Attention Dynamics on the Chinese Microblogging Site Sina Weibo

by
Hao Cui

Submitted to
Central European University
Department of Network and Data Science

In partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Network Science

Supervisor: János Kertész

Vienna, Austria
2023

I, the undersigned, Hao Cui, candidate for the degree of Doctor of Philosophy in Network Science at the Central European University Department of Network and Data Science, declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of works of others, and no part the thesis infringes on any person's or intstitution's copyright. I also declare that no part the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 30 September 2022

崔浩 Hao Cui

Signature

Abstract

In this thesis, I investigate attention dynamics on the Chinese microblogging site Sina Weibo by studying the popular hashtags on the real-time Hot Search List (HSL). The objective goal of my thesis is to provide insights into the emergence mechanisms and popularity dynamics of hashtags on the one hand, and the effects of interventions by the microblogging service providers on the other hand. Furthermore, I analyze the consequences of a strong external effect with long lasting impact on the popularity ranking list on the example of COVID-19. To achieve that, I investigate the evolution dynamics of the repost network of popular hashtags in the phase of their emergence, the ranking dynamics after the successful hashtags reach the system-wide level of popularity, and identify anomalous patterns that can be attributed to intentional measures taken by the service provider deviating from the principles of an objective ranking.

Firstly, I study the network dynamics in the prehistory of successful hashtags before they become popular. I show that the time of the day when the hashtag is born has an impact on the time needed to get to the HSL. Analyzing this time I distinguish two extreme categories which I label: a) “Born in Rome”, which means hashtags are mostly first created by super-hubs or reach super-hubs at an early stage during their propagation and thus have the chance to gain immediate wide attention from the broad public, and b) “Sleeping Beauty”, meaning the hashtags gain little attention at the beginning and reach system-wide popularity after a considerable time lag. The evolution of the repost networks of successful hashtags before getting to the HSL shows two types of growth patterns: “smooth” and “stepwise”. The former is usually dominated by a super-hub and the latter results from consecutive waves of contributions of smaller hubs. The repost networks of unsuccessful hashtags exhibit a simple pattern of evolution.

Secondly, I study the dynamics of hashtags after they appear on the HSL, including rank trajectory clustering, duration, and ranking dynamics. I characterize the rank dynamics by the time spent by hashtags on the list, the time of the day they appear there, the rank diversity, and by the ranking trajectories. I show how the circadian rhythm affects the popularity of hashtags, and observe categories of their rank trajectories by a machine learning classification algorithm. By analyzing patterns of the ranking dynamics I identify anomalies that are likely to result from the platform provider’s intervention into the ranking, including the anchoring of hashtags to certain ranks on the HSL. I propose a simple model of ranking that explains the mechanism of this anchoring effect.

Thirdly, I present the effect of COVID-19 on Weibo’s public attention dynamics by studying the correlations between different content categories of hashtags with real-world COVID situations in different periods of the pandemic. I show how the specific events, measures and developments during the epidemic affected the emergence of different kinds of hashtags and the ranking on the HSL. I give an analysis of how the hashtag topics changed during the investigated time span and conclude that there were three periods to distinguish during the time of observation. In period 1, I see strong topical correlations and clustering of hashtags; in period 2, the correlations are weakened, without a clustering pattern; in period 3, I see an increased clustering but not as strong as in period 1. I further explored the dynamics of HSL by measuring the ranking dynamics and the lifetimes of hashtags on the list. This way I could obtain information about the decay of attention, which is important for decisions about the temporal placement of governmental measures to achieve permanent awareness. Furthermore, I find an abnormally higher rank diversity in the top 15 ranks on HSL due to the COVID-19 related hashtags, indicating the possibility of intervention from the platform provider.

To my parents, who made my endeavors possible.

Acknowledgements

I had a wonderful four years of my PhD. I would like to acknowledge the people who have influenced me during this fantastic journey.

First of all, I must thank my supervisor Prof. János Kertész, who led me to science, trusted my ability, and consistently supported me throughout my PhD studies. I am very fortunate and grateful for being his student. I learned to see things from a broader scope.

I would extend numerous thanks to the great faculty members in our department. I would like to thank Prof. Gerardo Iñiguez for many useful discussions and suggestions about my thesis project. I thank Prof. Elisa Omodei, Prof. Márton Karsai, Prof. Júlia Koltai, Prof. Federico Battiston, Edward Lee for providing much useful advice and encouragement regarding my future career development. I thank Prof. Petra Kralj Novak, Prof. Tiago Peixoto, Prof. Balazs Vedres and Prof. Márton Pósfai for much useful advice and support for my study. I also thank Prof. Rossano Schifanella for his amazing lectures that aroused my interests in machine learning.

