leads and which follows. To do so, we now perform a Granger causality analysis to better investigate the temporal relationship between music virality and success.

4.4 Granger Causality

We now perform a Granger Causality (GC) analysis to verify whether virality can be used to indicate future success or vice versa. GC is a statistical test proposed in the context of econometrics for verifying the usage of one variable in forecasting another in time series data with a particular lag [30]. In short, it assesses whether the past values of a time series provide useful information for predicting future values of another time series beyond what can already be predicted from past values of the second time series alone.

To find GC between two time series X and Y, we perform a statistical test to assess whether including lagged values of X as predictors improves the forecasting of Y compared to a model that only includes lagged values of Y as predictors. If the inclusion of lagged values of X significantly improves the forecasting of Y, then we say that X *Granger-causes* Y. Specifically, GC tests the null hypothesis H_0 that X does not *Granger-causes* Y. If H_0 is rejected with a p-value below the predefined threshold, we accept the alternative hypothesis H_1 that X *Granger-causes* Y. In other words, the past values of X contain valuable information for predicting the future behavior of Y beyond what can be explained by the past values of Y alone.

Besides having "causality" in its name, GC does not imply causality in the traditional sense and should not be used to make direct causal inferences. Instead, GC helps analyze potential relationships between variables and is useful in predicting trends. Therefore, in this analysis, we are not verifying a causal relationship between music virality and success, which would require other extensive analyses. Rather, we are studying whether the viral trajectory of a song can serve as an indicator of future success (and vice

[5]Billboard: https://www.billboard.com/music/music-news/acraze-do-it-to-it-cherish-interview-1235048695/

versa). Next, we first present the possible outcome scenarios from GC analysis (Section 4.4.1). Then, we present the stationarity check and lag definitions (Section 4.4.2). Finally, we present and discuss the results of GC (Section 4.4.3).

4.4.1 Outcome Scenarios

GC is a unidirectional relationship, i.e., the fact that X *Granger-causes* Y does not necessarily imply that Y *Granger-causes* X. Therefore, to assess the impact of virality on musical success and vice versa, we explore four distinct scenarios from the results of GC in the song set, which are detailed next.

GS1. ***The song's virality can be used to forecast its success (Viral → Success).*** In this scenario, only the viral time series *Granger-causes* the success time series. In other words, changes or fluctuations in the viral time series precede and have a predictive influence on changes or fluctuations in the success time series.

GS2. ***The song's success can be used to forecast its virality (Success → Viral).*** This is the converse scenario of S1, where only the success time series *Granger-causes* the viral time series. This implies a relationship in which success plays a significant role in influencing the future virality of a song.

GS3. ***Both virality and success can be used to forecast each other (Viral ↔ Success).*** This scenario represents a bidirectional relationship, in which the two time series *Granger-cause* each other. This suggests a mutual influence between virality and success, with each factor impacting the other in a dynamic and interconnected manner.

GS4. ***There is no statistical relationship between a song's virality and its success.*** In this scenario, there is no GC relationship between both time series, meaning that changes or fluctuations in the viral time series cannot be used to predict or influence changes in the success time series, and vice versa.

4.4.2 Stationarity Check and Lag Definition

One prerequisite for employing the Granger Causality (GC) test is ensuring the stationarity of the time series, i.e., they should have a constant mean, variance, and no seasonal component. To verify this, we use two statistical tests: the Augmented Dickey-Fuller (ADF) and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) ones. Essentially, ADF evaluates the null hypothesis H_0 that the series is non-stationary due to a unit root [27], whereas KPSS tests the null hypothesis H_0 that the time series is stationary around a deterministic trend [46]. In both tests, we set the significance threshold at 0.05. If the resulting series remains non-stationary, it is not possible to run GC.

The next step is to calculate the causality lag for each song. To do so, we use a VAR (Vector Autoregression) model, which is a statistical model used to analyze the relationship between multiple time series variables. In a VAR model, each variable is modeled as a linear function of its own lagged values and the lagged values of other variables in the system. Here, lag refers to the number of past time periods included in the model's equation for each variable. To define the number of lags p, we initially estimate an unrestricted VAR(p) model and select the model that minimizes the information criteria of Akaike (AIC), Bayes (BIC), and Hannan-Quinn (HQIC) [94].

4.4.3 Results and Discussion

After checking stationarity and determining the causality lag for the virality and success time series of each song, we apply Granger Causality (GC) over such series by using the Python library *statsmodels* [95]. As a prerequisite for GC, both time series must be stationary. In addition, it is not possible to run GC in cases where the viral and success time series are exactly the same (i.e., mostly songs having a single entry on the charts). Therefore, our song set is reduced to 805, i.e., approximately 35% of the total.

Table 4.2 provides an overview of the GC results grouped by outcome scenarios. Overall, there are songs belonging to all four possible scenarios, indicating that there is no unique pattern governing the relationship between music virality and success in the Global charts. To better grasp such results, Table 4.3 presents specific examples of songs for each scenario, along with their corresponding causality lag values and the p-values obtained from the unidirectional GC tests for $Viral \rightarrow Success$ and $Success \rightarrow Viral$.

Most of the songs in the dataset present a bidirectional relationship between virality and success (i.e., $GS3$, in which both virality and success can be used to forecast each

Table 4.2: Summary of Granger Causality results (outcome scenario, number of songs, and respective percentage).

Outcome Scenario		**n**	**%**
GS1.	Viral → Success	8	0.99%
GS2.	Success → Viral	85	10.56%
GS3.	Viral ↔ Success	498	61.86%
GS4.	No Granger Causality relationship	214	26.58%
TOTAL		**805**	**100%**

Table 4.3: Results of the Granger Causality (GC) test (outcome scenario; song name and artists; the lag orders employed in the GC test; p-value of the GC test, where ** and * mean statistical significance at the 0% and 1% levels, respectively). Songs within each scenario are sorted in alphabetical order.

	Song	**Artist**	**Lag**	**Viral → Success**	**Success → Viral**
	Back To You	Selena Gomez	30	< 2.2e-16**	0.283
	Better	Khalid	2	< 2.2e-16**	0.777
GS1	Mi Gente	J Balvin, Willy William	59	0.004*	0.278
	Olha a Explosão	MC Kevinho	28	0.003*	0.409
	Supastars	Migos	3	< 2.2e-16**	0.967
	All I Want for Christmas Is You	Mariah Carey	12	0.109	0.001*
	Look What You Made Me Do	Taylor Swift	30	0.467	< 2.2e-16*
GS2	Psycho	Post Malone, Ty Dolla $ign	30	0.982	< 2.2e-16**
	Rebelde	RBD, Anahí, Dulce María, Maite Perroni, Christian Chávez, Christopher von Uckermann, Alfonso Herrera	10	0.946	< 2.2e-16**
	There's Nothing Holdin' Me Back	Shawn Mendes	17	0.757	0.002*
	Believer	Imagine Dragons	47	< 2.2e-16**	< 2.2e-16**
	Despacito	Luis Fonsi, Daddy Yankee	59	< 2.2e-16**	< 2.2e-16**
GS3	Estamos Bien	Bad Bunny	18	< 2.2e-16**	< 2.2e-16**
	Kill This Love	BLACKPINK	44	< 2.2e-16**	< 2.2e-16**
	Savage Remix	Megan Thee Stallion, Beyoncé	34	< 2.2e-16**	< 2.2e-16**
	Despacito - Remix	Luis Fonsi, Daddy Yankee, Justin Bieber	3	0.971	0.949
GS4	God's Plan	Drake	57	0.999	0.940
	INDUSTRY BABY	Lil Nas X, Jack Harlow	9	0.996	0.994
	My Universe	Coldplay, BTS	16	0.940	0.729
	no tears left to cry	Ariana Grande	23	0.793	0.950

other). This result suggests that, for such songs, each factor impacts the other in a dynamic and interconnected way. Examples include the songs "Savage Remix" by Megan Thee Stallion and Beyoncé, and "Believer" by Imagine Dragons. Such songs present parallels between their viral and success trajectories, as their shapes are similar and their peaks occur almost simultaneously.

Furthermore, fewer songs present a unidirectional relationship between virality and success (i.e., scenarios $GS1$ and $GS2$). For example, $GS1$ reveals that viral time series can be used to indicate future success for just eight songs. An example is the song "Mi Gente" by J Balvin and Willy William, considered the first Latin hit after "Despacito". Released on June 30, 2017, the song reached the top of the viral charts in less than a

week. Following such a trend, the song also gained positions on the success chart in the subsequent days, being the most streamed song worldwide on July 31st, exactly one month after its release.

However, there is another portion of songs (26.58%) in which GC reveals there is no relationship between their viral behavior and their success (i.e., scenario $GS4$). In this case, changes or fluctuations in the viral time series do not directly impact changes in the success time series, and vice versa. Therefore, based on such results, we cannot affirm that the future success of a song will always be influenced by its viral process. I.e., although there is a relationship between virality and success in many cases, these results cannot be generalized as a fact.

4.5 Causal Discovery

After exploring the relationship between virality and success using Granger Causality (GC), we advance the study of causality in this specific context by tackling the relationship between virality and musical success as a causal discovery task. In summary, causal discovery is an approach within the broader framework of causal inference that aims to infer the causal graph, whether completely or partially, of a structural causal model (SCM) based on observational or interventional data [92]. A SCM describes how each system variable is caused by the others (in addition to external noise terms), and its causal graph is a graphical and qualitative representation of the cause-and-effect relationships between all such variables.[6]

Formally, in a SCM from multivariate time series given by $\mathbf{V}_t = (V_t^i, \cdots V_t^N)$, where $t \in \mathbb{Z}$ represents time, the causal graph of this SCM is a directed graph G with vertices V_t^j for $V_t^j \in \mathbf{V}_t$ for all $t \in \mathbb{Z}$. The graph has edges $V_{t-\tau}^i \to V_t^j$ if and only if $V_{t-\tau}^i$ is a causal parent of V_t^i [92]. In our context, we can model each song as a specific system composed of two variables of interest (i.e., time series) representing virality and success. Therefore, the graph resulting from the causal discovery task can present edges between viral and successful vertices only if there is a lagged $(t-\tau)$ causal relation between them.

Similar to the previous section, we first discuss the possible outcome scenarios for the causal discovery (Section 4.5.1). Then, we present the experimental setup, i.e., the causal algorithm, stationary check, and lag definition (Section 4.5.2). Finally, we present and discuss our results (Section 4.5.3).

[6]For more details and formal definitions, see [81, 92].

4.5.1 Outcome Scenarios

The causal discovery task produces a directed graph representing the causal relationships between variables. Thus, relationships here are also unidirectional, resulting in four possible **discovery outcome scenarios** (DS):

DS1. Success is caused by virality, i.e., there are edges $V_{t-\tau}^{viral} \rightarrow V_t^{success}$ in the causal graph;

DS2. Virality is caused by success, i.e., there are edges $V_{t-\tau}^{success} \rightarrow V_t^{viral}$ in the causal graph;

DS3. Both virality and success cause each other, i.e., there are edges $V_{t-\tau}^{viral} \rightarrow V_t^{success}$ and $V_{t-\tau}^{success} \rightarrow V_t^{viral}$ in the causal graph;

DS4. There is no causal relationship between a song's virality and its success.

4.5.2 Experimental Setup

To detect the causal graph between a song's virality and its success, we use PCMCI, a constraint-based method developed for large-scale time series datasets [91]. We model and choose such a method from the QAD-template [92], which is based on Pearl's Causal Hierarchy [82] and the Causal Inference Engine [7]. Furthermore, to perform PCMCI, it is necessary to assume the stationarity of the time series. In addition, the model follows the assumptions of the causal Markov condition and causal faithfulness. The former says that, given the direct causes of a variable, that variable is conditionally independent of its non-effects, whereas the latter says that the conditional independence relationships encoded in the model are faithful to the underlying causal structure [110].

Finally, we set 14 days as the maximum lag for the PCMCI algorithm. This choice is justified according to domain knowledge and because we want to investigate the close temporal impact between virality and success. Furthermore, we use the partial correlation test to verify the independence between variables and define 0.05 as the significance level to obtain the causal graphs.

Table 4.4: Summary of Causal Discovery results (outcome scenario, number of songs, and respective percentage).

Outcome Scenario		n	%
DS1.	Success is caused by virality	119	7.50%
DS2.	Virality is caused by success	120	7.57%
DS3.	Both virality and success cause each other	515	32.47%
DS4.	No causal relationship	832	52.46%
TOTAL		**1586**	**100%**

4.5.3 Results and Discussion

In this chapter, we use the implementation of the PCMCI method of the *Tigramite* package in Python.[7] Similar to Granger Causality (GC), such a method assumes system stationarity. Therefore, we only consider songs whose viral and success time series are stationary, reducing our set of songs to 1,586, i.e., 69% of the total.

Table 4.4 summarizes the results of the causal graphs obtained for each song mapped into possible scenarios. Unlike Granger Causality, most songs do not show a causal relationship between virality and success (i.e., $DS4$ scenario). In other words, in the causal graphs of these songs, there are no edges $V_{t-\tau}^{viral} \rightarrow V_t^{success}$ or $V_{t-\tau}^{success} \rightarrow V_t^{viral}$. In these cases, both success and virality may be influenced by external factors that were not mapped in the system. Examples include the songs "...Ready For It?" by Taylor Swift and "All The Stars" by Kendrick Lamar with SZA. In fact, "All The Stars" was released as part of the soundtrack album of the movie *Black Panther*, and the movie's popularity (which is not measured by streams) may have influenced the song's success.

Still, there are songs with a causal relationship between virality and success. For instance, 119 songs are in the scenario $DS1$, in which success is caused by virality. One example is the song "DDU-DU DDU-DU" by K-pop girl group BLACKPINK, released in 2018. As it is common with pop artists, the song's initial virality was driven by a dedicated fan base, but the song reached breakthrough status, eventually achieving massive success. An indicative of this success is that the song marked the group's debut in the Billboard Hot 100, the main chart in the United States.

There are also cases where virality is caused by success (i.e., $DS2$), such as the song "Veneno" by the Brazilian singer Anitta. The song's success in Brazil pushed its virality not only in other Latin American countries but also worldwide. Furthermore, 32.47% of the songs are in the $DS3$ scenario, in which both virality and success cause each other. Examples include big hits such as "Perfect Duet" by Ed Sheeran and Beyoncé, "Downtown" by Anitta and J Balvin, and "Shake It Off" by Taylor Swift.

[7] https://github.com/jakobrunge/tigramite/

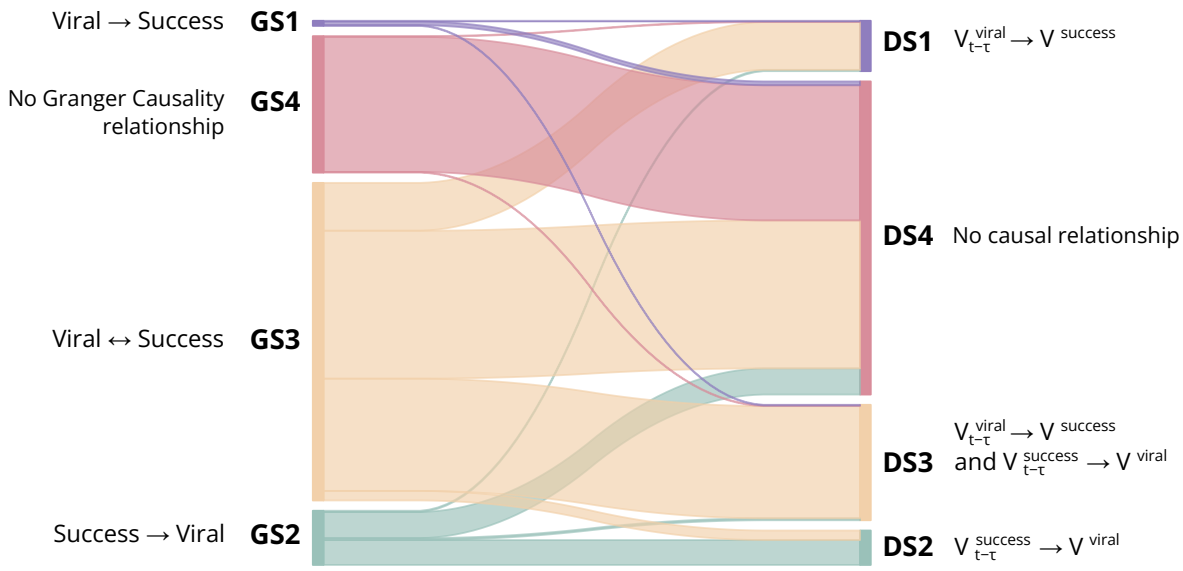


Figure 4.6: Correspondence flow between Granger (left) and discovery (right) outcome scenarios.

4.6 Correspondence Analysis

Despite being distinct methods in essence, using both Granger Causality and causal discovery helps revealing relevant information about how virality and musical success are related. Whereas the former is a purely statistical method that informs whether one variable can be used to forecast another, the latter is part of a more sophisticated causal inference framework, returning a graph with the relationships between the variables. Therefore, causal discovery allows a deeper study of causality in the musical context and can be used as an extension or complement to Granger Causality analysis.

In this sense, it is possible to relate the output scenarios of the two methods, even though Granger causality cannot be used alone to make causal inferences. For example, scenario $GS1$ (virality Granger-causes success) can be related to scenario $DS1$ (success is caused by virality), scenario $GS2$ relates to $DS2$, and so on. As the last step in our analysis, Figure 4.6 shows the correspondence between the outcomes of the two methods for the 805 songs over which both methods were executed. It illustrates a flow from one set of scenarios to another, i.e., Granger scenarios to discovery scenarios.

In general, there is a correspondence between the mapped scenarios. Most songs show the same behavior in both Granger Causality and Causal Discovery. The largest exception is the set of songs from the $GS3$ scenario (i.e., virality *Granger-causes* success and vice versa). When analyzing the causal discovery result for such songs, most of them are within the $DS4$ scenario (i.e., no causal relationship), whereas the rest are divided between $DS3$ (the corresponding scenario) and $DS1$. Such a behavior can be explained by

the nature of the methods, since causal discovery has much stronger assumptions within a more robust causal inference framework.

Finally, both approaches confirm that not all songs present an explicit causal relationship between their viral and success processes. In other words, although there is a temporal relationship between such processes for some songs, we cannot generalize that virality always precedes success or vice versa. Our findings reinforce the importance of analyzing virality and success as distinct processes, despite being closely related. Moreover, other external factors can influence the viral and success behavior of songs, including the debate and sharing on other social platforms. Such factors are not directly conveyed by the metrics provided by streaming services, which are just one part of the complex and diverse music industry.

4.7 Overall Considerations

Content virality has become an increasingly relevant phenomenon in the digital ecosystem, and for music, it has been shaping the consumption and commercial success of tracks. This chapter investigates the temporal connection between musical virality and success on digital platforms. Specifically, we build time series to represent both viral and successful behaviors on streaming charts. From such series, we first perform a correlation analysis to check the synchrony between virality and success. Next, we apply Granger Causality to check whether virality can be used to forecast success (and vice versa). We then use Causal Discovery to infer the causal relationships between the two variables.

The results show it is not possible to affirm there is always synchrony between virality and success. Moreover, virality and success can indeed be used as predictors of each other, but such a behavior cannot be generalized to all songs. In other words, there are cases where virality can be an indicator of future success; but not all songs show such a relationship. These findings are confirmed by the causal discovery analysis, which reveals that virality may cause success (and vice versa) for some songs, but not all.

Overall, our findings reinforce the characteristics of virality and success as two distinct yet interconnected facets of musical popularity. Understanding the symbiotic relationship between such concepts and their manifestation on social platforms is fundamental to explaining music consumption trends in the digital age. Our comprehensive analyses represent a first step towards understanding such complex phenomena, highlighting the importance of exploring not only the individual aspects of virality and success but also their dynamic interactions.

Limitations. The main limitation of this chapter is that the dataset comprises data from a single digital streaming platform, potentially excluding trends and dynamics from other platforms or offline music consumption. Moreover, the dataset considered does not include important music markets such as China and South Korea, i.e., the world's 5th and 7th top music markets in 2023, respectively.[8]

[8]IFPI Global Music Report 2023: https://globalmusicreport.ifpi.org/

Chapter 5

Music Virality as a Contagion Process

The analyses from the previous chapter revealed that virality and success, while interconnected, do not always evolve in synchrony. For some, virality may precede and even influence future success, whereas in others the two phenomena follow independent paths. Such findings reinforce the notion that virality and success represent distinct yet interrelated facets of the broader concept of musical popularity, often shaped by the dynamics of social media platforms. Building on this perspective, we now investigate how such dynamics can be represented through the lens of contagion processes in the digital era.