I am fortunate to have our beloved coordinator Olga Peredi, PhD students, postdocs and visitors in our department. I extend special thanks to my student buddy Manran Zhu, Postdoc Lisette Espín Noboa who provided me loving support and encouragement. I thank the elder cohort who welcomed and guided me at the beginning of my PhD, Luis Guillermo Natera Orozco, Tamer Khraisha, Júlia Perczel, Matteo Neri, Srebrenka Letina, Johannes Wachs, Dávid Deritei, Orsolya Vásárhelyi. I thank colleagues in my cohort: Juliana Pereira, Rafiazka Hilman, Teodoro Criscione, Abdullah Alrhmoun, who provided an inspiring environment for me to study and grow. I thank the younger cohorts who are active in extracurricular activities: Maria Gabriela Juncosa Calahorrano, Felipe Vaca, Ludovico Napoli, Isabela Rosario Villamil, Elsa Andres, Adriana Manna, Luka Blagojevic, Sandeep Chowdhary, Leonardo Di Gaetano, Leonardo Rizzo, Yijing Chen, Onkar Sadekar, Júlia Számely, Sina Sajjadi, Martí Medina Hernández, Clara Eminente, Helcio Felippe, Sebastian Kusch, Timur Naushirvanov, Lorenzo Betti for making my PhD life colorful. Outside of CEU, my peers also influenced me during schools and conferences. I had a memorable time with Rebeka O. Szabó and Milán Janosov during the CSS summer school in Berlin. I learned from Pablo Lozano during the CCS conference in Lyon. I had great collaboration with Jan Bachmann during the complex system winter workshop at Saline Royale in France.

I thank my friends Zhitian, Imane, Ruichuan, Vanessa, Lea, Jasmine, Momo, Ting, Gulshan, Romayssa who have broadened my worldview outside of my field to the real society.

I thank my Master's thesis supervisor Prof. Tamás Prőhle at Eötvös Loránd University who introduced me to Bayesian networks that later made it possible for me to pursue a PhD program in network science.

Lastly, I thank my parents, who love me wholeheartedly and provide me with endless encouragement for me to pursue my goals. They are the greatest parents. I dedicate this work to them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and problem statement

With the development of technology transformation, social media has become inevitable source of information in people's daily life. There are trending contents on social media everyday, attracting a lot of attention from the public. Netizens can be connected in social networking sites such as the well-known microblogging sites Twitter and the Chinese equivalent Sina Weibo.

Users on microblogging sites generate hundreds of millions of posts per day, some of which contain one or multiple hashtags referring to the topic of the posts. In the flood of information, hashtags achieve different levels of popularity at a certain time, and the most popular hashtags in the whole system are depicted in ranking lists to inform users. Twitter Trends and Sina Weibo Hot Search List (HSL) are the examples of microblogging ranking lists which show in real time the most popular hashtags that gain wide attention in the whole microblogging system. These lists serve as indication of public interest and attention [7, 25] and, at the same time, they trigger collective awareness [8] of the latest trending topics or events emerging in the world. These trends or hot topics can originate from natural reaction to real-world events [96] or from manipulation by companies and organizations [76].

Contents generated from an ordinary user are very likely to be buried in the flooding newly-generated online information and never become a trending topic. This raises a natural question, how do these trending topics become popular? Is there a way to gain some understanding of successful hashtags by studying the emergence of their popularity from a network science perspective? After the popular hashtags succeed on the HSL, their popularity fluctuates with time as shown by the dynamics of their scores and rank changes on the HSL. What patterns can be observed from the ranking dynamcis of the popular hashtags on the HSL? Is there a connection between the ranking patterns and the history before they appear on the HSL? For a hashtag, being popular enough on the HSL can be desirable since massive public attention could result in financial gains and many other benefits. Apart from well-known tools of direct advertisements or propaganda [132, 90, 9, 25, 26], individuals, companies, and political actors are tempted to make use of hidden manipulation in order to attract public attention. Hashtag popularity can also be undesirable in the case of a negative influence. Are we able to find evidence of any forms of interventions towards these popular hashtags? Last but not least, what influence do exogenous factors and contingencies exert on some properties of the microblogging complex system? Unexpected situations, shocks, such as the COVID-19 pandemic can dramatically influence our lives. How does the pandemic influence public attention dynamics on Chinese social media?