Indeed, the spread of songs across such digital platforms shares several similarities with the propagation of infectious diseases. For instance, as diseases propagate through contact between individuals, songs spread through social interactions, with users sharing songs with friends or reposting them on social media. In a broader context, previous work has applied epidemic models to describe the diffusion of various types of online content. In the music context, such studies investigate the dynamics of song popularity represented as downloads [60, 89] or views [51, 93].

Given the growing role of social interactions in driving music consumption, modeling the music virality and success on streaming platforms as contagion processes represents a powerful framework for understanding the complex dynamics of such phenomena. Hence, in this chapter, we investigate whether epidemic models can effectively represent music popularity on social platforms. Specifically, we address the following questions: *"Are epidemic models suitable for representing music popularity on streaming platforms?"* *"How accurately can such models forecast the popularity trajectories of songs?"*

Following the analyses from the previous chapter, we also use time series derived from data obtained from Spotify charts (Section 5.1). Regarding the methodology, we: (i) apply epidemic models to streaming data to capture music popularity dynamics (Section 5.2); (ii) introduce a wave-based modeling approach that better reflects the nature of viral diffusion on streaming (Section 5.3); and (iii) evaluate the forecasting performance of our method against traditional time-series forecasting methods (Section 5.4).

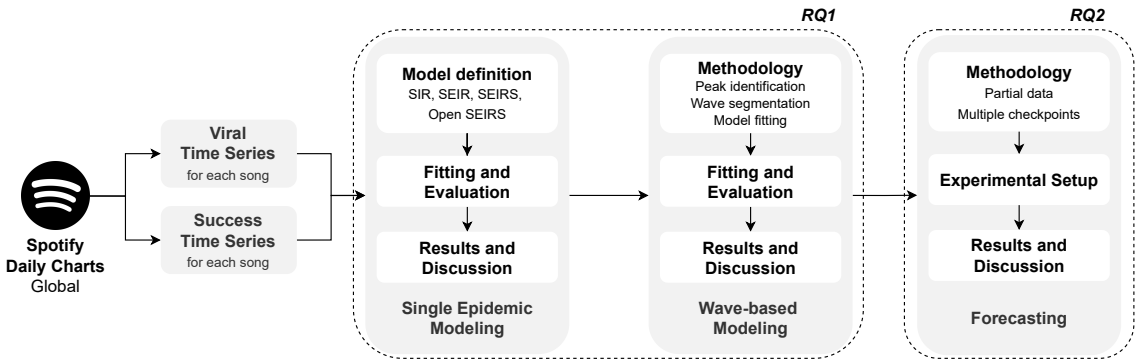


Figure 5.1: Overview of the analysis conducted using our data and time series modeling.

5.1 Data and Time Series Modeling

Similar to the previous chapter, we rely only on Spotify data to represent music popularity. For each market (and for the Global aggregate), Spotify provides two daily charts: the Top 200, representing success through the most-streamed songs, and the Viral 50, reflecting virality based on play growth, sharing speed, and discovery patterns.[1] Again, we use the daily Global charts covering January 2017 to March 2022,[2] from which we build time series for each song to capture its popularity trajectory. Popularity is measured through the rank score, computed from the chart position and set to zero whenever the song does not appear on the charts (see Sections 4.1 and 4.2).

After building the time series, we perform some preprocessing steps to prepare them to be the input of the epidemic models. First, we apply a min-max normalization in all time series to be within the interval $[0, 0.5]$. We choose such limits because since the virality time series will represent the evolution of individuals within the infected state (see the next section), not all individuals may be infected simultaneously in a closed epidemic (i.e., when the population is fixed). Therefore, we assume that the maximum proportion of infected individuals is 50%. In addition, when the songs did not reach the charts, we set the normalized rank score to 0.001, meaning that the song had some popularity, but it was not enough to reach the charts.[3]

Next, inspired by the work of Sachak-Patwa et al. [93], we take a simple moving average of each time series to reduce the noise and smooth the fluctuations in popularity. For the sake of this chapter, we consider a simple moving average of seven days. Moreover, we only consider songs that were present in the charts for more than a week. Therefore,

[1] Spotify: https://support.spotify.com/us/artists/article/charts/

[2] On March 2022, Spotify Charts changed its platform, and it was no longer possible to download the CSV files with the charts. Again, it was still the only platform that provided so many options of charts.

[3] Other normalization strategies are also possible, such as adding the noise first and then scaling by the maximum possible rank score.

our dataset contains 1,647 viral and 1,725 successful (i.e., hit) songs. There is an overlap between the two sets, as some songs appear at least once on both charts. From such data, we perform three distinct analyses to answer our two research questions, which are illustrated in Figure 5.1 and detailed next.

5.2 Single Epidemic Modeling

In this section, we address the first research question (*Are epidemic models suitable for representing music popularity on streaming platforms?*) by modeling music popularity using single epidemic models. In other words, we consider the song's viral/success trajectory as a single epidemic process. Our assumption is that music popularity is closely related to its consumption patterns. Specifically, when an individual is impacted by a song (either positively or negatively) upon listening, they are more likely to engage in social sharing – whether by recommending it or simply discussing it with others. Following prior work on video popularity [51, 89], we consider compartmental models to represent such a phenomenon. We aim to verify whether this modeling type can reflect music popularity and, if yes, what is the best model for it. In Section 5.2.1, we present the models used in this analysis. Then, Section 5.2.2 details the evaluation setup of the fitting process. Finally, Section 5.2.3 presents and discusses the results for all considered models.

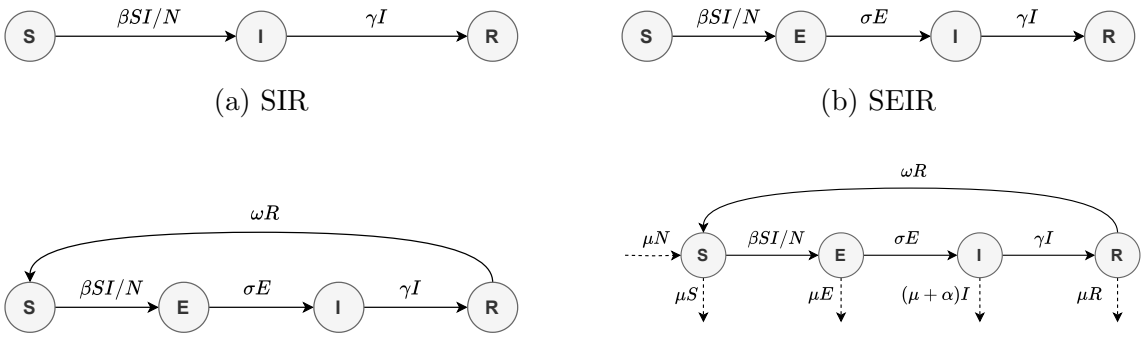
5.2.1 Model Definition

Living in society promotes the adoption of behaviors, emotions, and conditions by individuals through contact, a process formally defined as **social contagion**. This concept describes how ideas, behaviors, or emotional states can spread within a population in a way similar to the transmission of infectious agents in epidemiology. The most familiar application of contagion models involves disease dynamics, exemplified by the recent COVID-19 pandemic. However, the scope of social contagion extends well beyond this, incorporating the dissemination of information, knowledge, and collective digital behaviors [2, 59, 122].

There are several models for representing contagion processes. The most traditional ones (i.e., **compartmental models**) consider only pairwise relationships between individuals and assume population homogeneity when representing the infection mecha-

Table 5.1: Summary of the model states and parameters used in this thesis.

Term	Description	SIR	SEIR	SEIRS	Open SEIRS
	States				
S	Susceptible individuals	✓	✓	✓	✓
E	Exposed individuals		✓	✓	✓
I	Infected individuals	✓	✓	✓	✓
R	Recovered individuals	✓	✓	✓	✓
	Parameters				
β	Infection rate (from S or E to I)	✓	✓	✓	✓
γ	Recovery rate (from I to R)	✓	✓	✓	✓
σ	Latency rate (from S to E)		✓	✓	✓
ω	Loss-of-immunity rate (from R back to S)			✓	✓
μ	Birth/death rate				✓
α	"Death due to infection" rate				✓



(a) SIR (b) SEIR

(c) SEIRS (d) Open SEIRS

Figure 5.2: Compartmental models considered in this thesis.

nism [33]. Here, we consider four different compartmental models for our music popularity time series: SIR, SEIR, SEIRS, and Open SEIRS [12, 13]. Although our time series do not directly measure the number of individuals impacted by a song, they can reasonably be considered as proxies for the infection curve in such models.

Regarding notation, each model is described by a system of differential equations, in which each equation represents the variation of a state over time t. We use italic uppercase letters to denote the model compartments (i.e., the possible states of individuals in the system) and the total population N. Moreover, the parameters representing the rate of change of each state are denoted by lowercase Greek letters. Table 5.1 summarizes the states and parameters from all models, while Figure 5.2 illustrates their dynamics.

SIR model. This model considers a three-state epidemic, in which individuals can be either susceptible (S), infected (I), or recovered (R). It considers a fixed population of $N = S + I + R$ and a closed epidemic, i.e., there is no increase or decrease in the total population over time (Figure 5.2a). The number of individuals in each state is a function of time t, with transitions from susceptible to infected occurring at rate β, and from infected to recovered at rate γ. Here, susceptible means users who have not been exposed

to a given song, but may do so in the future. In contrast, infected individuals are those who are actively contributing to the spread of a song by streaming (for success) or sharing (for virality). Finally, the recovered state means that a person who has lost interest in a song and stopped consuming it. The rates of change of each state are given by Equations 5.1, 5.2, and 5.3.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\beta SI}{N} \quad (5.1) \qquad \frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\beta SI}{N} - \gamma I \quad (5.2) \qquad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I \quad (5.3)$$

SEIR model. This model builds upon SIR by adding a new state E between susceptible and infected, in which people are exposed before being actually infected (Figure 5.2b). In our context, an exposed individual means someone who has encountered the song indirectly but has not yet actively engaged with it. In other words, this additional state allows capturing the delay between initial exposure and active engagement. Individuals transition from the exposed to the infected state at a rate σ, while the rest of the dynamics follow similar principles to the SIR model. The rate of change for each state is defined by Equations 5.4, 5.5, 5.6, and 5.7.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\beta SI}{N} \quad (5.4) \qquad \frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \gamma I \quad (5.6)$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\beta SI}{N} - \sigma E \quad (5.5) \qquad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I \quad (5.7)$$

SEIRS model. It extends SEIR by allowing individuals in the recovered state to return to the susceptible state, introducing the possibility of reinfection (Figure 5.2c). In our context, this reflects the scenario in which users who have previously lost interest in a song may re-engage with it after some time. The transition from recovered to susceptible occurs at a rate ω, and the dynamic transitions between all four states are described by Equations 5.8, 5.9, 5.10, and 5.11.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \frac{\beta SI}{N} + \omega R - \mu S \quad (5.8) \qquad \frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \gamma I - (\mu + \alpha)I \quad (5.10)$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\beta SI}{N} - \sigma E - \mu E \quad (5.9) \qquad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \omega R - \mu R \quad (5.11)$$

Open SEIRS model. It introduces population dynamics into the previous SEIRS model, allowing individuals to enter and exit the system over time (Figure 5.2d). In

our context, new listeners join the platform (births), and others become inactive or leave (deaths). The birth and death rates are represented by μ, and while births are only accounted for in the susceptible state, deaths can happen in all of them. There is also an additional "death due to infection" rate α on the infected state, which can represent users who were actively streaming/sharing the song, but permanently disengaged from it due to saturation or shifts in taste. The rest of the dynamics are similar to the simpler SEIRS model (Equations 5.12, 5.13, 5.14, and 5.15).

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu N - \frac{\beta S I}{N} + \omega R - \mu S \tag{5.12}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\beta S I}{N} - \sigma E - \mu E \tag{5.13}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \gamma I - (\mu + \alpha) I \tag{5.14}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \omega R - \mu R \tag{5.15}$$

5.2.2 Model Fitting and Evaluation

We now define the initial conditions required for fitting the four models to each time series. Since the time series are normalized, we set the total population N to 1.0 in all cases. The initial number of infected individuals, I_0, corresponds to the first observed value in the time series. For all models, the susceptible population at time zero is defined as $S_0 = N - I_0$. All other compartments, i.e., exposed (E_0) and recovered (R_0) are initialized as zero, because we assume that a song's popularity starts with no prior exposure.

We use the *SciPy* Python library[4] to estimate the models' parameters for each popularity time series. We use the least squares approach to perform parameter fitting. For each parameter (i.e., β, γ, σ, ω, μ, α), we set an initial guess of 0.5 and a lower bound of 0 to ensure valid values. Moreover, to evaluate the models' accuracy, we use the Root Mean Squared Error (RMSE) over the whole time period (including when the song is not in the chart), which quantifies the deviation between the observed data and the fitted curve. In our analysis, RMSE values range from 0, indicating a perfect fit, and typically approach 0.5 in poor fits, given the normalized time series. However, since the epidemic model curves are not strictly bounded during fitting, they may exceed the normalized range, meaning there is no fixed upper limit for this metric.

[4]SciPy: https://scipy.org/

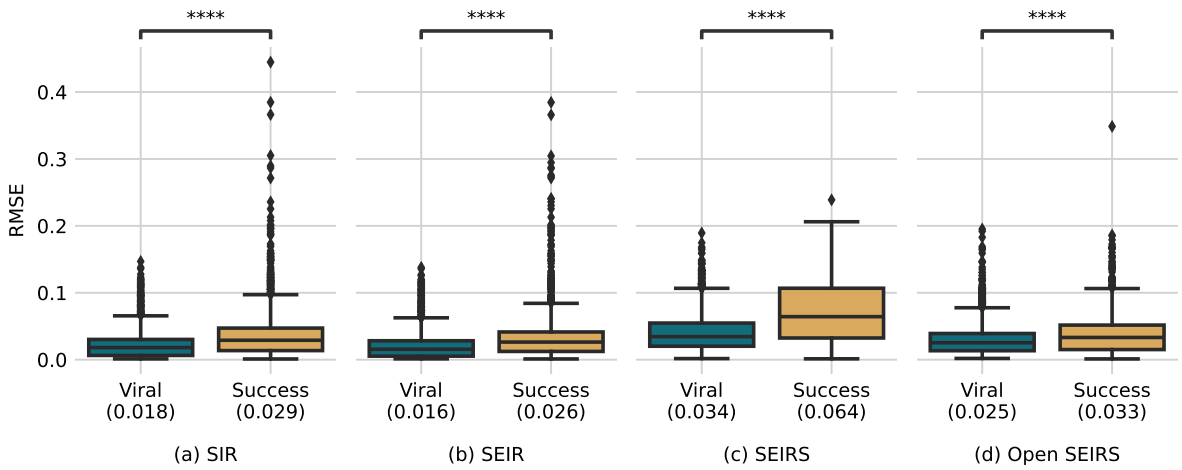


Figure 5.3: RMSE for virality and success curves using single epidemic models. Values in parentheses are the median values. Significance is calculated using the Mann-Whitney U test: **** indicates $p \leq 0.0001$.

5.2.3 Results and Discussion

To verify whether epidemic models are suitable for representing music popularity, we evaluate the fitting results for both virality and success time series. Figure 5.3 illustrates this comparison by showing the distribution of the RMSE values grouped by the four considered epidemic models. The results show that all models performed better for virality time series when compared with the success ones, suggesting that the viral sharing of a song follows a more epidemic-like pattern than its listening behavior measured by streams. In fact, online virality reflects a fast and short-term sharing behavior by definition [31], while success may also be related to external factors (e.g., marketing, artist popularity, playlist placement) that the traditional epidemic models do not capture.

Furthermore, when focusing specifically on the virality time series, SEIR produces the best overall fitting performance among the four models considered (median RMSE of 0.016 for virality). Such a finding aligns with the results of previous work on online content popularity, highlighting the SEIR effectiveness in capturing both the initial popularity growth and its subsequent decline [93].

However, there are songs for which the considered epidemic models struggle to capture multiple and spaced peaks of virality over time. Even the SEIRS and Open SEIRS models, which allow individuals to be reinfected (i.e., to return to the susceptible state), tend to reach an equilibrium in the long term. An example is the song "Mon Amour - Remix" by Zzoilo and Aitana (Figure 5.4), which has two explicit viral moments after its release. However, none of the four models can capture them correctly, highlighting the need for more sophisticated approaches that capture such complex dynamics.

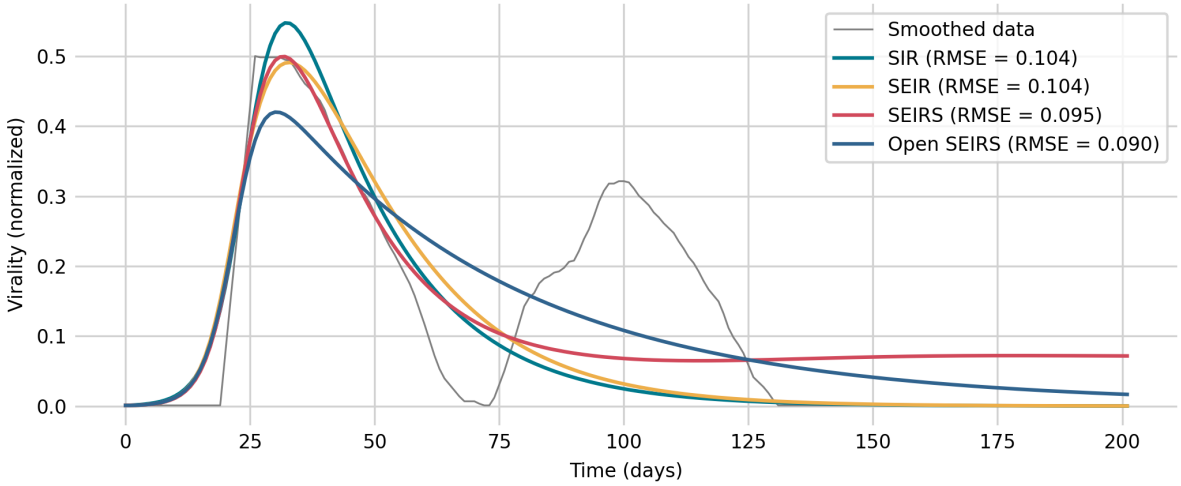


Figure 5.4: Virality time series with its respective model fits for the song "Mon Amour - Remix" by Zzoilo and Aitana.

Recalling our research question, epidemic models can represent music popularity on streaming platforms to some extent, being more suitable for representing virality rather than success. This reflects a key difference between the two processes: while the first is fast and ephemeral (similar to several infectious diseases), the latter may be longer and influenced by other external factors that are not easily captured by simple epidemic models. However, despite the good results of epidemic models for virality (particularly SEIR), the existence of songs with multiple viral moments opens space for questioning the limitations of the models used and proposing more flexible approaches for representing music virality over time.

5.3 Wave-based Epidemic Modeling

We now propose a novel wave-based approach to model music virality on streaming platforms. From now on, we focus exclusively on virality rather than success, as the latter is not well captured by epidemic models. The central assumption of our approach is that each wave of music virality can be analyzed as a distinct epidemic, potentially leveraged by different factors (e.g., a remix, a viral trend, or renewed exposure on social media). The motivation comes from viral infection epidemics, such as COVID-19, which manifested in multiple waves over time. Each wave was driven by a different virus variant (e.g., original strain, delta, and omicron), each with its own transmission dynamics and characteristics. Our wave-based approach is detailed in Section 5.3.1, followed by the evaluation setup in Section 5.3.2, and the results and discussion in Section 5.3.3.

5.3.1 Methodology

Our proposed approach is based on the assumption that every song may have multiple virality periods (waves), each one with its own dynamics. Initially, the main goal of this method is to model and understand the dynamics of music virality, and therefore it works *a posteriori*, i.e., we rely on the complete time series in the fitting process. In short, our approach is composed of three main steps: (i) peak identification, (ii) wave segmentation and adjustment, and (iii) epidemic model fitting, which are described next.

Peak identification. From the preprocessed time series, we use a peak detection method[5] to identify significant local maximum points that represent distinct moments of virality. Candidate peaks are local maxima of the time series. To ensure that each peak corresponds to an independent event, we define a minimum distance of 30 days between consecutive peaks. If the consecutive candidate peaks are closer than 30 days, we discard the lower ones. The procedure is repeated iteratively over all candidate peaks.

Wave segmentation and adjustment. The peak detection method also returns the left and right base points for each identified peak, which we initially consider as the start and end of each virality wave. Such bases correspond to the lowest points surrounding the peak and are determined by scanning outward from the peak until reaching a local minimum on each side. First, we set a minimum width of 7 days for each wave (when $rank\_score > 0.001$), to filter out short-lived spikes that do not represent sustained viral behavior. Moreover, overlaps between waves may occur, especially when peaks are close together. We address this by adopting an independent wave approach where each wave is treated as a separate and self-contained event, which is not affected by adjacent waves. Specifically, when there is an intersection between two waves (i.e., when the right base of the first wave is after the left base of the second one), we shift the starting point of the second wave to immediately follow the end of the first. If the resulting wave has length zero (i.e., it is entirely within the prior), we discard it. We do this because we assume that one wave does not receive any impact from the past, nor does it impact the future.

Epidemic model fitting. Having clearly defined waves, we fit an epidemic model to each one independently. We use the SEIR model based on our analysis of model performance for virality (see Section 5.2). This is also in line with previous works that state that such a model captures the onset of each wave better [93].

[5]We use the `find_peaks` function of the SciPy package: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html

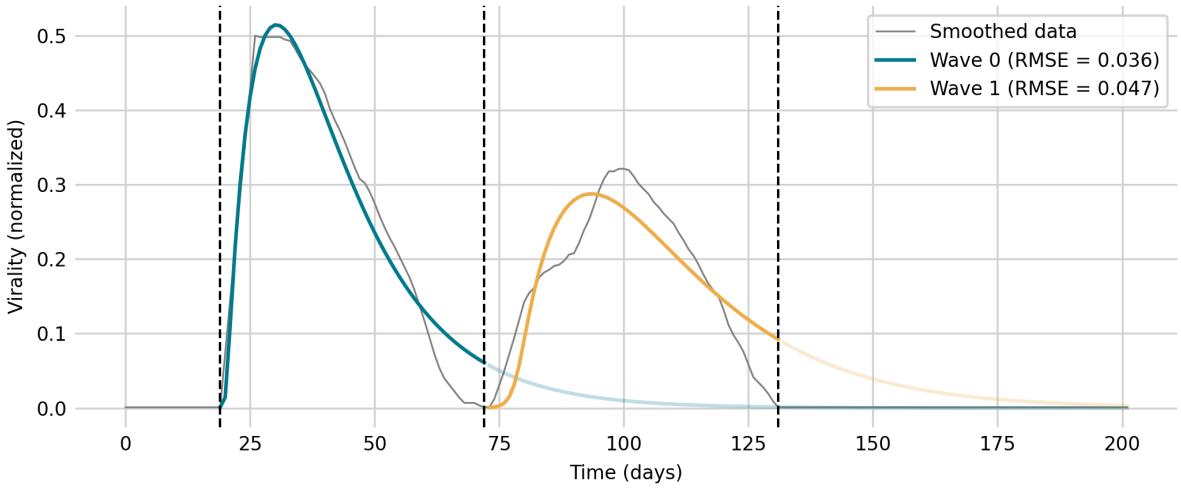


Figure 5.5: Virality time series with the wave-based SEIR fit for the song "Mon Amour - Remix" by Zzoilo and Aitana. The vertical dashed lines delimit the waves.

5.3.2 Fitting and Evaluation

To implement our wave-based approach, we use the `find_peaks` function from the *SciPy* library to identify individual moments of virality within each time series. Once the waves are segmented and adjusted, we fit an independent SEIR model to each wave separately, and the initial conditions follow the same setup described in the single-model approach (Section 5.2.2).

To evaluate the performance of each wave fit, we compute the RMSE considering only the segment of the time series between the wave's defined start and end points. Such an evaluation allows assessing how well the model captures each individual virality moment. Then, for songs that have multiple waves, we report the overall performance using the average RMSE across all waves.

5.3.3 Results and Discussion

Since we now fit the SEIR model to each virality wave independently, some songs cannot be considered due to limitations in the model fitting process, namely, the absence of waves after adjustments. As a result, the dataset used for our wave-based analysis comprises 1,045 viral songs (63.4% of the original set). The median value for the average RMSE across all fitted songs is 0.061, indicating a generally good alignment between the model and the observed data. Whereas this value is slightly higher than the median RMSE

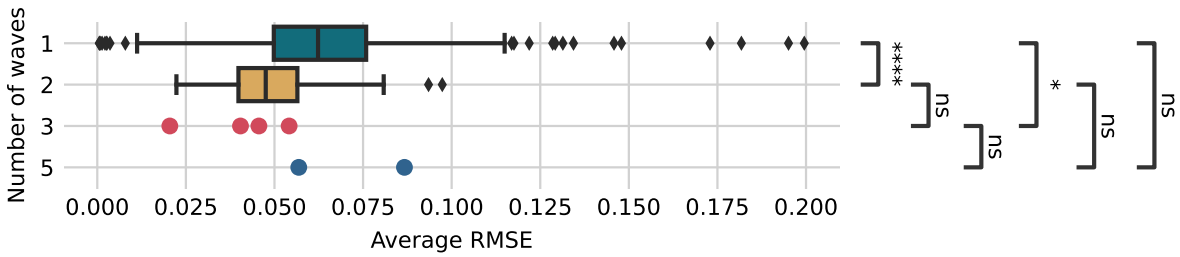


Figure 5.6: Average RMSE distribution grouped by the number of identified waves. Significance is calculated using the Mann-Whitney U test: * for $0.01 < p \leq 0.05$; ** for $0.001 < p \leq 0.01$; *** for $0.0001 < p \leq 0.001$; **** for $p \leq 0.0001$; and 'ns' otherwise.

Table 5.2: Descriptive statistics of the SEIR parameters in our wave-based approach.

Primary SEIR parameters				
	Min.	**Mean**	**Median**	**Max.**
Average infection rate (β)	8.235×10^{-2}	991.970	15.770	1.018×10^{6}
Average recovery rate (γ)	1.605×10^{-7}	0.175	0.113	6.150
Average latency rate (σ)	1.300×10^{-2}	2269.111	0.223	6.271×10^{5}
Derived parameters				
	Min.	**Mean**	**Median**	**Max.**
Average infectious period ($1/\gamma$)	0.162	1.202×10^{4}	9.181	6.230×10^{6}
Average R_0 (β/γ)	1.092	3.342×10^{4}	161.042	3.353×10^{7}

of the single-model approach, our wave-based method offers a more realistic representation of virality patterns by capturing multiple engagement periods. For example, our approach can capture both virality waves on the song "Mon Amour - Remix", which is not possible using the previous approach (Figure 5.5).

Regarding the number of waves, the vast majority of songs (994, or approximately 95%) have only one identified wave. Therefore, only a smaller portion has more complex dynamics, with 45 songs having two waves, four songs having three, and only two songs reaching five distinct virality waves. The average and median wave lengths are 38 and 32 days, respectively. Figure 5.6 presents the distribution of the average RMSE by the number of identified waves. For songs with three and five waves, we show the individual points instead of boxplots due to the small number of samples. In general, there is no statistically significant difference in RMSE across most groups, except for a slight difference between songs with one and two or three waves. However, the median values remain relatively close, indicating that our approach maintains a consistent performance even when the number of waves increases.

SEIR Parameters. A major strength of using epidemic models such as SEIR lies in the interpretability of their parameters, which may offer valuable insights into the dynamics of music consumption. Table 5.2 presents descriptive statistics for the primary SEIR parameters (i.e., β, γ, σ). Since a song may have multiple virality waves, we choose to

Table 5.3: Top 5 songs with highest average infection rate β.

Song	**Artists**	β	γ	σ	**RMSE**
Glorious	Macklemore, Skylar Grey	1.013×10^6	0.030	0.013	0.148
Adan y Eva	Paulo Londra	369.256	0.023	0.064	0.087
Dark Red	Steve Lacy	291.631	0.071	0.070	0.090
Notion	The Rare Occasions	240.981	0.031	0.066	0.102
a lot	21 Savage	236.768	0.052	0.068	0.097

report the average value of each parameter per song. Such parameters are usually within the range $[0, 1]$ but there is not a real upper bound since they depend heavily on the shape and scale of each time series.

In our context, the average infection rate (β) measures how fast a song spreads among users, making it a key parameter when analyzing how fast it goes viral. A higher β indicates that a song spreads very quickly in the population, possibly due to strong word-of-mouth combined with marketing strategies. The median value of more than 15 suggests that songs gain traction relatively quickly. However, notice that this may partly reflect: (i) the lack of data before songs enter the Viral 50 chart, limiting our view of early virality growth, and (ii) the normalization to 0.5, likely overestimating the fraction of infected people at the peak. Table 5.3 contains the five songs with the highest average infection rates. The extreme values happen because all such songs have a high virality rank score at the beginning of the wave, requiring a high β to fit the curve accurately.

The average recovery rate (γ) reflects how quickly users lose interest in a song once they have started to engage with it, and higher values mean that people lose interest more quickly. The median value of 0.113 suggests that engagement tends to last for a reasonable period before fading. This leads to longer tails, i.e., long periods in the low positions of the chart. In contrast, the average latency rate (σ) captures how fast individuals move from the exposure to a song to the active engagement with it, with higher values meaning a faster adoption. The median value of 0.223 indicates that, in general, people take some time after being exposed to a song before deciding to share it.

Derived parameters. From the primary SEIR parameters, we can also derive meaningful insights. For example, the infectious period ($1/\gamma$) estimates how long a user stays engaged with a song after discovering it. The median value of around nine days aligns with our previous findings that viral songs usually stay popular for a week or two before fading (see Chapter 3). Another important parameter is the basic reproduction number ($R_0 = \beta/\gamma$), which represents how many new users a single engaged person is expected to influence. A median R_0 of 161.04 suggests that viral songs have strong contagious potential, reinforcing the idea that music virality is a phenomenon with similar mechanisms to traditional epidemics.

Overall, our wave-based approach for modeling music virality complements the findings from the previous section and answers our first research question (*Are epidemic*

models suitable for representing music popularity on streaming platforms?). Indeed, this approach can effectively represent the dynamics of music virality on streaming platforms, especially when songs have multiple periods of virality. Moreover, given its interpretability and ability to reflect song diffusion patterns, our epidemic approach may also be a valuable tool for forecasting music consumption trends, a hypothesis that we explore next.

5.4 Forecasting Virality Behavior

We now use the wave-based approach to address the second research question (*How accurately can epidemic models forecast the popularity trajectories of songs?*). Motivated by the parallels between music virality and real-world epidemics such as COVID-19, we explore the potential of our approach to forecast virality trends using only partial time series data. Inspired by prior work on online social dynamics [87], we aim to understand whether early virality signals can be used to anticipate a song's future trajectory on streaming platforms. After presenting our forecasting methodology (Section 5.4.1), we detail our experimental setup (Section 5.4.2) and discuss the results (Section 5.4.3).

5.4.1 Methodology

Our methodology operates at the individual wave level to evaluate SEIR model's forecasting capabilities in the context of music virality. This choice is aligned with our wave-based approach, in which each viral moment is treated independently, mirroring the behavior of separate outbreaks in epidemiological models. Unlike systems that may require dynamic detection of waves, here we assume the waves are already identified a-priori, based on the complete time series. This setup allows us to verify how well the SEIR model can forecast a song's future virality once a wave has already begun. In practical terms, this simulates a scenario in which a song starts gaining traction on social media or streaming platforms, and we want to guess how far and fast it might spread.

To perform this forecasting, we use partial information from the original smoothed virality time series. Specifically, for a given time point t after the beginning of the wave, we use all observed data from the start of the wave up to time t to fit a new SEIR model. This fitted model is then used to predict the subsequent wave evolution beyond t. By repeating this process across multiple time checkpoints within each wave, we can evaluate

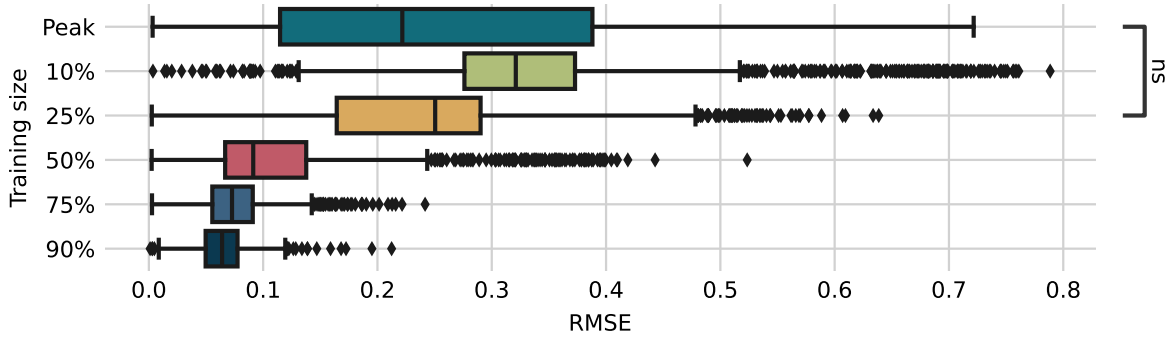


Figure 5.7: Forecast RMSE for different partial data sizes. Unless specified with 'ns', all pairs of distributions are statistically different with $p \leq 0.0001$ (Mann-Whitney U test).

how early and how accurately the model can forecast the complete virality pattern.

5.4.2 Experimental Setup

To evaluate the forecasting performance of the SEIR model, we perform two main experiments at the wave level. In the first experiment, we simulate different forecast horizons by fitting the model using partial data from the beginning of the wave up to specific cut-off points. Specifically, we consider up to 10%, 25%, 50%, 75%, 90%, and the peak point of the virality curve. With this experiment, we aim to verify the earliest point at which there is a reasonable forecasting and how the performance evolves as more data becomes available.

Following the literature on time series forecasting [43], the second experiment complements the first one and compares SEIR against three baseline forecasting methods: (i) **ARIMA**, a classical statistical model; (ii) **SVR** (Support Vector Regression), which captures non-linear trends; and (iii) **Prophet**, a tool designed for handling time series with seasonality and irregularities [113]. For each baseline, we use the implementation provided by libraries *pmdarima*[6], *scikit-learn*[7], and *prophet*,[8] respectively, all with default parameter settings. To ensure a consistent and fair comparison of the models' performance, we evaluate the forecasting using RMSE only over the remaining wave portion that was not used during the fitting process (i.e., after the cut-off time t).

[6] https://alkaline-ml.com/pmdarima/
[7] https://scikit-learn.org/
[8] https://facebook.github.io/prophet/

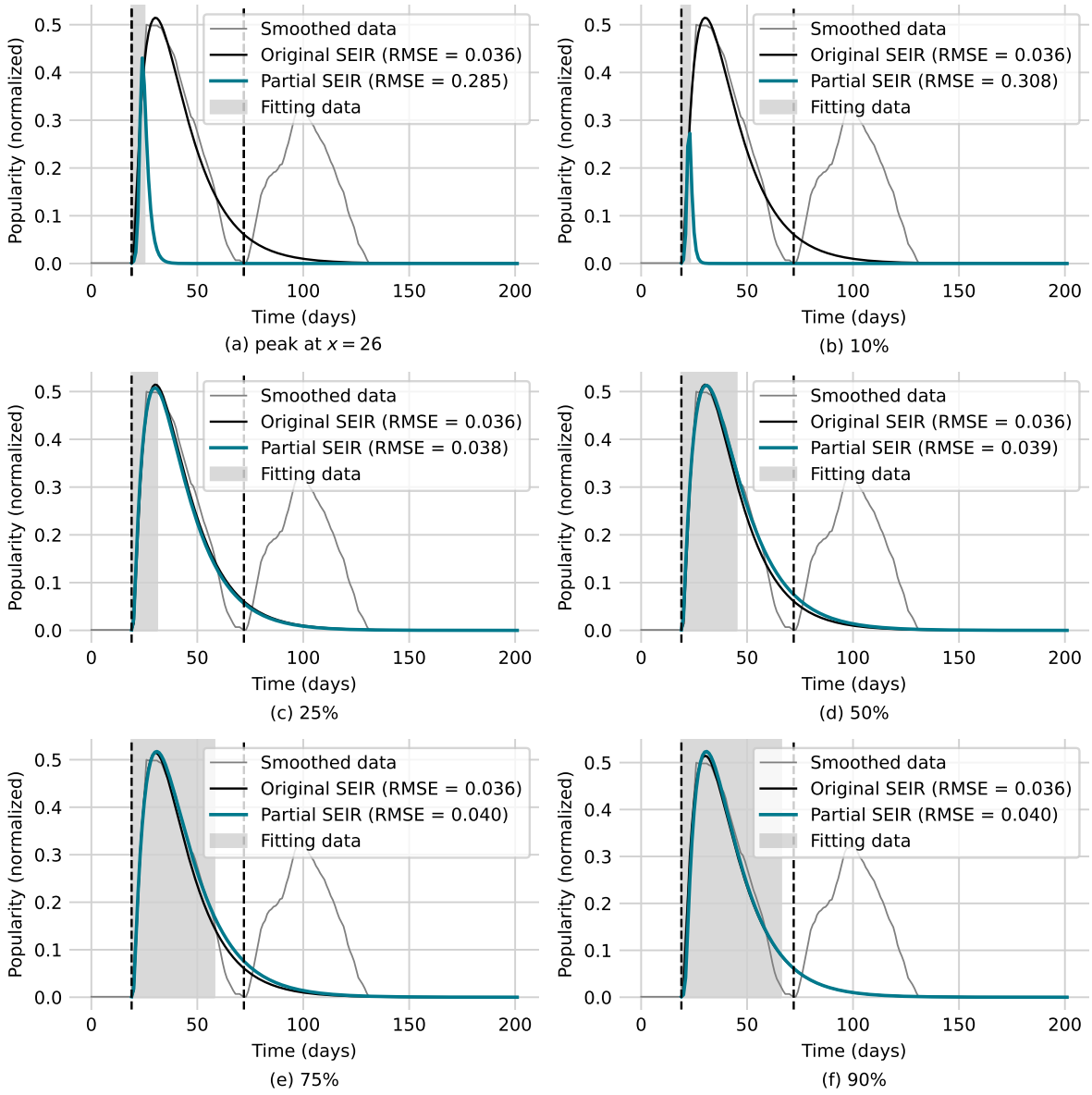


Figure 5.8: Forecast with different partial data sizes for the first virality wave of the song "Mon Amour - Remix" by Zzoilo and Aitana. Vertical dashed lines delimit the wave.

5.4.3 Results and Discussion

The forecasting results are summarized in Figure 5.7, which presents the RMSE distribution for each fitting data size. As expected, increasing the amount of available data generally improves the forecasting performance, and all pairwise comparisons between training data sizes have statistically significant differences, except the pair between Peak and 25%. Indeed, the peaks in median occur at 33% of the wave time. In addition, using 50% of the wave to fit the SEIR model produces a median RMSE of 0.091, which is only slightly higher than the 0.061 obtained using the whole wave. This suggests that the

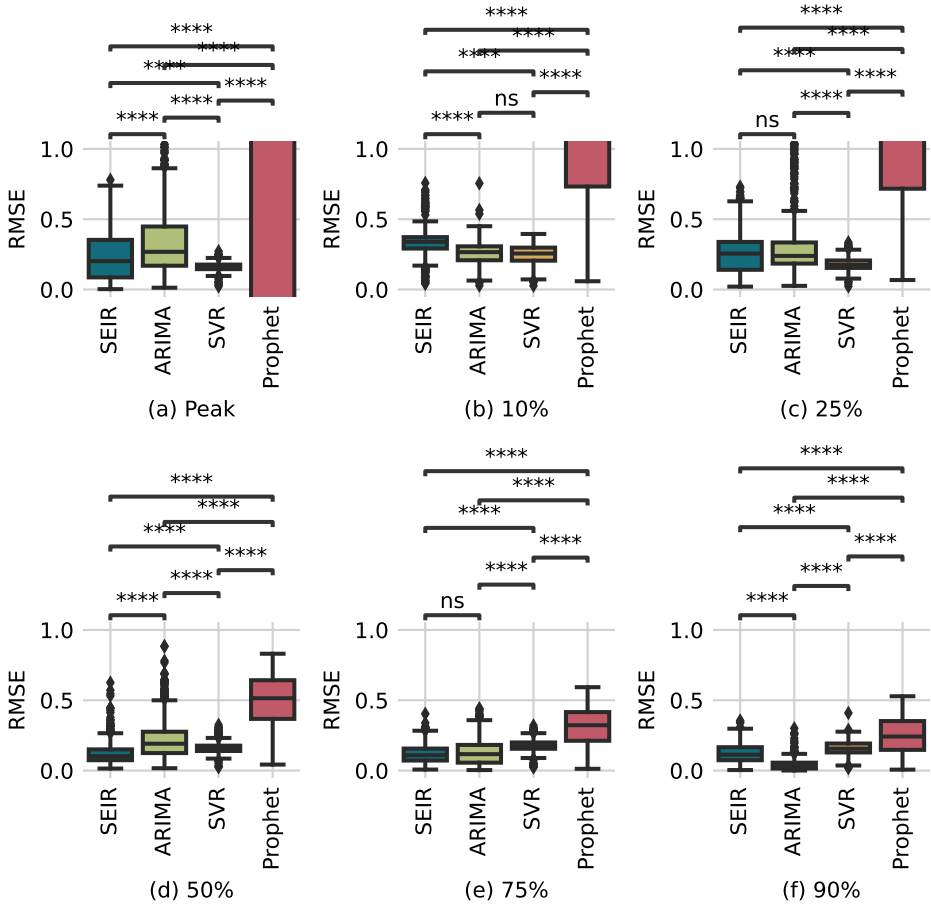


Figure 5.9: Forecasting performance for our approach with baselines. Significance is calculated using the Mann-Whitney U test: * for $0.01 < p \leq 0.05$; ** for $0.001 < p \leq 0.01$; *** for $0.0001 < p \leq 0.001$; **** for $p \leq 0.0001$; and 'ns' otherwise.

model can still produce acceptable forecasts of a song's virality even with partial data. An example is illustrated by Figure 5.8, which contains the forecast results for the first virality wave of the song "Mon Amour - Remix".

For the comparison with baseline models, we consider only the subset of 548 songs for which all models (SEIR, ARIMA, SVR, and Prophet) successfully completed the forecast. Figure 5.9[9] reveals that no single model dominates across all training conditions. However, SEIR performs consistently well, outperforming others at Peak and 50%. In contrast, ARIMA tends to be better when more data is available (especially at 90%), capturing the final virality decays at the end of waves. SVR and Prophet show competitive results in certain conditions, but they generally have higher variability than SEIR.

Overall, our findings answer the second research question (*How accurately can such models forecast the popularity trajectories of songs?*), indicating that our epidemic

[9]We set the upper limit of the y-axis to 1 to prevent distortions caused by extreme errors, particularly in Figures 5.9(a–c), in which Prophet failed to capture the underlying dynamics and produced significantly large forecasting errors.

approach can forecast music virality with reasonable accuracy, especially once a wave has begun. Its performance is generally comparable to traditional time-series forecasting methods such as ARIMA, with the advantage of offering interpretable parameters that provide insights into the underlying dynamics of music consumption. For instance, β, γ, and σ reveal how quickly a song spreads, how long users stay engaged, and how fast exposure turns into active engagement.

5.5 Overall Considerations

In this chapter, we explore the use of epidemic models to represent and forecast music popularity on streaming platforms, namely Spotify. By distinguishing virality from long-term success, we first evaluate how classic epidemic models fit both types of popularity trajectories. The analysis reveals that such models are better suited to capturing virality, as they naturally align with the fast, contagious spread of content driven by social interactions. In contrast, long-term success follows more stable and sustained patterns that are not fully captured by traditional epidemic dynamics.

Building on this insight, we propose a novel wave-based modeling approach designed to capture multiple bursts of popularity, i.e., independent periods in which a song gains public attention, often due to remixes, challenges, or renewed exposure on social media. Our results demonstrate that epidemic-based approaches not only represent viral dynamics with high interpretability but also achieve forecasting performance comparable to traditional time-series methods. All such findings highlight the potential of contagion-based frameworks for understanding music consumption, with practical applications for online trend detection and marketing strategy development.

Limitations. The main limitations of our analyses are related to some design choices and the input data. First, our forecasting experiments rely on static, pre-identified wave boundaries, which constrain real-time applicability since the onset of a new wave is not known in advance. Then, the discrete nature and undisclosed calculation method of Spotify charts (i.e., the basis of our time series) limit the interpretability of our results. Moreover, the analyses in this chapter are restricted to Spotify data, particularly the Viral 50 charts, which capture only part of the dynamics of music popularity. To verify the generalization of our approach, it is essential to test it on other platforms, such as TikTok, which have been driving music virality in recent years.

Chapter 6

Music Virality on Other Platforms: A Case Study on TikTok

In the previous chapter, we demonstrated that epidemiological models are suitable for capturing the dynamics of music virality on streaming platforms, namely Spotify. Indeed, the methodology allows the derivation of interpretable parameters that describe specific processes of this phenomenon, such as the speed at which a song spreads, the period it remains viral (i.e., the time in which a song is in virality charts), and other temporal aspects of its popularity trajectory. The proposed wave-based approach is also proven to capture the multiple virality peaks that songs may have on Spotify, highlighting even more the suitability of epidemic modeling to this problem.

However, the main limitation of such an approach is that it was designed and validated only for Spotify data. Nowadays (as of December 2025), short-video platforms such as TikTok stand out as key venue where music virality can be observed. Such platforms allow songs to be easily embedded in content, serving either as background tracks for everyday videos, dance challenges, or as part of promotional content published by artists and record labels. Therefore, music is often consumed passively, as users repeatedly encounter tracks while scrolling through content. This constant exposure enables songs to become widely recognizable, even without active searching or intentional listening.

The importance of TikTok in driving music virality is also highlighted by recent statistics. According to the platform itself, in 2024, 84% of the songs that reached the Billboard Global 200[1] first went viral on TikTok.[2] An example is the song "Gata Only" by Chilean artists FloyyMenor and Cris Mj, which became the number one track on TikTok worldwide, comprising over 50 million video creations and 1.3 billion Spotify streams.[3] The song climbed to #2 on Spotify's Global chart and also became a hit in the United States, debuting at number #98 on the Billboard Hot 100 (i.e., the main song ranking in the US), later peaking at #27 within five weeks.

[1]The Billboard Global 200 is a weekly chart that ranks the world's best-performing songs based on a combination of digital sales and streaming data from over 200 territories worldwide.

[2]https://newsroom.tiktok.com/tiktok-and-luminate-release-latest-music-impact-report

[3]https://www.musicbusinessworldwide.com/tiktok-reveals-its-top-songs-of-2024-says-that-13-of-16-no-1-hits-in-the-us-this-year-are-linked-to-trends-on-its-platform/

Building on such insights, this chapter deepens the analyses of the previous chapter on RG3 (*Model music virality as a process of social contagion for a deeper understanding of the dynamics of music viralization on social platforms.*) by investigating whether epidemiological modeling can still be used for representing music virality on online platforms other than Spotify. Specifically, we present a case study using TikTok data to replicate our methodology on them and also to assess whether the behavior aligns with our previous findings, thus testing the generalizability of the results obtained previously.

The remainder of this chapter follows a structure similar to that adopted in the previous one. Section 6.1 describes the data acquisition using TikTok Research API and the building process of the virality time series. The songs analyzed here correspond to a subset of those used in the Spotify experiments, and Section 6.2 contains a correlation analysis on the time series obtained from both platforms. In Section 6.3, we apply the classical epidemic models to the TikTok series, whereas Section 6.4 extends the analysis with the wave-based modeling approach. Section 6.5 repeats the forecasting experiments to evaluate predictive performance in this new context. Finally, Section 6.6 summarizes the main findings and the implications of our results.

6.1 Data Collection and Time Series Modeling

To perform this case study on TikTok, we aim to maintain methodological consistency with the analyses presented in the previous chapter. This alignment allows us to directly compare the performance of epidemiological models when applied to Spotify and TikTok data, ensuring that observed differences are due to platform dynamics rather than methodological variations. Here, we use the TikTok Research API,[4] which provides access to metadata from publicly available videos on the platform. Through this interface, it is possible to retrieve all videos that include a specific song, identified by its corresponding TikTok audio identifier. It is worth noting that, beyond the official audio IDs (typically linked to artists' verified profiles), users may upload other versions of the same song that are later reused in other videos. In this chapter, we restrict our analyses to the official audio IDs only. As a result, videos using non-official versions of a song may not be included in our dataset. Furthermore, there may also be songs with few or no videos associated with their official TikTok ID.

Our initial set of songs consists of the 1,045 tracks for which we successfully fitted epidemiological models using Spotify data (see Section 5.3). However, due to request limits imposed by the API (i.e., 1000 requests and 100,000 records per day), we employ

[4] https://developers.tiktok.com/products/research-api/

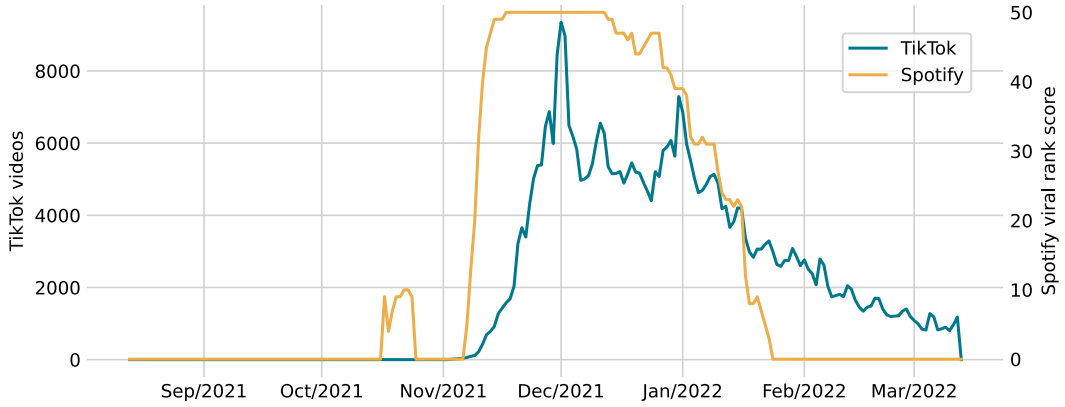


Figure 6.1: Virality time series for the song "abcdefu" by GAYLE. There are distinct y-axis scales for TikTok videos (left) and Spotify chart position (right).

a stratified sampling strategy, selecting 10% of these songs based on their Spotify popularity.[5] This resulted in a subset of 104 tracks. For each of them, we collect the list of short videos published between January 2017 (or the track's release date, if later) and March 2022, in order to match the time span used in the Spotify analyses. In total, our dataset contains 4,581,565 videos across all songs, with a median of 4,364 videos per track, ranging from a minimum of six to a maximum of 773,873.

From the collected video lists, we now build time series representing each song's virality on TikTok. Ideally, the most direct measure of this phenomenon would be the number of views over time, as this reflects the number of users exposed to the song. However, the Research API provides only cumulative view counts (i.e., a single static number) rather than the historical evolution of views. Therefore, we represent music virality on TikTok by using the daily number of new videos created with a given track. This proxy captures the dynamics of how often songs are reused by the community and allows for consistency with the temporal modeling framework adopted in this thesis.

An illustrative example is shown in Figure 6.1, which presents the time series of daily TikTok videos of the song "abcdefu" by GAYLE. After its release in August 2021, it gained traction on the platform starting in November, reaching its peak daily number of videos in early December. Moreover, its virality on Spotify measured by the rank score (see Section 4.2) follows a similar trajectory, peaking in the viral charts slightly earlier. This alignment may suggest a potential correlation between the two platforms, which we further investigate in the next section.

[5]The popularity score is retrieved from the Spotify Web API (https://developer.spotify.com/documentation/web-api) and ranges from 0 to 100. Such a variable updates over time and, in our dataset, reflects the collection date (March 2022).

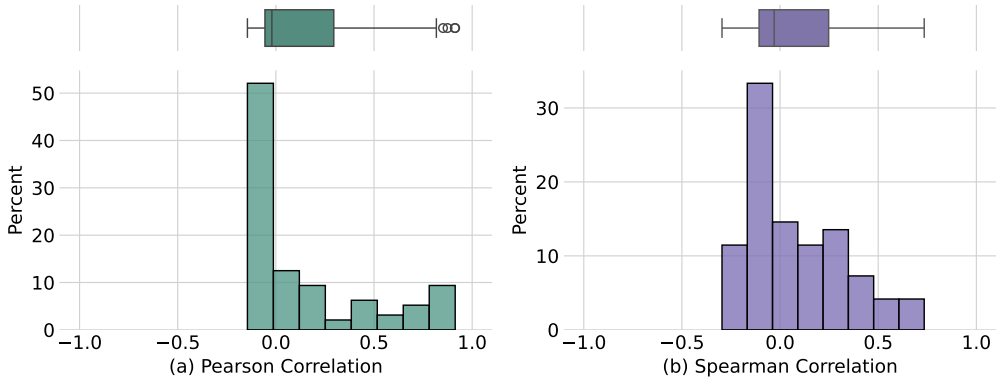


Figure 6.2: Distribution of the (a) Pearson and (b) Spearman correlation coefficients for TikTok number of videos and Spotify virality rank score.

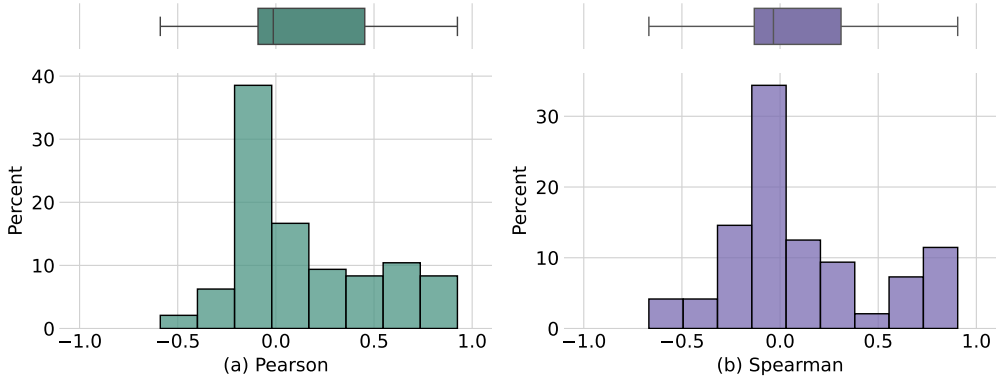


Figure 6.3: Distribution of the (a) Pearson and (b) Spearman correlation coefficients for TikTok number of videos and Spotify success rank score.

6.2 Correlation Analysis with Spotify

In this section, we investigate the relationship between the number of videos on TikTok and song virality on Spotify to assess whether a phenomenon observed on one platform is consistently mirrored on the other. In other words, we test whether a statistical relationship exists between these two dimensions of music popularity. Following a similar methodology to Section 4.3, we calculate correlation coefficients for all songs in our sample using both Pearson's r and Spearman's ρ, which capture linear and monotonic relationships, respectively.

Figure 6.2 presents the distribution of correlation values across all songs regarding music virality. Overall, there is no clear linear or monotonic correlation between the TikTok and Spotify time series, with median values of -0.020 (Pearson) and -0.030 (Spearman). For comparison, we observe a similar pattern when contrasting TikTok time series with Spotify's success (Figure 6.3), with medians of -0.013 and -0.033, respectively. Such results suggest that both platforms capture distinct aspects of music popularity,

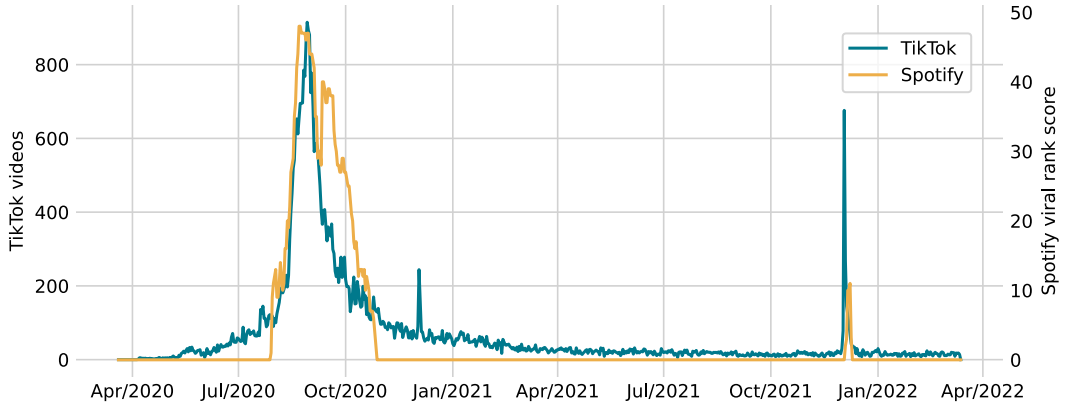


Figure 6.4: Virality time series for the song "Heather" by Conan Gray. There are distinct y-axis scales for TikTok videos (left) and Spotify chart position (right).

potentially reflecting differences in how users interact with music content (e.g., active listening vs. passive exposure or engagement through short videos vs. playlist curation).

Nevertheless, there are songs with strong correlation values between TikTok and Spotify performance. Specifically, 19 songs have significant linear correlations, and seven present significant monotonic correlations (i.e., coefficient greater than 0.5). A notable example is "Heather" by Conan Gray, released in March 2020 (Figure 6.4). Its TikTok and Spotify curves follow very similar trajectories, producing $r = 0.911$ and $\rho = 0.583$. For such songs, it is reasonable to infer that TikTok activity may directly influence Spotify popularity, or vice versa, reflecting a strong relationship between both platforms.

However, the general lack of significant correlations for most songs may indicate that TikTok and Spotify tend to capture distinct dimensions of the music virality phenomenon. Whereas there are overlaps in certain cases, the broader picture suggests complementary rather than redundant roles. Our findings emphasize the value of examining platform-specific virality dynamics, illustrating how applying epidemic models to TikTok time series can complement analyses conducted on Spotify and other platforms with distinct interaction patterns.

6.3 Single Epidemic Modeling

In this section, we model the virality of songs on TikTok using classic epidemic models, treating each trajectory as a single contagion process. We use such compartmental models to capture how songs spread through user engagement, reflected by the daily number of video creations. Section 6.3.1 details the models and their evaluation in this

context, whereas Section 6.3.2 presents and discusses the fitting results.

6.3.1 Model Definition and Setup

We consider the same models from Chapter 5 (i.e., SIR, SEIR, SEIRS, and Open SEIRS) for TikTok time series, in which the variable of interest is the daily number of videos created using a given song. Although this measure does not directly represent the number of individuals exposed to the song, it captures how frequently the community reuses the track, serving as a proxy for its propagation curve in epidemic models.

SIR model. Here, the susceptible state (S) corresponds to TikTok users who have not yet been exposed to or engaged with a song. Infected users (I) are those who are actively disseminating the song by creating videos with it. Finally, recovered users (R) are those who once reused the track but lost interest and stopped contributing to its diffusion.

SEIR model. The exposed state (E) captures users who have been indirectly exposed to the song. For example, when a user has watched a video containing the song but has not yet created their own content. This additional state allows modeling the delay between exposure and active engagement through video creation.

SEIRS model. This extension accounts for the possibility of reinfection, which in TikTok corresponds to users who, after previously losing interest, return to create new videos with the same song. This captures situations where trends re-emerge after a period of inactivity, typically driven by memes, viral challenges, or renewed promotional efforts.

Open SEIRS model. Finally, this model incorporates population dynamics. Here, "births" correspond to new TikTok users who may encounter and reuse a song, whereas "deaths" correspond to users becoming inactive or leaving the platform. Additional "deaths due to infection" can represent users who previously contributed to a song's spread but disengaged permanently, either due to saturation or shifts in trend dynamics.

For fitting the epidemic models to TikTok time series, we adopt the same procedure used in the Spotify analysis (see Sections 5.1 and 5.2). All series are normalized within the $[0, 0.5]$ interval, and we consider the total population as $N = 1.0$, the initial infected I_0 set to the first observed value, and all other compartments initialized to zero. Model parameters are estimated with the *SciPy* library using least squares, starting from initial guesses of 0.5 and nonnegative bounds. Model accuracy is evaluated through Root Mean

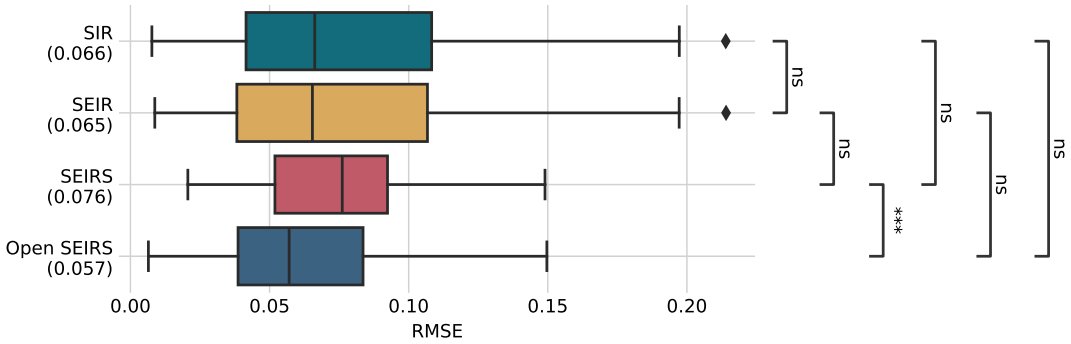


Figure 6.5: RMSE for TikTok virality curves using single epidemic models. Values in parentheses are the median values. Significance is calculated using the Mann-Whitney U test: *** indicates $0.0001 < p \leq 0.001$.

Squared Error (RMSE), which ranges from 0 (perfect fit) to typically 0.5 for poor fits, acknowledging that fitted curves are not strictly bounded within the normalized range.

6.3.2 Results

Similar to the experiments on Spotify time series in the previous chapter, we now fit the four classical epidemic models to verity their suitability for representing music virality on TikTok. From our sample of 104 songs, such models could be successfully applied to 95 (91.3% of the total). The remaining songs were excluded because they had very few or no videos linked to the official audio, with most activity instead tied to user-generated audio IDs (which were discarded from our analysis, as detailed in the previous section).

Figure 6.5 presents the distribution of RMSE values for the four considered models. Overall, the distributions are very close, with median errors ranging from 0.065 to 0.076. According to the Mann–Whitney U test, which assesses statistical differences between two distributions, there are no significant differences across most model comparisons, suggesting that they perform similarly. The only exception is in the comparison between SEIRS and Open SEIRS, where the null hypothesis of equal distributions is rejected with $0.0001 < p \leq 0.001$. Therefore, Open SEIRS produces slightly lower errors and can be considered more suitable for TikTok virality series than the prior.

Such a result contrasts with the findings from the previous chapter with Spotify data, in which the SEIR model is the one with the lowest median error. This divergence highlights the importance of analyzing TikTok independently, as this platform captures specific dimensions of the music virality phenomenon that differ from those observed in streaming environments. For instance, TikTok dynamics are highly volatile and driven by

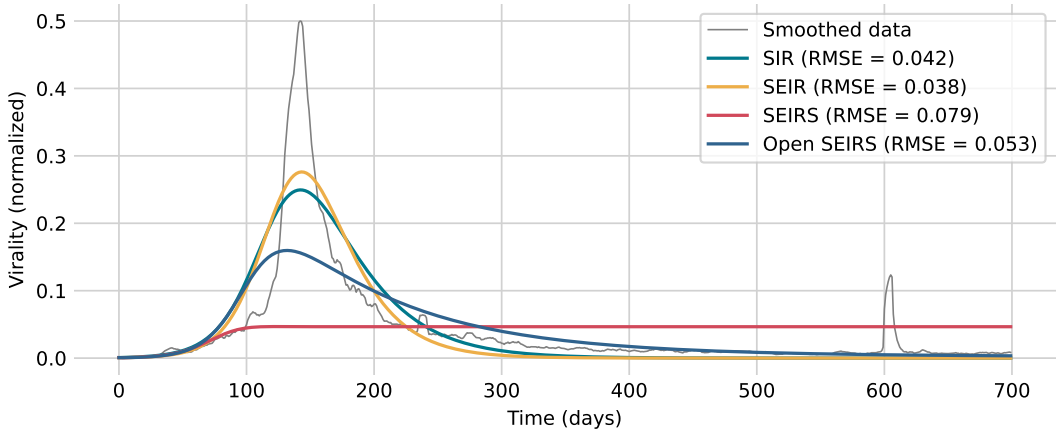


Figure 6.6: Virality time series with its respective model fits for the song "Heather" by Conan Gray.

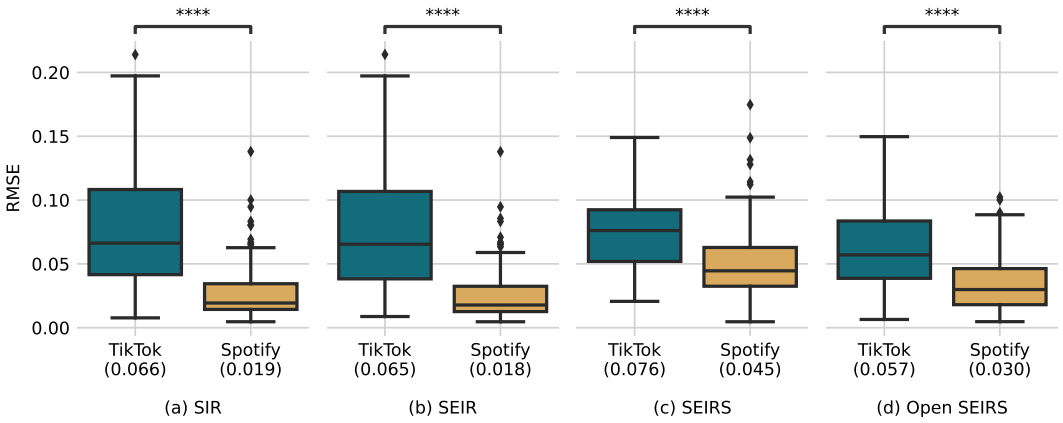


Figure 6.7: RMSE for TikTok and Spotify virality curves. Values in parentheses are the median values. Significance is calculated using the Mann-Whitney U test: **** for $p \leq 0.0001$.

constant user participation in trends, challenges, and meme cycles, making reinfection and open population dynamics captured by Open SEIRS relevant for explaining how songs regain attention or reach new audiences through new content waves. The platform's recommendation algorithm is also important in this context, since it acts as a key driver of feedback loops that amplify exposure and trigger successive waves of engagement.

However, there are also songs for which a single model cannot capture the full dynamics of virality on TikTok. Even the SEIRS and Open SEIRS models, which allow reinfection, fail to represent multiple viral periods, especially when they are spaced apart. The song "Heather" by Conan Gray (Figure 6.6) is a good example of this. After an initial surge of videos about 130 days after its release, the song had another viral moment (although on a smaller scale) around day 600, which is not captured by any of the models.

Cross-platform comparison. Figure 6.7 shows the distribution of RMSE values across models, not only for TikTok data but also in comparison with the results obtained for the viral time series of the same songs on Spotify. For all models considered, the error is consistently higher for TikTok, indicating a poorer overall fit to this platform's data. For instance, the median RMSE for the Open SEIRS model (which performs best on TikTok) is nearly twice as large as the corresponding value for Spotify. Moreover, for the SEIR model, which achieves the best performance on Spotify, the median RMSE on TikTok is more than three times higher than that on Spotify (0.065 versus 0.018). Such differences may reflect distinct virality patterns between the two platforms, reinforcing the correlation analysis discussed in Section 6.2.

6.4 Wave-based Epidemic Modeling

The limitations of single epidemic models in capturing multiple viral resurgences highlight the need for a more flexible framework. In this section, we extend the wave-based epidemic modeling approach originally proposed for Spotify on Chapter 5 to TikTok data. By decomposing a song's virality time series into a sequence of independent epidemic waves, we aim to assess whether this methodology can also capture both primary bursts of virality and later revivals on TikTok, thus taking a first step toward the generalization of this approach across platforms. After detailing the modeling and evaluation in Section 6.4.1, we present and discuss the results in Section 6.4.2.

6.4.1 Modeling and Evaluation

For TikTok, we use the same wave-based epidemic modeling methodology proposed for Spotify in the previous chapter (see Section 5.3 for more details). The approach assumes that songs may present multiple virality periods, which are modeled as independent epidemic waves. In short, the methodology involves three steps: (i) identifying significant peaks in the virality time series using a peak detection algorithm; (ii) segmenting the series around each peak to define the start and end of waves, while ensuring a minimum wave duration and resolving overlaps between adjacent waves; and (iii) fitting an SEIR model independently to each wave. Despite not performing best in the previous experiment, we keep the SEIR model in the wave-based approach since wave segmentation removes the

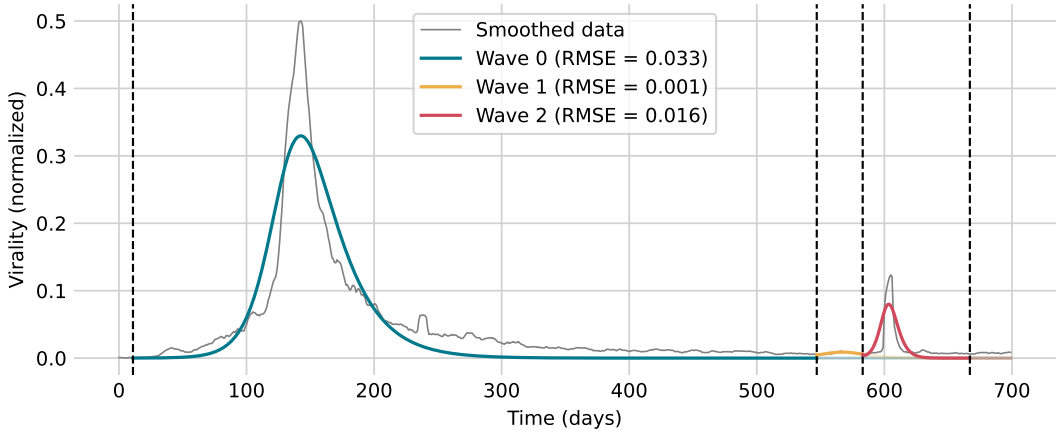


Figure 6.8: Virality time series with the wave-based SEIR fit for the song "Heather" by Conan Gray. The vertical dashed lines delimit the waves.

need to consider reinfection, with each wave modeled independently. Moreover, previous work supports SEIR as particularly effective for capturing the onset of such waves [93].

The fitting and evaluation process also mirrors the one applied to Spotify. We detect peaks with the `find_peaks` function from the *SciPy* library, define the corresponding waves, and fit an SEIR model to each wave separately using the same initialization setup as in Section 5.2.2. Model performance is evaluated by computing the RMSE for each wave, and for songs with multiple waves, we report the average RMSE across all of them.

6.4.2 Results and Discussion

Our wave-based approach produces a median average RMSE of 0.025, a value lower than all of the single epidemic models tested in the previous section. This indicates a better alignment between the fitted curves and the observed TikTok data, and such an improvement shows that the methodology allows for a more realistic representation of virality patterns by capturing multiple engagement periods in the platform. For instance, for the song "Heather" by Conan Gray, our approach identified three distinct waves of virality. Although the existence of the second wave may be considered debatable given its relatively small peak compared to the other two, the viral periods are much more clearly delimited, as illustrated by Figure 6.8.

Regarding wave duration, the average and median values are 100 and 45 days, respectively. Such values are larger than those observed for Spotify, which may reflect either platform-specific dynamics or modeling choices in the time series building process. In other words, TikTok time series capture the number of videos created on the platform

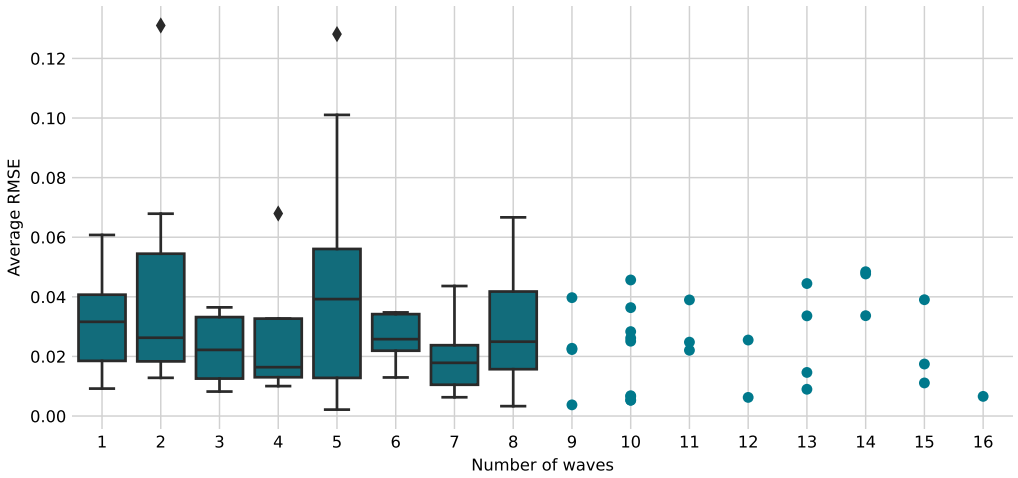


Figure 6.9: Average RMSE distribution grouped by the number of identified waves.

Table 6.1: Descriptive statistics of the SEIR parameters for TikTok time series using our wave-based approach.

	SEIR parameters			
	Min.	**Mean**	**Median**	**Max.**
Average infection rate (β)	0.345	3.617×10^{2}	4.432	3.294×10^{4}
Average recovery rate (γ)	8.772×10^{-3}	0.694	0.372	4.185
Average latency rate (σ)	3.062×10^{-3}	1.452×10^{5}	9.181×10^{3}	6.233×10^{6}
	Derived parameters			
	Min.	**Mean**	**Median**	**Max.**
Average infectious period ($1/\gamma$)	0.638	4.036×10^{2}	12.023	1.804×10^{4}
Average R_0 (β/γ)	1.109	1.372×10^{4}	98.039	1.276×10^{6}

every day, whereas Spotify only provides information about a song's virality once it enters the daily Viral 50 chart. Hence, its virality on days outside the chart remains unknown.

As for the number of waves, TikTok time series have considerably more viral periods than those on Spotify (see Section 5.3.3). The median is six waves per song, with a minimum of one and a maximum of 16. Figure 6.9 shows the distribution of average RMSE values grouped by the number of identified waves. For songs with more than nine waves, we show individual points instead of boxplots due to the small sample size. Finally, after applying the Mann-Whitney U test, no statistically significant differences were found between all of the distributions.

SEIR Parameters. We now move to the analysis of each parameter of the SEIR model, which provides meaningful insights into the mechanisms of music virality on TikTok. Table 6.1 reports descriptive statistics for the estimated parameters (i.e., β, γ, σ) on TikTok, in which each value is the average across all detected waves of a song. Once again, parameters are generally within the interval $[0, 1]$, but there is no strict upper bound, as the values are determined by the underlying shape and scale of the time series.

The infection rate (β) captures how quickly a song spreads through the platform.

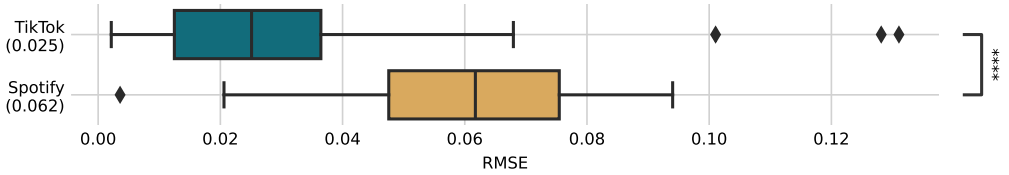


Figure 6.10: RMSE for TikTok and Spotify virality curves using the wave-based approach. Values in parentheses are the median values. Significance is calculated using the Mann-Whitney U test: **** for $p \leq 0.0001$.

On TikTok, the median $\beta = 4.432$ shows that songs can reach new audiences fast, which may be driven by challenges, memes, and content creation. Whereas the distribution includes very large outliers (with maximum values above 10^4), these may reflect explosive moments when songs suddenly go viral after being picked up by influential users or trends. However, higher β values could also result from normalizing the time series to 0.5, which may lead to overestimating the proportion of individuals infected at the peak.

The recovery rate (γ) indicates how fast users lose interest once engaged. With a median of 0.372, this value suggests a considerable time engagement: users do not move instantly on to new content, although they do it faster when compared to Spotify (see Section 5.3.3). Still, the long tail in the distribution shows that certain songs can maintain attention for longer periods, especially when repurposed in multiple trend contexts.

Moreover, the latency rate (σ) measures how quickly users move from exposure to active engagement. The very high median value ($\sigma = 9.181 \times 10^3$) confirms that adoption on TikTok is almost instantaneous: once a user sees or hears a song, they may reuse it almost immediately in their own content. This aligns with the platform's design, where creative participation (remixing, duets, or challenges) accelerates virality far beyond what is typically seen in other music platforms.

Derived parameters. The infectious period ($1/\gamma$) has a median of about 12 days, meaning users remain engaged with a song for nearly two weeks before interest fades. In addition, the basic reproduction number ($R_0 = \beta/\gamma$) shows a median of 98, indicating that a single engaged user can, on average, lead to nearly a hundred new adoptions (due to their followers or the exposure of the videos on the "for you" page). Such a high R_0 underscores the explosive viral potential of TikTok, where songs can gain traction in completely new waves of engagement. However, such results should be interpreted with caution, as the normalization applied to the time series may overestimate peak engagement and, consequently, the parameter values.

Cross-platform comparison. When comparing the results of the wave-based approach between TikTok and Spotify, Figure 6.10 shows that this method produced lower values of RMSE for TikTok curves, indicating that it better captures the virality dynamics specific

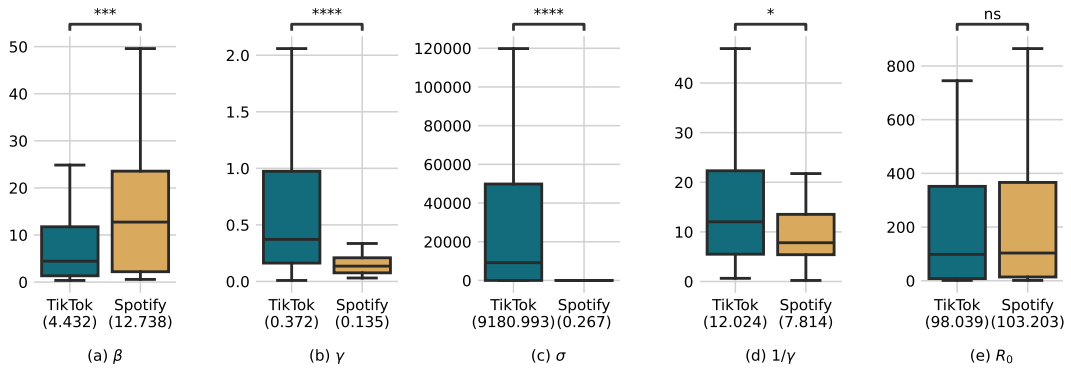


Figure 6.11: Parameter values for TikTok and Spotify virality curves using the wave-based approach. Outliers are ommitted for readability purposes. Values in parentheses are the median values. Significance is calculated using the Mann-Whitney U test: * for $0.01 < p \leq 0.05$; *** for $0.0001 < p \leq 0.001$; **** for $p \leq 0.0001$; and 'ns' otherwise.

to this platform. Furthermore, there are some SEIR parameters with distinct distributions across both platforms, as illustrated in Figure 6.11. For example, viral curves on Spotify have a higher infection rate (β), suggesting that once a song begins to gain visibility, it may spread more efficiently through recommendation mechanisms and playlist-driven exposure. In contrast, the recovery (γ) and latency (σ) rates are significantly higher on TikTok, which may indicate that user engagement cycles are shorter and that content is consumed and replaced at a faster pace.

Regarding the latter parameter, the median σ value is several orders of magnitude higher, suggesting that individuals move almost instantaneously from the exposed to the infected state. In practical terms, this may reflect how quickly users create or interact with new videos once they are exposed to a trend, as well as potential effects of the temporal granularity and other data-related issues. The infectious period ($1/\gamma$) is slightly longer on TikTok (12 days) than on Spotify (7.8 days). Although this difference is statistically weak, its effect size is significant. Finally, for the basic reproduction number (R_0), no statistical difference was observed between platforms, suggesting that the intrinsic potential for a song to go viral is independent of the platform itself.

6.5 Forecasting Song Virality

Similar to the analysis performed for Spotify in the previous chapter, we now evaluate our approach's ability to forecast virality trends on TikTok. Specifically, this section investigates whether a song's future trajectory can be anticipated using only its

initial performance data on the platform. We detail the experimental setup in Section 6.5.1 and discuss the results in Section 6.5.2.

6.5.1 Experimental Setup

Recalling the methodology proposed in Section 5.1, we focus on individual, pre-identified viral waves. For any given checkpoint t within a wave, we fit an SEIR model using only the data observed from the wave's start up to that point. This model is then used to predict the wave's subsequent trajectory. Repeating this process at various checkpoints allows us to assess how early and accurately the model can forecast a song's complete virality pattern based on limited initial data.

To evaluate the SEIR model's forecasting performance, we run two complementary experiments. The first one analyzes how predictive accuracy evolves with more data by fitting the model to progressively larger portions of a wave, i.e., using data up to the 10%, 25%, 50%, 75%, 90%, and peak virality points. In addition, the second experiment compares our model against the same three traditional forecasting methods: **ARIMA**, **Support Vector Regression (SVR)**, and **Prophet**. To ensure a fair comparison across all models, we measure the performance using the Root Mean Squared Error (RMSE) calculated only on the unseen data after the respective cut-off point for each model.

6.5.2 Results

Figure 6.12 illustrates the results of the first experiment, which evaluates the impact of the forecast horizon. It shows the evolution of the Root Mean Squared Error (RMSE) as different fractions of the time series are used as the model input. Similar to the analysis performed for Spotify, the forecast error decreases as more data becomes available. In particular, when using only 50% of the initial wave data, the SEIR model achieves a median RMSE of 0.027, which is very close to the 0.025 obtained when fitting the model with the full viral curve. This finding reinforces that wave-based modeling is robust and effective even with partial data from TikTok, indicating the possibility of predicting a song's viral behavior considerably in advance once a wave has begun.

Regarding the second experiment, Figure 6.13 compares the performance of our

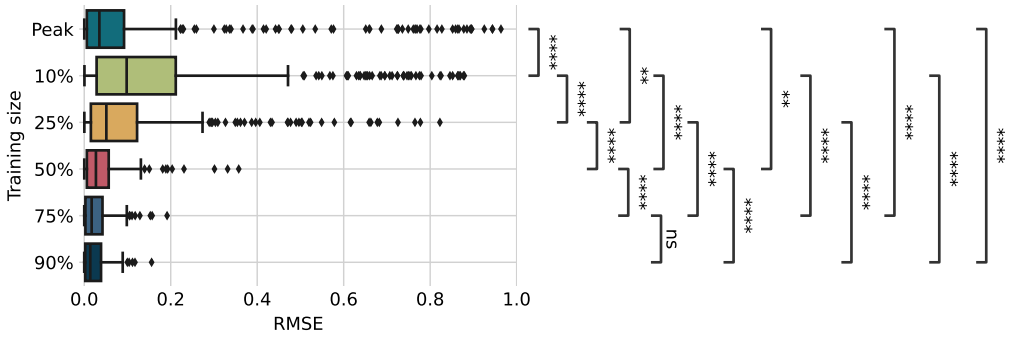


Figure 6.12: Forecast RMSE for TikTok with different partial data sizes. Significance is calculated using the Mann-Whitney U test: ** for $0.001 < p \leq 0.01$; **** for $p \leq 0.0001$; and 'ns' for non-significant.

SEIR-based method with the three baseline models for time series forecasting.[6] In all of the analyzed scenarios, Prophet is the one with the highest median error. In particular, in the scenario using data up to the wave's peak, it fails significantly, generating RMSE values above 1. In contrast, the error distributions for the three remaining models (i.e., SEIR, ARIMA, and SVR) were quite close, often with no statistically significant differences. Therefore, SEIR consistently ranks among the best-performing approaches across all scenarios, with particular emphasis on its performance at the 50% and 90% scenarios.

Cross-platform comparison. Figure 6.14 expands our analyses by directly comparing the SEIR forecasting performance across platforms. Specifically, we only use the Spotify time series for the songs present in our TikTok sample. The results reveal that the proposed wave-based approach consistently outperforms the models fitted on Spotify data for all forecasting horizons, suggesting that the dynamics captured in TikTok time series are more predictable once the wave structure is partially known. This may be attributed to the distinctive viral mechanisms of TikTok, where engagement patterns follow clearer virality bursts compared to those typically observed in streaming environments.

Overall, the experimental results demonstrate that epidemiological models represent a promising alternative for forecasting the viral behavior of songs on TikTok after a wave has started. The SEIR model produces comparable, and in certain scenarios, superior performance to traditional time-series models such as ARIMA. Furthermore, our approach has the fundamental advantage of employing interpretable parameters, such as the transmission rate and the infection period, which not only enable forecasting but also reveal intrinsic and important characteristics of the viral phenomenon's dynamics.

[6]To avoid distortions from extreme errors, we set the y-axis upper limit to 1 in Figures 6.13(a–c). This adjustment is necessary because Prophet sometimes fails to capture the underlying dynamics, resulting in unusually large forecasting errors.

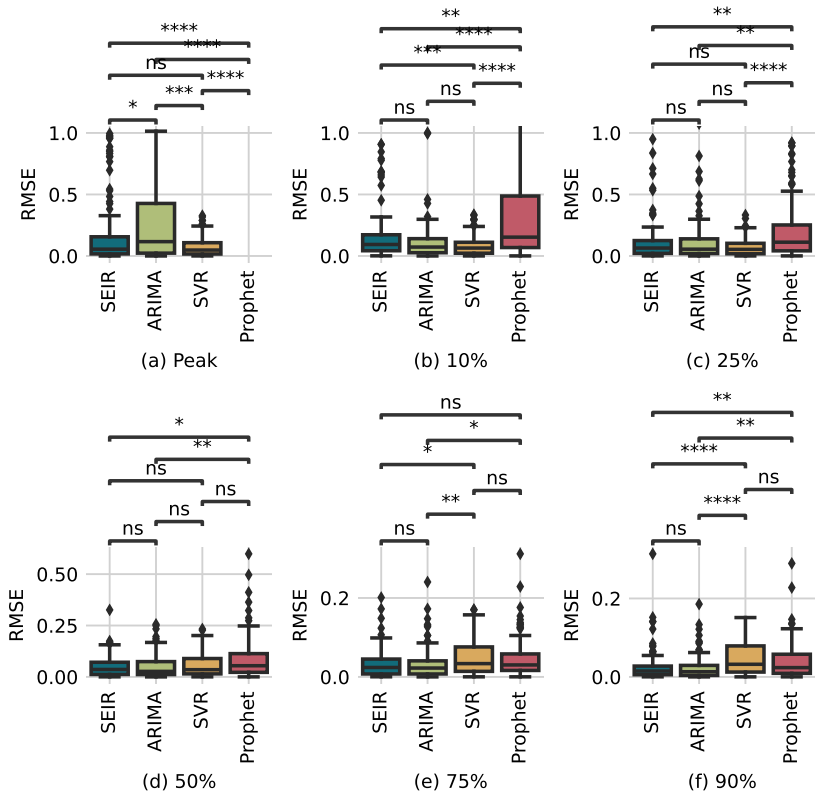


Figure 6.13: Forecasting performance for our approach on TikTok data with baselines. Significance is calculated using the Mann-Whitney U test: * for $0.01 < p \leq 0.05$; ** for $0.001 < p \leq 0.01$; *** for $0.0001 < p \leq 0.001$; **** for $p \leq 0.0001$; and 'ns' otherwise. Note the difference in the y-axis scale.

6.6 Overall Considerations

In this chapter, we go deeper on the investigation regarding modeling music virality on social platforms as a contagion process. Following the methodology with Spotify data from the previous chapter, we perform a case study applying the same methodology to data from TikTok, one of the main platforms driving the phenomenon of online virality. Using data collected via the platform's Research API, we build time series for a sample of the songs used in the Spotify analysis and replicate the same analyses to verify whether this methodology is also effective for other platforms.

The results demonstrate that the virality dynamics observed on TikTok are different from those on Spotify. Therefore, both platforms should be analyzed separately as they highlight distinct dimensions of the viral phenomenon. Nevertheless, the application of epidemiological models (both the simple and the wave-based approaches) produces good results that are close to the real data, offering interpretable parameters that provide valuable insights into the diffusion of these songs. Finally, the forecasting experiments re-

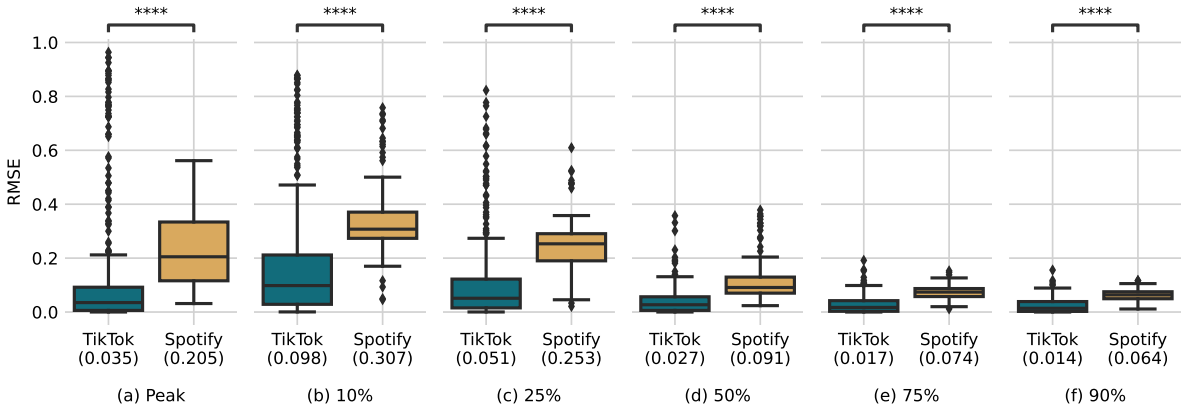


Figure 6.14: Forecast RMSE for TikTok and Spotify for different partial data sizes. Significance is calculated using the Mann-Whitney U test: **** for $p \leq 0.0001$.

veal that our model's performance is comparable to that of traditional ones, allowing the analysis of a song's viral behavior once its virality wave has begun. Overall, the success of this case study represents an important step toward generalizing the modeling of the viral phenomenon as a social contagion process.

Limitations. Our analyses have some limitations. The wave-based approach is sensitive to noise in the time series, which may still produce spurious waves despite smoothing. The song sample is relatively small and should be expanded with more recent data. Finally, daily video counts may not perfectly capture TikTok virality, which metrics like views or listening history (currently unavailable via the API) could represent more accurately.

Chapter 7

Concluding Remarks

The rise of streaming and social media has not only transformed how music is consumed but has also generated unprecedented data, enabling a deeper analysis of the mechanisms behind music popularity. Specifically, it allows for a granular examination of two distinct facets of music popularity: the rapid and ephemeral virality and sustained mainstream success. While often used interchangeably, the dynamics related to these two phenomena and their intrinsic relationship have remained largely unquantified. Analyzing the music ecosystem from the dual perspectives of virality and success is therefore highly relevant to understanding its contemporary dynamics.

In this thesis, we addressed this gap with the main goal of understanding the phenomenon of music viralization on social platforms and its relationship with mainstream success. To do so, we quantitatively characterized viral and hit songs, modeled their temporal and causal relationships, and investigated the contagion processes underlying their diffusion. Our results revealed that virality is primarily driven by artist- and time-related features rather than musical content itself, that viral exposure can in some cases anticipate later mainstream success (and vice versa), and that epidemiological models effectively capture the short-term propagation of songs on social platforms. Together, all such findings offer a comprehensive and data-driven understanding of how virality emerges, interacts with, and sometimes predicts success. Next, we present this thesis' main contributions, organized according to the three research goals (RGs) that guided it.

RG1. Quantitative Characterization of Viral and Hit Songs. Based on chart data, we perform a quantitative and data-driven comparison of hit and viral songs to identify similarities and differences in their characteristics. After an exploratory characterization, we distinguish viral from hit songs through a classification task that evaluates both intrinsic and extrinsic features. We conclude that relying solely on acoustic or other intrinsic features is not enough for this task. Extrinsic features such as artist-related and temporal features are the most influential, increasing the likelihood of a song becoming viral. Overall, our results highlight that hit and viral songs represent two distinct yet interconnected dimensions of music popularity.

RG2. Causal Relationship Between Virality and Success. We model the temporal evolution of musical virality and success using time series that represent song performance on streaming platform charts. We then perform three distinct analyses for evaluating the temporal relationship between these two phenomena: *(i)* correlation analysis, *(ii)* Granger causality, and *(iii)* causal discovery. Our results indicate that virality and success do not always show synchrony, and while they can sometimes forecast each other, this behavior is not present across all songs. Causal discovery analysis confirms that for some songs, in which virality may lead to success (and vice versa). Overall, our findings reflect the importance of exploring not only the individual aspects of virality and success but also their dynamic interactions.

RG3. Music Virality as a Contagion Process. From the time series, we explore the use of epidemic models to represent, explain, and forecast music popularity. We introduce a wave-based approach that captures multiple independent bursts of popularity, which is not possible using classic epidemic models. Using streaming data from Spotify and TikTok, we evaluate our approach's ability to fit and forecast virality over time, comparing its performance with traditional time-series forecasting methods. Our results show that this approach successfully models the contagion process on both platforms, offering interpretable parameters that provide meaningful insights about the viral phenomenon. Moreover, it offers forecast accuracy comparable to conventional time series algorithms, with the additional benefit of providing interpretable parameters that shed light on the underlying diffusion processes.

Overall, this thesis represents an important framework for understanding music popularity in the digital era, offering key insights into the distinct phenomena of virality and success. We contribute to the fields of Hit Song Science and social computing by establishing a data-driven distinction between viral and hit songs, uncovering their complex causal relationship, and validating the use of epidemiological models to forecast online contagion. Such findings benefit not only researchers but also the broader music industry. In particular, they can serve as a basis for enabling the early detection of viral trends to inform A&R and marketing strategies, ultimately helping artists and labels develop more effective promotional approaches in the digital music market. Therefore, all such insights contribute to a deeper understanding of how music consumption is shaped in the streaming age, in which the rapid dissemination of content through social networks is increasingly central to the music industry's dynamics.

Methodological Note on Data Validity. A challenge inherent to studies leveraging streaming and social media platforms lies in the proprietary nature of their data. Specifically, the Spotify charts, used here as primary metric for mainstream success and virality, operate as a *black box*. Although the success charts are defined as the ranking of the

most-streamed songs, we acknowledge that the lack of transparency regarding the exact parameters of the viral ranking algorithm constitutes a limitation. However, we methodologically defend the use of such charts as they represent the best publicly available proxy for large-scale consumption and engagement volume. Therefore, they serve as a robust indicator of popularity in the digital era.

7.1 Research Products

This section lists the products of the thesis research. We first present the direct products derived from each chapter (Section 7.1.1). Then, we list the main byproducts that are related to the whole research, but have not entered the thesis (Section 7.1.2).

7.1.1 Direct Products

This section presents the direct products resulting from this research, including publications and other materials that emerged directly from the development of the thesis.

Chapter 2. A Literature Review on Music Virality.

1. Oliveira, G. P.; Couto da Silva, A. P.; Moro, M. M. Music virality on social platforms: A literature review. Vórtex Music Journal, v. 13, p. 1-32, 2025. DOI. [76].

Chapter 3. A Quantitative Characterization of Viral and Hit Songs.

2. Oliveira, G. P.; Couto da Silva, A. P.; Moro, M. M. A quantitative comparison of viral and hit songs in the Brazilian music market. Vórtex Music Journal, v. 12, p. 1-29, 2024. DOI. [72].

3. Oliveira, G. P.; Couto da Silva, A. P.; Moro, M. M. What makes a viral song? Unraveling music virality factors. In: Proceedings of the 16th ACM Web Science Conference, p. 181-190, 2024. DOI. [73].

Chapter 4. On the Causal Relationship Between Music Virality and Success.

4. Oliveira, G. P.; Couto da Silva, A. P.; Moro, M. M. Analyzing the temporal relation between virality and success in the Brazilian music market. In: Proceedings of the 51st Integrated Software and Hardware Seminar, p. 157-168, 2024. DOI. [71]. **Best Paper Award.**

5. Oliveira, G. P.; Couto da Silva, A. P.; Moro, M. M. On the Causal Relationship Between Music Virality and Success. IEEE Access, v. 14, p. 122782-122791, 2025. DOI. [75].

Chapter 5. Music Virality as a Contagion Process.

6. Oliveira, G. P.; Vassio, L.; Couto da Silva, A. P.; Moro, M. M. Modeling music popularity as an epidemic: insights from the Brazilian market. In: Proceedings of the 14th Brazilian Workshop on Social Network Analysis and Mining, p. 79-92, 2025. DOI. [78].

7. Oliveira, G. P.; Vassio, L.; Couto da Silva, A. P.; Moro, M. M. Contagious rhythms: A wave-based epidemic approach for music virality on social platforms. In: 17th International Conference on Advances in Social Network Analysis and Mining, 2025. DOI. [77].

Besides the publications, an important outcome of this chapter is the **novel wave-based approach** developed to model music virality, which was further applied to TikTok data (Chapter 6). This model, however, has the potential to be extended beyond Spotify and TikTok to other platforms, and may also be explored as a framework for modeling broader social phenomena, such as meme and behavior dissemination.

7.1.2 Byproducts

Besides the publications directly related to the thesis, the knowledge acquired from this research has also contributed to other co-advisorships, publications, and datasets in social and music computing. Items are listed from the most recent to the oldest.

Undergraduate Final Project Co-advisorships.

1. Bernardo Roberto Andrade Silva (UFMG 2025). *Análise de dados de circulação viral e epidemiológicos de municípios que trabalham com o MI-Aedes.*

2. Gabriela Assunção Fonseca (UFMG 2024). *Explorando a Influência Externa na Popularidade Musical: Google Trends como Proxy para Viralidade e Sucesso no Spotify.*

    - **Paper:** Fonseca, G. A.; Oliveira, G. P.; Couto da Silva, A. P. Influência Externa na Popularidade Musical: Google Trends como Indicador de Viralidade e Sucesso no Spotify (*External Influence on Music Popularity: Google Trends as an Indicator of Virality and Success on Spotify*). In: Procs. 14th Brazilian Workshop on Social Network Analysis and Mining, p. 93-105, 2025. DOI. [25].

3. Jorge Henrique F. da Silva (UFMG 2024). *Detecção de tópicos frequentes em músicas virais no Spotify.*

Dataset. MGD+: an enhanced dataset on musical success over time in global and regional markets with enhanced artist and genre collaboration information. DOI.

- **Paper:** Seufitelli, D. B.; Oliveira, G. P.; Silva, M. O.; Moro, M. M. MGD+: An Enhanced Music Genre Dataset with Success-based Networks. In: Proceedings of the 5th Dataset Showcase Workshop, 2023. p. 36-47. DOI. [97].

Further Publications. *On music and gender*:

1. Silva, M. O.; Oliveira, G. P.; Moro, M. M. Data Insights on Gender Representation: Analyzing the Book and Music Industries. In: Companion Proceedings of the 39th Brazilian Symposium on Databases, p. 338-347, 2024. DOI. [107].

2. Silva, M. O.; Oliveira, G. P.; Moro, M. M. Premiação das mulheres na literatura e na música: análises de dados da Billboard e do Goodreads (*Women in Literature and Music Awards: Analyzing Billboard and Goodreads Data*). In: A Internet como campo de disputas de gênero, p. 185-197. 2024. DOI. [108].

On other aspects of musical success:

3. Oliveira, G. P.; Moro, M. M. Frequent Genre Mining on Hit and Viral Songs. Journal of Information and Data Management, v. 16, p. 136-145, 2025. DOI. [62].

4. Silva, M. O.; Oliveira, G. P.; Seufitelli, D. B.; Moro, M. M. Temporal Success Analyses in Music Collaboration Networks: Brazilian and Global Scenarios. Vórtex Music Journal, v. 11, p. 1-27, 2023. DOI. [106].

5. Silva, M. O.; Oliveira, G. P.; Seufitelli, D. B.; Moro, M. M. Collaboration-Aware Hit Song Prediction. J. on Interactive Systems, v. 14, p. 201-214, 2023. DOI. [105].

6. Silva, M. O.; Oliveira, G. P.; Seufitelli, D. B.; Moro, M. M. Collaboration as a Driving Factor for Hit Song Classification. In: Procs. 28th Brazilian Symposium on Web and Multimedia, p. 66-74, 2022. DOI. [103]. **Best Paper Runner-up.**

7. Paula, B. C. M.; Oliveira, G. P.; Moro, M. M. Mood Analysis during the COVID-19 Pandemic in Brazil through Music. In: Companion Proceedings of the 28th Brazilian Symposium on Web and Multimedia, p. 53-56, 2022. DOI. [80].

8. Melo-Gomes, L.; Seufitelli, D. B.; Oliveira, G. P.; Silva, Mariana O.; Moro, M. M. Análise do Sucesso Musical no Brasil Utilizando Dados do Twitter (*Analysis of Musical Success in Brazil Using Twitter Data*). In: Companion Proceedings of the 37th Brazilian Symposium on Databases, p. 40-46. 2022 DOI. [58].

9. Seufitelli, D. B.; Oliveira, G. P.; Silva, M. O.; Barbosa, G. R. G.; Melo, B. C.; Botelho, J. E.; Moro, M. M. From Compact Discs to Streaming: A Comparison of Eras within the Brazilian Market. Vórtex Music Journal, v. 10, p. 1-28, 2022. DOI. [96].

On social aspects in other domains:

10. Oliveira, G. P.; Moura, A. F. C.; Batista, N. A.; Brandão, M. A.; Hora, A.; Moro, M. M. How do developers collaborate? Investigating GitHub heterogeneous networks. Software Quality Journal, v. 31, p. 211-241, 2023. DOI. [69].

11. Silva, M. O.; Oliveira, G. P.; Moro, M. M. Analyzing Character Networks in Portuguese-language Literary Works. In: In Proceedings of the 12th Brazilian Workshop on Social Network Analysis and Mining, p. 115-126, 2022. DOI. [104].

12. Oliveira, G. P.; Paiva, B. F.; Couto da Silva, A. P.; Moro, M. M. Characterizing the Diffusion of Misinformation Regarding the CoronaVac Vaccine in Brazil. In: Procs. Brazilian Workshop on Social Network Analysis and Mining, 2022. DOI. [64].

In addition to developing research on music and other collaborative domains, the knowledge acquired during the PhD was also applied to research in the area of digital government, contributing to scientific production in this field [15, 65–68, 74].

7.2 Future Work

Based on our research on music popularity, we now identify and discuss potential research directions on this subject. More than open research problems, such topics are

relevant for understanding the factors behind music virality on social platforms and can also provide meaningful insights about the viral dynamics of other types of content.

Complex contagion modeling. The modeling of music popularity diffusion can be extended beyond the compartmental models adopted in this work, which consider only pairwise relationships between individuals and assume population homogeneity when representing the infection mechanism [33]. For example, there are cases in which it is necessary to consider additional dimensions that may impact transmission, resulting in *multilayer models* that account for multiple network dimensions [5]. In addition, more recent studies investigating the transmission of behaviors use *higher-order models*, such as hypergraphs,[1] to model processes in which contagion requires contact with two or more sources of activation [35]. Therefore, research on music popularity could benefit from such modeling to explain aspects beyond those captured in the current analyses, e.g., the combined influence of advertising, recommendation systems, and cross-platform interactions.

Recommendation algorithms. Most of the current social networks are based on a personalized content feed so that people can see content that is more suited to their tastes. Therefore, recommendation algorithms play a central role in the viralization of online content. Studies such as Ivanov et al. [36] focus on content recommendation itself, but little is known about the impact of recommendation algorithms on popular platforms such as TikTok and Spotify, as well as how recommendation drives the subsequent success of such songs. Therefore, diving into this particular focus could provide more insights into the connection between song virality and commercial success.

Community analysis. Still, regarding content dissemination, the existence of personalized feeds makes specific content go viral in certain niches or user communities. For example, content related to reading and books in general resonates a lot in the community that became known as BookTok.[2] Regarding music, future work can compare the mass viralization of songs with the same process in more segmented groups, for example analyzing geographic and cultural borders.

Platform influence analysis. Although the phenomenon of music going viral occurs regardless of platform, the unique characteristics of each platform (e.g., public, recommendation algorithm, interaction dynamics, etc.) may influence how and why a given piece of content goes viral. For example, videos on YouTube may go viral differently than songs on TikTok or Spotify. Therefore, studying such dynamics comparatively could help

[1] In graph theory, a hypergraph is a generalization of a standard graph, where an edge can connect any number of vertices.

[2] https://www.bbc.com/news/uk-england-67555175

to better understand how different media formats and social interactions impact music dissemination online.

Ethical aspects. Recommendation algorithms play a key role in content virality within social platforms, especially on those in which the user's feed is not only composed of content posted by people they follow but also other relevant recommended content (e.g., "for you" pages). Therefore, it is reasonable to say that music virality can be artificially boosted by bots (i.e., automated accounts) and invasive marketing strategies, which raises questions about authenticity and transparency. Hence, exploring such ethical challenges would contribute to understanding how to balance virality with fair and responsible practices in the music industry.

References

[1] Fabian Abel, Ernesto Diaz-Aviles, Nicola Henze, Daniel Krause, and Patrick Siehndel. Analyzing the blogosphere for predicting the success of music and movie products. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 276–280. IEEE Computer Society, 2010. doi: 10.1109/ASONAM.2010.50. URL https://doi.org/10.1109/ASONAM.2010.50.

[2] Thayer Alshaabi, David Rushing Dewhurst, Joshua R. Minot, Michael V. Arnold, Jane Lydia Adams, Christopher M. Danforth, and Peter Sheridan Dodds. The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009-2020. *EPJ Data Sci.*, 10(1): 15, 2021. doi: 10.1140/epjds/s13688-021-00271-0. URL https://doi.org/10.1140/epjds/s13688-021-00271-0.

[3] Carlos Soares Araujo, Marco Cristo, and Rafael Giusti. Predicting music popularity on streaming platforms. In *Proceedings of the 17th Brazilian Symposium on Computer Music*, pages 141–148. SBC, 2019. doi: 10.5753/sbcm.2019.10436. URL https://doi.org/10.5753/sbcm.2019.10436.

[4] Carlos V. S. Araujo, Rayol M. Neto, Fabíola G. Nakamura, and Eduardo F. Nakamura. Predicting music success based on users' comments on online social networks. In *Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web*, pages 149–156. ACM, 2017. doi: 10.1145/3126858.3126885. URL https://doi.org/10.1145/3126858.3126885.

[5] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009. doi: 10.1073/pnas.0906910106. URL https://doi.org/10.1073/pnas.0906910106.

[6] Nicola Barbieri and Francesco Bonchi. Influence maximization with viral product design. In *SIAM International Conference on Data Mining 2014, SDM 2014*, volume 1, page 55 – 63, 2014. doi: 10.1137/1.9781611973440.7. URL https://doi.org/10.1137/1.9781611973440.7.

[7] Pizzolittos Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. USA*, 113(27):7345–7352, 2016. doi: 10.1073/PNAS.1510507113. URL https://doi.org/10.1073/pnas.1510507113.

[8] Hemilly Bastos, Débora Moreira Giunti, Larissa Benvindo, Alexandre Nascimento, and Luana Inocêncio. Trends no TikTok e sua influência no streaming musical: os casos Doja Cat e Olivia Rodrigo. In *Anais do 44º Congresso Brasileiro de Ciências da Comunicação*, 2021.

[9] Miguel Baños-Gonzalez, Hector Canorea Tiralaso, and Mario Rajas Fernandez. The broadcast of the music video on YouTube. Analysis of the viral capacity of the video clip. *Revista Latina de Comunicación Social*, 1(77):117–141, 2020. ISSN 1138-5820. doi: 10.4185/RLCS-2020-1452. URL https://doi.org/10.4185/RLCS-2020-1452.

[10] Marco Biasioli. "It wasn't our song anymore": Molchat Doma, the death of the reader and the birth of the TikToker. *IASPM Journal*, 14(1):151 – 171, 2024. doi: 10.5429/2079-3871(2024)v14i1.10en. URL https://doi.org/10.5429/2079-3871(2024)v14i1.10en.

[11] Kerstin Bischoff, Claudiu S. Firan, Mihai Georgescu, Wolfgang Nejdl, and Raluca Paiu. Social knowledge-driven music hit prediction. In *Proceedings of the 5th International Conference Advanced Data Mining and Applications*, pages 43–54. Springer, 2009. doi: 10.1007/978-3-642-03348-3_8. URL https://doi.org/10.1007/978-3-642-03348-3_8.

[12] Ottar N Bjørnstad et al. Modeling infectious epidemics. *Nature methods*, 17(5):455–457, 2020. doi: 10.1038/s41592-020-0822-z. URL https://doi.org/10.1038/s41592-020-0822-z.

[13] Ottar N Bjørnstad et al. The seirs model for infectious disease dynamics. *Nature methods*, 17(6):557–559, 2020. doi: 10.1038/s41592-020-0856-2. URL https://doi.org/10.1038/s41592-020-0856-2.

[14] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS]*, pages 601–608. MIT Press, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/296472c9542ad4d4788d543508116cbc-Abstract.html.

[15] Michele A. Brandão, Arthur P. G. Reis, Bárbara M. A. Mendes, Clara A. B. Almeida, Gabriel P. Oliveira, Henrique Hott, Larissa D. Gomide, Lucas L. Costa, Mariana O. Silva, Anisio Lacerda, and Gisele L. Pappa. Plus: A semi-automated

pipeline for fraud detection in public bids. *Digital Government: Research and Practice*, 5(1):1–16, 2024. doi: 10.1145/3616396. URL https://doi.org/10.1145/3616396.

[16] Aniello Castiglione, Giovanni Cozzolino, Francesco Moscato, and Vincenzo Moscato. Cognitive analysis in social networks for viral marketing. *IEEE Trans. Ind. Informatics*, 17(9):6162–6169, 2021. doi: 10.1109/TII.2020.3026013. URL https://doi.org/10.1109/TII.2020.3026013.

[17] Meeyoung Cha, Juan Antonio Navarro Pérez, and Hamed Haddadi. The spread of media content through blogs. *Social Network Analysis and Mining*, 2(3):249 – 264, 2012. doi: 10.1007/s13278-011-0040-x. URL https://doi.org/10.1007/s13278-011-0040-x.

[18] Jen Shiau Chou, Masanao Ochi, Takeshi Sakaki, Ken Nagahama, Kanji Sakai, Junichiro Mori, and Ichiro Sakata. Constructive approach for early extraction of viral spreading social issues from twitter. In *Proceedings of the 12th ACM Conference on Web Science*, pages 96–105. ACM, 2020. doi: 10.1145/3394231.3397899. URL https://doi.org/10.1145/3394231.3397899.

[19] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013. doi: 10.4324/9780203771587. URL https://doi.org/10.4324/9780203771587.

[20] Andi Coulter. Marketing Agile Artists: How Music Labels Can Leverage TikTok's Virality. *Journal of the Music and Entertainment Industry Educators Association*, 22(1):135–161, 2022. ISSN 15597334. doi: 10.25101/22.5. URL https://doi.org/10.25101/22.5.

[21] Ruth Dhanaraj and Beth Logan. Automatic prediction of hit songs. In *ISMIR*, pages 488–491, London, UK, 2005. ISMIR. URL http://ismir2005.ismir.net/proceedings/2024.pdf.

[22] Maura Edmond. Here we go again: Music videos after youtube. *Television and New Media*, 15(4):305 – 320, 2014. doi: 10.1177/1527476412465901. URL https://doi.org/10.1177/1527476412465901.

[23] Femi Eromosele. What Happens When a Music Video Goes Viral? Gastrocomedy and Prosumer Recreations of Timaya's I Can't Kill Myself. *Journal of African Cultural Studies*, 33(4):424 – 440, 2021. doi: 10.1080/13696815.2020.1756756. URL https://doi.org/10.1080/13696815.2020.1756756.

[24] Lauren K. Fink et al. Viral tunes: changes in musical behaviours and interest in coronamusic predict socio-emotional coping during COVID-19 lockdown. *Humanities*

and Social Sciences Communications, 8(1), 2021. doi: 10.1057/s41599-021-00858-y. URL https://doi.org/10.1057/s41599-021-00858-y.

[25] Gabriela A. Fonseca, Gabriel P. Oliveira, and Ana Paula Couto da Silva. Influência externa na popularidade musical: Google trends como indicador de viralidade e sucesso no spotify. In *Proceedings of the 14th Brazilian Workshop on Social Network Analysis and Mining*, pages 93–105, Porto Alegre, RS, Brasil, 2025. SBC. doi: 10.5753/brasnam.2025.8784. URL https://doi.org/10.5753/brasnam.2025.8784.

[26] Joana Freitas. 'Make Classical Music Great Again': Contemporary Music, Masculinity, and Virality in Memetic Media in Online Spaces. *Contemporary Music Review*, 41(4):429 – 444, 2022. doi: 10.1080/07494467.2022.2087392. URL https://doi.org/10.1080/07494467.2022.2087392.

[27] Wayne A Fuller. *Introduction to statistical time series*. John Wiley & Sons, 2009.

[28] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, USA, 2019. ISBN 978-1-491-96229-9.

[29] Carlos Gomes, Daniel Schneider, Katia Moraes, and Jano Moreira de Souza. Crowdsourcing for music: Survey and taxonomy. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, SMC*, pages 832–839. IEEE, 2012. doi: 10.1109/ICSMC.2012.6377831. URL https://doi.org/10.1109/ICSMC.2012.6377831.

[30] Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969. doi: 10.2307/1912791. URL https://doi.org/10.2307/1912791.

[31] Marco Guerini, Carlo Strapparava, and Gözde Özbal. Exploring text virality in social networks. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 506–509. The AAAI Press, 2011. doi: 10.1609/icwsm.v5i1.14169. URL https://doi.org/10.1609/icwsm.v5i1.14169.

[32] Marco Guerini, Alberto Pepe, and Bruno Lepri. Do linguistic style and readability of scientific abstracts affect their virality? In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2012. doi: 10.1609/icwsm.v6i1.14305. URL https://doi.org/10.1609/icwsm.v6i1.14305.

[33] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4): 599–653, 2000. doi: 10.1137/S0036144500371907. URL https://doi.org/10.1137/S0036144500371907.

[34] Akram Sadat Hosseini and Steffen Staab. Emotional framing in the spreading of false and true claims. In *Proceedings of the 15th ACM Web Science Conference*, pages 96–106, Austin, TX, USA, 2023. ACM. doi: 10.1145/3578503.3583611. URL https://doi.org/10.1145/3578503.3583611.

[35] Iacopo Iacopini, Giovanni Petri, Alain Barrat, and Vito Latora. Simplicial models of social contagion. *Nature communications*, 10(2485), 2019. doi: 10.1038/s41467-019-10431-6. URL https://doi.org/10.1038/s41467-019-10431-6.

[36] Sergei Ivanov, Konstantinos Theocharidis, Manolis Terrovitis, and Panagiotis Karras. Content recommendation for viral social influence. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 565–574. ACM, 2017. doi: 10.1145/3077136.3080788. URL https://doi.org/10.1145/3077136.3080788.

[37] Lu Jiang, Yajie Miao, Yi Yang, Zhen-Zhong Lan, and Alexander G. Hauptmann. Viral video style: A closer look at viral videos on youtube. In *International Conference on Multimedia Retrieval*, page 193. ACM, 2014. doi: 10.1145/2578726.2578754. URL https://doi.org/10.1145/2578726.2578754.

[38] Christian Kahl. Create attention to attract attention - viral marketing of digital music in social networks. In *18th Americas Conference on Information Systems*. Association for Information Systems, 2012. URL http://aisel.aisnet.org/amcis2012/proceedings/SocialIssues/9.

[39] Christian Kahl and Andreas Albers. How to unleash the virus - social networks as a host for viral music marketing. In *IEEE 15th Conference on Business Informatics*, pages 47–54. IEEE Computer Society, 2013. doi: 10.1109/CBI.2013.16. URL https://doi.org/10.1109/CBI.2013.16.

[40] Jigisha Kamal, Pankhuri Priya, M R Anala, and G R Smitha. A classification based approach to the prediction of song popularity. In *ICSES*, pages 1–5, Chennai, India, 2021. IEEE Computer Society. doi: 10.1109/ICSES52305.2021.9633884. URL https://doi.org/10.1109/ICSES52305.2021.9633884.

[41] Amir Khatibi, Fabiano Belém, Ana Paula Couto da Silva, Jussara M. Almeida, and Marcos André Gonçalves. Fine-grained tourism prediction: Impact of social and environmental features. *Inf. Process. Manag.*, 57(2):102057, 2020. doi: 10.1016/J.IPM.2019.102057. URL https://doi.org/10.1016/j.ipm.2019.102057.

[42] Quyu Kong, Marian-Andrei Rizoiu, Siqi Wu, and Lexing Xie. Will this video go viral: Explaining and predicting the popularity of youtube videos. In *Companion of*

The Web Conference, pages 175–178. ACM, 2018. doi: 10.1145/3184558.3186972. URL https://doi.org/10.1145/3184558.3186972.

[43] Vaia I. Kontopoulou, Athanasios D. Panagopoulos, et al. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, 15(8):255, 2023. doi: 10.3390/fi15080255. URL https://doi.org/10.3390/fi15080255.

[44] Haris Krijestorac, Rajiv Garg, and Vijay Mahajan. Cross-platform spillover effects in consumption of viral content: A quasi-experimental analysis using synthetic controls. *Inf. Syst. Res.*, 31(2):449–472, 2020. doi: 10.1287/ISRE.2019.0897. URL https://doi.org/10.1287/isre.2019.0897.

[45] Sachin Kumar and Alok Nikhil Jha. Fake news goes viral! determination and analysis of virality of socially relevant events in digital governance. In *15th International Conference on Theory and Practice of Electronic Governance*, pages 376–380. ACM, 2022. doi: 10.1145/3560107.3560165. URL https://doi.org/10.1145/3560107.3560165.

[46] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54 (1-3):159–178, 1992. doi: 10.1016/0304-4076(92)90104-Y. URL https://doi.org/10.1016/0304-4076(92)90104-Y.

[47] Elke B. Lange and Klaus Frieler. Challenges and opportunities of predicting musical emotions with perceptual and automatized features. *Music Perception: An Interdisciplinary Journal*, 36(2):217–242, 2018. doi: 10.1525/mp.2018.36.2.217. URL https://doi.org/10.1525/mp.2018.36.2.217.

[48] Daniel Le Compte and Daniel Klug. "it's viral!" - A study of the behaviors, practices, and motivations of tiktok users and social activism. In *Companion Publication of the 2021 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 108–111. ACM, 2021. doi: 10.1145/3462204.3481741. URL https://doi.org/10.1145/3462204.3481741.

[49] Eric T. Lehman. "Washing Hands, Reaching Out" - Popular Music, Digital Leisure and Touch during the COVID-19 Pandemic. *Leisure Sciences*, 43(1-2, SI):273–279, 3 2021. ISSN 0149-0400. doi: 10.1080/01490400.2020.1774013. URL https://10.1080/01490400.2020.1774013.

[50] Hanchao Li, Zhouhemu Tang, Xiang Fei, Kuo-Ming Chao, Ming Yang, and Chaobo He. A survey of audio MIR systems, symbolic MIR systems and a

music definition language demo-system. In *14th IEEE International Conference on e-Business Engineering*, pages 275–281. IEEE Computer Society, 2017. doi: 10.1109/ICEBE.2017.51. URL https://doi.org/10.1109/ICEBE.2017.51.

[51] Yifei Li and Li Shao. Using an epidemiological model to explore the interplay between sharing and advertising in viral videos. *Scientific Reports*, 14(1), 2024. doi: 10.1038/s41598-024-61814-9. URL https://doi.org/10.1038/s41598-024-61814-9.

[52] Chen Ling, Jeremy Blackburn, Emiliano De Cristofaro, and Gianluca Stringhini. Slapping cats, bopping heads, and oreo shakes: Understanding indicators of virality in tiktok short videos. In *Proceedings of the 14th ACM Web Science Conference*, pages 164–173. ACM, 2022. doi: 10.1145/3501247.3531551. URL https://doi.org/10.1145/3501247.3531551.

[53] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 4765–4774, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[54] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of math. statistics*, pages 50–60, 1947.

[55] Lucy March. Coming out online: Memetic authenticity in Rebecca Black's "Friday (Remix)". *Popular Communication*, 22(1):33 – 46, 2024. doi: 10.1080/15405702.2023.2287739. URL https://doi.org/10.1080/15405702.2023.2287739.

[56] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 11th ACM Conference on Web Science*, pages 173–182, Boston, MA, USA, 2019. ACM. doi: 10.1145/3292522.3326034. URL https://doi.org/10.1145/3292522.3326034.

[57] Maximilian Mayerl, Michael Vötter, Günther Specht, and Eva Zangerle. Pairwise learning to rank for hit song prediction. In *Datenbanksysteme für Business, Technologie und Web (BTW 2023)*, volume P-331 of *LNI*, pages 555–565. Gesellschaft für Informatik e.V., 2023. doi: 10.18420/BTW2023-26. URL https://doi.org/10.18420/BTW2023-26.

[58] Luiza de Melo-Gomes, Danilo B. Seufitelli, Gabriel P. Oliveira, Mariana O. Silva, and Mirella M. Moro. Análise do Sucesso Musical no Brasil Utilizando Dados do Twitter. In *Companion Proceedings of the 37th Brazilian Symposium on Databases*,

pages 40–46, Porto Alegre, RS, Brasil, 2022. SBC. doi: 10.5753/sbbd_estendido.2022.21841. URL https://doi.org/10.5753/sbbd_estendido.2022.21841.

[59] Michael Muhlmeyer, Shaurya Agarwal, and Jiheng Huang. Modeling social contagion and information diffusion in complex socio-technical systems. *IEEE Syst. J.*, 14(4):5187–5198, 2020. doi: 10.1109/JSYST.2020.2993542. URL https://doi.org/10.1109/JSYST.2020.2993542.

[60] Marily Nika, Thomas Wilding, Dieter Fiems, Koen De Turck, and William J. Knottenbelt. Going multi-viral: Synthedemic modelling of internet-based spreading phenomena. In *Proceedings of the 8th International Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS*, page 50 – 57, 2015. doi: 10.4108/icst.valuetools.2014.258221. URL http://doi.org/10.4108/icst.valuetools.2014.258221.

[61] Williams E Nwagwu and Ayobola Akintoye. Influence of social media on the uptake of emerging musicians and entertainment events. *Information Development*, 2023. doi: 10.1177/02666669221151162. URL https://doi.org/10.1177/02666669221151162.

[62] Gabriel P. Oliveira and Mirella M. Moro. Frequent genre mining on hit and viral songs. *Journal of Information and Data Management*, 16(1):136–145, 2025. doi: 10.5753/jidm.2025.4676. URL https://doi.org/10.5753/jidm.2025.4676.

[63] Gabriel P. Oliveira, Mariana O. Silva, Danilo B. Seufitelli, Anisio Lacerda, and Mirella M. Moro. Detecting collaboration profiles in success-based music genre networks. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, pages 726–732, Montreal, Canada, 2020.

[64] Gabriel P. Oliveira, Beatriz F. Paiva, Ana Paula Couto da Silva, and Mirella M. Moro. Characterizing the diffusion of misinformation regarding the coronavac vaccine in brazil. In *Proceedings of the 11th Brazilian Workshop on Social Network Analysis and Mining*, pages 204–215, Porto Alegre, RS, Brasil, 2022. SBC. doi: 10.5753/brasnam.2022.223173. URL https://sol.sbc.org.br/index.php/brasnam/article/view/20529.

[65] Gabriel P. Oliveira, Arthur P. G. Reis, Felipe A. N. Freitas, Lucas L. Costa, Mariana O. Silva, Pedro P. V. Brum, Samuel E. L. Oliveira, Michele A. Brandão, Anisio Lacerda, and Gisele L. Pappa. Detecting inconsistencies in public bids: An automated and data-based approach. In *Proceedings of the 28th Brazilian Symposium on Multimedia and Web*, pages 182–190, New York, NY, USA, 2022. ACM. doi: 10.1145/3539637.3558230. URL https://doi.org/10.1145/3539637.3558230.

[66] Gabriel P. Oliveira, Arthur P. G. Reis, Bárbara M. A. Mendes, Clara A. Bacha, Lucas L. Costa, Gabriel L. Canguçu, Mariana O. Silva, Victor Caetano, Michele A. Brandão, Anisio Lacerda, and Gisele L. Pappa. Ferramentas open-source de qualidade de dados para licitações públicas: Uma análise comparativa. In *Proceedings of the 37th Brazilian Symposium on Databases*, pages 116–127, Porto Alegre, RS, Brasil, 2022. SBC. doi: 10.5753/sbbd.2022.224351. URL https://doi.org/10.5753/sbbd.2022.224351.

[67] Gabriel P. Oliveira, Bárbara M. A. Mendes, Clara A. Bacha, Lucas L. Costa, Larissa D. Gomide, Mariana O. Silva, Michele A. Brandão, Anisio Lacerda, and Gisele L. Pappa. Assessing data quality inconsistencies in brazilian governmental data. *Journal of Information and Data Management*, 14(1), 2023. doi: 10.5753/jidm.2023.3220. URL https://doi.org/10.5753/jidm.2023.3220.

[68] Gabriel P. Oliveira, Bárbara M. A. Mendes, Camila S. Braz, Lucas L. Costa, Mariana O. Silva, Michele A. Brandão, Anisio Lacerda, and Gisele L. Pappa. Ranqueamento de licitações públicas a partir de alertas de fraude. In *Proceedings of the 12th Brazilian Workshop on Social Network Analysis and Mining*, pages 204–215, Porto Alegre, RS, Brasil, 2023. SBC. doi: 10.5753/brasnam.2023.232105. URL https://doi.org/10.5753/brasnam.2023.232105.

[69] Gabriel P. Oliveira, Ana Flávia C. Moura, Natércia A. Batista, Michele A. Brandão, Andre Hora, and Mirella M. Moro. How do developers collaborate? investigating github heterogeneous networks. *Software Quality Journal*, 31(1):211–241, 2023. doi: 10.1007/s11219-022-09598-x. URL https://doi.org/10.1007/s11219-022-09598-x.

[70] Gabriel P. Oliveira, Mariana O. Silva, Danilo B. Seufitelli, Gabriel R. G. Barbosa, Bruna C. Melo, and Mirella M. Moro. Hot streaks in the music industry: identifying and characterizing above-average success periods in artists' careers. *Scientometrics*, 128(11):6029–6046, 2023. doi: 10.1007/S11192-023-04835-X. URL https://doi.org/10.1007/s11192-023-04835-x.

[71] Gabriel P. Oliveira, Ana Paula Couto da Silva, and Mirella M. Moro. Analyzing the temporal relation between virality and success in the brazilian music market. In *Proceedings of the 51st Integrated Software and Hardware Seminar*, pages 157–168, Porto Alegre, RS, Brasil, 2024. SBC. doi: 10.5753/semish.2024.2656. URL https://doi.org/10.5753/semish.2024.2656.

[72] Gabriel P. Oliveira, Ana Paula Couto da Silva, and Mirella M. Moro. A quantitative comparison of viral and hit songs in the brazilian music market. *Vórtex Music*

Journal, 12:1–29, 2024. doi: 10.33871/vortex.2024.12.8727. URL https://doi.org/10.33871/vortex.2024.12.8727.

[73] Gabriel P. Oliveira, Ana Paula Couto da Silva, and Mirella M. Moro. What makes a viral song? Unraveling music virality factors. In *Proceedings of the 16th ACM Web Science Conference*, pages 181–190, New York, NY, USA, 2024. ACM. ISBN 9798400703348. doi: 10.1145/3614419.3644011. URL https://doi.org/10.1145/3614419.3644011.

[74] Gabriel P. Oliveira, Mariana O. Silva, Lucas G. L. Costa, Marco Tulio Dutra, and Gisele L. Pappa. ICPSet: A structured dataset of public procurement items. In *Proceedings of the 6th Dataset Showcase Workshop*, pages 103–113, Porto Alegre, RS, Brasil, 2024. SBC. doi: 10.5753/dsw.2024.243826. URL https://doi.org/10.5753/dsw.2024.243826.

[75] Gabriel P. Oliveira, Ana Paula Couto da Silva, and Mirella M. Moro. On the causal relationship between music virality and success. *IEEE Access*, 13:122782–122791, 2025. doi: 10.1109/ACCESS.2025.3589173. URL https://doi.org/10.1109/ACCESS.2025.3589173.

[76] Gabriel P. Oliveira, Ana Paula Couto da Silva, and Mirella M. Moro. Music virality on social platforms: A literature review. *Vórtex Music Journal*, 13:1–32, 2025. doi: 10.33871/vortex.2025.13.10848. URL https://doi.org/10.33871/vortex.2025.13.10848.

[77] Gabriel P. Oliveira, Luca Vassio, Ana Paula Couto da Silva, and Mirella M. Moro. Contagious rhythms: A wave-based epidemic approach for music virality on social platforms. In *Social Networks Analysis and Mining - 17th International Conference, ASONAM 2025, Niagra Falls, Canada, August 25-28, 2025*, Lecture Notes in Computer Science. Springer, 2025.

[78] Gabriel P. Oliveira, Luca Vassio, Ana Paula Couto da Silva, and Mirella M. Moro. Modeling music popularity as an epidemic: insights from the brazilian market. In *Proceedings of the 14th Brazilian Workshop on Social Network Analysis and Mining*, pages 79–92, Porto Alegre, RS, Brasil, 2025. SBC. doi: 10.5753/brasnam.2025.8760. URL https://doi.org/10.5753/brasnam.2025.8760.

[79] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021. doi: 10.1136/bmj.n71. URL http://doi.org/10.1136/bmj.n71.

[80] Bruna C. M. Paula, Gabriel P. Oliveira, and Mirella M. Moro. Mood analysis during the covid-19 pandemic in brazil through music. In *Companion Proceedings of the 28th Brazilian Symposium on Web and Multimedia*, pages 53–56, Porto Alegre, RS, Brasil, 2022. SBC. doi: 10.5753/webmedia_estendido.2022.227063. URL https://sol.sbc.org.br/index.php/webmedia_estendido/article/view/21984.

[81] Judea Pearl. *Causality*. Cambridge university press, 2009.

[82] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[83] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011. doi: 10.5555/1953048.2078195. URL https://dl.acm.org/doi/10.5555/1953048.2078195.

[84] Elia Pizzolitto. Music in business and management studies: a systematic literature review and research agenda. *Management Review Quarterly*, 74:1439–1472, 2024. doi: 10.1007/s11301-023-00339-3. URL https://doi.org/10.1007/s11301-023-00339-3.

[85] Leandro Rodrigues Ramos, Karin Satie Komati, Francisco de Assis Boldt, and Jefferson Oliveira Andrade. Geração semiautomática de valores de referência para identificação de obstruções em lingotamento contínuo. In *Proceedings of the 47th Integrated Software and Hardware Seminar*, pages 116–127, Cuiabá, Brazil, 2020. SBC. doi: 10.5753/semish.2020.11322. URL https://doi.org/10.5753/semish.2020.11322.

[86] Jing Ren and Robert J. Kauffman. Understanding music track popularity in a social network. In *European Conference on Information Systems*, pages 374–388, Guimarães, Portugal, 2017. AIS. URL http://aisel.aisnet.org/ecis2017_rp/25.

[87] Bruno Ribeiro. Modeling and predicting the growth and death of membership-based websites. In *23rd International World Wide Web Conference*, pages 653–664. ACM, 2014. doi: 10.1145/2566486.2567984. URL https://doi.org/10.1145/2566486.2567984.

[88] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference*

on Web Search and Data Mining, pages 399–408. ACM, 2015. doi: 10.1145/2684822.2685324. URL https://doi.org/10.1145/2684822.2685324.

[89] Dora P. Rosati, Matthew H. Woolhouse, Benjamin M. Bolker, and David J. D. Earn. Modelling song popularity as a contagious process. *Proceedings of the Royal Society A*, 477(2253):20210457, 2021. doi: 10.1098/rspa.2021.0457. URL https://doi.org/10.1098/rspa.2021.0457.

[90] Daniel Folador Rossi et al. Identificação de estáticas em poços de petróleo utilizando motifs. In *SEMISH*, pages 308–319, João Pessoa, Brazil, 2023. SBC. doi: 10.5753/semish.2023.230748. URL https://sol.sbc.org.br/index.php/semish/article/view/25083.

[91] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019. doi: 10.1126/sciadv.aau4996. URL https://doi.org/10.1126/sciadv.aau4996.

[92] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023. doi: 10.1038/s43017-023-00431-y. URL https://doi.org/10.1038/s43017-023-00431-y.

[93] Rahil Sachak-Patwa, Nabil T. Fadai, and Robert A. Van Gorder. Understanding viral video dynamics through an epidemic modelling approach. *Physica A: Statistical Mechanics and its Applications*, 502:416 – 435, 2018. doi: 10.1016/j.physa.2018.02.083. URL https://doi.org/10.1016/j.physa.2018.02.083.

[94] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[95] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*, pages 92–96, Austin, USA, 2010. scipy.org. doi: 10.25080/MAJORA-92BF1922-011. URL https://doi.org/10.25080/Majora-92bf1922-011.

[96] Danilo B. Seufitelli, Gabriel P. Oliveira, Mariana O. Silva, Gabriel R. G. Barbosa, Bruna C. Melo, Juliana E. Botelho, Luíza Melo-Gomes, and Mirella M. Moro. From Compact Discs to Streaming: A Comparison of Eras within the Brazilian Market. *Revista Vórtex*, 10(1):1–28, 2022. doi: 10.33871/23179937.2022.10.1.2. URL https://doi.org/10.33871/23179937.2022.10.1.2.

[97] Danilo B. Seufitelli, Gabriel P. Oliveira, Mariana O. Silva, and Mirella M. Moro. MGD+: An Enhanced Music Genre Dataset with Success-based Networks. In *Proceedings of the 5th Dataset Showcase Workshop*, pages 36–47. SBC, 2023. doi: 10.5753/dsw.2023.233826. URL https://doi.org/10.5753/dsw.2023.233826.

[98] Danilo B. Seufitelli, Gabriel P. Oliveira, Mariana O. Silva, Clarisse Scofield, and Mirella M. Moro. Hit song science: a comprehensive survey and research directions. *J. New Music Res.*, 52(1):41–72, 2023. doi: 10.1080/09298215.2023.2282999. URL https://doi.org/10.1080/09298215.2023.2282999.

[99] Ravi S. Sharma and Tushar Pandey. The impact of electronic word-of-mouth in the distribution of digital goods. *Webology*, 8(1):1–13, 2011. ISSN 1735-188X.

[100] Ravi S. Sharma, Miguel Morales-Arroyo, and Tushar Pandey. The emergence of electronic word-of-mouth as a marketing channel for the digital marketplace. *Journal of Information, Information Technology, and Organizations*, 6:41 – 61, 2011.

[101] Seungkyu Shin and Juyong Park. On-chart success dynamics of popular songs. *Advances in Complex Systems*, 21(3-4):1850008, 2018. doi: 10.1142/S0219525918500 08X. URL https://doi.org/10.1142/S021952591850008X.

[102] Benjamin Shulman, Amit Sharma, and Dan Cosley. Predictability of popularity: Gaps between prediction and understanding. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*, pages 348–357. AAAI Press, 2016. doi: 10.1609/icwsm.v10i1.14748. URL http://doi.org/10.1609/icwsm.v10i1.14748.

[103] Mariana O. Silva, Gabriel P. Oliveira, Danilo B. Seufitelli, Anísio Lacerda, and Mirella M. Moro. Collaboration as a driving factor for hit song classification. In *Brazilian Symposium on Multimedia and Web*, pages 66–74. ACM, 2022. doi: 10.1145/3539637.3556993. URL https://doi.org/10.1145/3539637.3556993.

[104] Mariana O. Silva, Gabriel P. Oliveira, and Mirella M. Moro. Analyzing character networks in portuguese-language literary works. In *Proceedings of the 12th Brazilian Workshop on Social Network Analysis and Mining*, pages 115–126, Porto Alegre, RS, Brasil, 2023. SBC. doi: 10.5753/brasnam.2023.230585. URL https://doi.org/10.5753/brasnam.2023.230585.

[105] Mariana O. Silva, Gabriel P. Oliveira, Danilo B. Seufitelli, and Mirella M. Moro. Collaboration-aware hit song prediction. *Journal on Interactive Systems*, 14(1):201–214, 2023. doi: 10.5753/jis.2023.3137. URL https://doi.org/10.5753/jis.2023.3137.

[106] Mariana O. Silva, Gabriel P. Oliveira, Danilo B. Seufitelli, and Mirella M. Moro. Temporal Success Analyses in Music Collaboration Networks: Brazilian and Global Scenarios. *Revista Vórtex*, 11(2):1–27, 2023. doi: 10.33871/23179937.2023.11.2.7185. URL https://doi.org/10.33871/23179937.2023.11.2.7185.

[107] Mariana O. Silva, Gabriel P. Oliveira, and Mirella M. Moro. Data insights on gender representation: Analyzing the book and music industries. In *Companion Proceedings of the 39th Brazilian Symposium on Databases*, pages 338–347, Porto Alegre, RS, Brasil, 2024. SBC. doi: 10.5753/sbbd_estendido.2024.243743. URL https://doi.org/10.5753/sbbd_estendido.2024.243743.

[108] Mariana O. Silva, Gabriel P. Oliveira, and Mirella M. Moro. Premiação das mulheres na literatura e na música: análises de dados da billboard e do goodreads. In Cristina Scheibe Wolff and Elaine Schmitt, editors, *A internet como campo de disputas de gênero*, pages 185–197. Cultura e Barbárie, 2024. doi: 10.29327/536640 7.1-17. URL https://doi.org/10.29327/5366407.1-17.

[109] Abhishek Singhi and Daniel G. Brown. Can song lyrics predict hits. In *International Symposium on Computer Music Multidisciplinary Research*, pages 457–471, Plymouth, UK, 2015. The Laboratory of Mechanics and Acoustics. URL https://cmmr2019.prism.cnrs.fr/Docs/proceedingsCMMR2015.pdf.

[110] Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[111] Shzr Ee Tan. Singapore Takes the 'Bad' Rap: A State-Produced Music Video Goes 'Viral'. *Ethnomusicology Forum*, 18(1):107–130, 2009. ISSN 17411912; 17411920. doi: 10.1080/17411910902793915. URL https://doi.org/10.1080/17411910902793915.

[112] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *J. of Lang. and Social Psychology*, 29(1):24–54, 2010. doi: 10.1177/0261927X09351676. URL https://doi.org/10.1177/0261927X09351676.

[113] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018. doi: 10.1080/00031305.2017.1380080. URL https://doi.org/10.1080/00031305.2017.1380080.

[114] Ramine Tinati, Thanassis Tiropanis, and Leslie Carr. An approach for using Wikipedia to measure the flow of trends across countries. In *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*,

page 1373 – 1377, 2013. doi: 10.1145/2487788.2488177. URL https://doi.org/10.1145/2487788.2488177.

[115] Thanassis Tiropanis, Wendy Hall, Jon Crowcroft, Noshir S. Contractor, and Leandros Tassiulas. Network science, web science, and internet science. *Commun. ACM*, 58(8):76–82, 2015. doi: 10.1145/2699416. URL https://doi.org/10.1145/2699416.

[116] Eleana Tsiara and Christos Tjortjis. Using twitter to predict chart position for songs. In *IFIP Artificial Intelligence Applications and Innovations*, pages 62–72. Springer, 2020. doi: 10.1007/978-3-030-49161-1_6. URL https://doi.org/10.1007/978-3-030-49161-1_6.

[117] Michael Vötter, Maximilian Mayerl, Günther Specht, and Eva Zangerle. HSP datasets: Insights on song popularity prediction. *Int. J. Semantic Comput.*, 16 (2):233–255, 2022. doi: 10.1142/S1793351X22400104. URL https://doi.org/10.1142/S1793351X22400104.

[118] Kevin Wallsten. "Yes we can": How online viewership, blog discussion, campaign statements, and mainstream media coverage produced a viral video phenomenon. *Journal of Information Technology and Politics*, 7(2-3):163 – 181, 2010. doi: 10.1080/19331681003749030. URL https://doi.org/10.1080/19331681003749030.

[119] Niels Werber, Daniel Stein, Jörg Döring, Veronika Albrecht-Birkner, Carolin Gerlitz, Thomas Hecken, Johannes Paßmann, Jörgen Schäfer, Cornelius Schubert, and Jochen Venus. Getting Noticed by Many: On the Transformations of the Popular. *Arts*, 12(1), 2023. ISSN 2076-0752. doi: 10.3390/arts12010039. URL https://doi.org/10.3390/arts12010039.

[120] Vishnu Srinivasa Murthy Yarlagadda and Shashidhar G. Koolagudi. Content-based music information retrieval (CB-MIR) and its applications toward the music industry: A review. *ACM Comput. Surv.*, 51(3):45:1–45:46, 2018. doi: 10.1145/3177849. URL https://doi.org/10.1145/3177849.

[121] Shlomo Yitzhaki. Relative deprivation and the gini coefficient. *The Quart. J. of Economics*, 93(2):321–324, 1979. doi: 10.2307/1883197. URL https://doi.org/10.2307/1883197.

[122] Kenji Yokotani and Masanori Takano. Social contagion of cyberbullying via online perpetrator and victim networks. *Comput. Hum. Behav.*, 119:106719, 2021. doi: 10.1016/J.CHB.2021.106719. URL https://doi.org/10.1016/j.chb.2021.106719.

[123] Eva Zangerle, Michael Vötter, Ramona Huber, and Yi-Hsuan Yang. Hit song prediction: Leveraging low- and high-level audio features. In *Proceedings of the 20th*

International Society for Music Information Retrieval Conference, pages 319–326, Delft, The Netherlands, 2019. ISMIR.

Appendix A

Characterization Details of Viral and Hit Songs

This appendix presents the details of the characterization of hit and viral songs presented in Chapter 3. Specifically, we present the translation of the Portuguese terms that appear in the topics (Section A.1).

A.1 Translation of Portuguese terms

Here, we present the translation of Portuguese terms of Tables 3.8 and 3.9. Such terms carry several meanings that are directly linked to the Brazilian and other Portuguese-speaking countries' cultures, which may not be translated accurately into English.

A.1.1 Global

H4. sit, take, grab, can, young woman, house, but, call, then, put

H5. yeah, here, want, floor, do, today, butt, now, can, was

H6. love, people, heart, man, know, to miss (someone or something), life, everything, because, kiss

V4. sit, take, love, everything, want, here, house, wh*re, today, four

V5. play, love, so, people, want, shake, life, butt, like that, know

V6. yeah, want, love, can, grab, drag, face, slap, but, people

A.1.2 Brazil

H10. sit, take, want, play, today, floor, butt, butt, then, go down

H11. yeah, everything, life, today, then, here, God, want, love, forever

H12. love, people, want, life, heart, everything, to miss (someone or something), nothing, time, mouth

V10. sit, take, want, play, want, go down, yeah, today, can, call

V11. love, everything, life, people, want, today, here, time, because, nothing

V12. butt, then, want, punch, everything, deny, mouth, floor, four, put