1.2 Significance of the problem

Social networking sites have risen in the information age and have become an inseparable part of our everyday lives. Social media is central in attracting people's attention and shaping public perception [36, 43]. Public attention is important in many aspects, including education, politics, marketing, and pandemic management [4, 31]. Popular contents on social media are approximations of public concerns and attention. Therefore, understanding the dynamics of popularity on social media platforms is of great interest. Understanding the emergence of popular topics plays important role in business intelligence, governance, and trend predictions in the real world [63, 96]. Understanding the public attention dynamics during the pandemic serves an essential role in pandemic management, contributing to minimizing the effect. As service providers have financial interests and may be under political pressure, the objectivity of the ranking lists on social media and the truth of their content can be questionable. Similarly to the fight against fake news, the fight against manipulation of public attention is in the interest of the society and it also needs the tools of detecting interventions. Deepening the knowledge of public attention dynamics on social media promotes instant and effective communications among governments, health experts and the public, helping the government to monitor public opinion, maintain the stability of the society as well as develop and deliver more effective measures to minimize the societal effects in the case of a pandemic.

1.3 Contributions

In this thesis, I intend to give answers to the previously raised questions by studying Sina Weibo. In terms of the popularity emergence of the hashtags, I focus on how they become popular. I firstly unfold the different routes leading successful hashtags to the HSL by studying the evolution patterns of the whole repost networks as well as their giant components during the prehistory, i.e., the time period from the birth of the hashtags till their first appearance on the HSL. I characterize two extreme types of popularity emergence mechanisms namely “Born in Rome” and “Sleeping Beauty” depending on the length of the prehistory. I have found that the birth timing of the hashtags play an important role in determining whether they will be popular enough to appear on the HSL.

As a next step, I dig deeper into the ranking dynamics of hashtags on the HSL. I describe the patterns of the ranking dynamics, uncover their relationship with the circadian mode of Weibo activity, and also establish a link to the prehistory of the hashtags. I identify anomalous ranking characteristics of the HSL which can be attributed to the interventions into the ranking by the service provider and propose a model of anchoring to explain the anomalies. The aim of the identification of regularities in the ranking dynamics was twofold: First, contribution to the quantitative characterization of the dynamics of public attention in order to better understand its mechanism, and second, finding signatures of interventions by the service provider.

Lastly, I try to understand how a major event with long standing effect influences the hashtag dynamics. I focused on how COVID-19 related hashtags acquired popularity on Sina Weibo during the first four months of the pandemic and partly dominated it. In this study, I have provided a novel approach to studying and quantifying the attention dynamics by taking advantage of the real-time HSL on Sina Weibo. Instead of capturing the public attention using segmented key words from retrieved posts, we adopted the “rank diversity” [66] measure on Weibo HSL to describe the rank dynamics of real-time popular hashtags which can serve as an indication of public attention in the Sina Weibo system. I have been able to track the evolution of public attention in different periods during the pandemic, follow how the attention of the population shifted from one group of topics to another and study the changing correlation patterns of different COVID-related topic categories with the evolving COVID-19 situation in the world. I obtain information about the decay of attention by exploring the ranking dynamics and hashtag lifetimes on the list, which is important for decisions about the temporal placement of governmental measures to achieve permanent awareness. Furthermore, I discovered abnormally higher rank diversity in the top 15 ranks on HSL due to the COVID-19 related hashtags, revealing the possibility of algorithmic or direct intervention from the platform provider. On the one hand, such study contributes to the understanding of the dynamics of public attention on social media and how it reflects the dynamics of the public thoughts and behaviors. On the other hand, identifying the patterns of online attention dynamics and the relationship to events and measures during pandemic may contribute to its efficient management.

1.4 Thesis structure

This thesis is structured in the following way:

- Chapter 2: In this chapter, I give an introduction of microblogging sites and their ranking lists of popular topics. Specifically, I review the emergence of popularity, popularity competition, interventions on popularity, and the impact of COVID-19 pandemic.
- Chapter 3: In this chapter, I investigate the emergence mechanisms of popular hashtags on Sina Weibo by studying their repost network evolution dynamics in the prehistory before they become popular enough to appear on the HSL.
- Chapter 4: In this chapter, I focus on the dynamics of successful hashtags after they appear on the Sina Weibo HSL by studying the patterns of their rank trajectories, dynamics of duration and ranking. I pinpoint signatures of anomalies from the observations and propose a ranking model with anchoring effect to simulate the interventions on the ranking list by the platform provider.
- Chapter 5: In this chapter, I conduct the study of attention dynamics on the Sina Weibo HSL during the COVID-19 pandemic by separating different observation periods and semantic contents of the hashtags on the HSL. I also made a comparison of the hashtag ranking dynamics before and after the outbreak.
- Chapter 6: In this chapter, I conclude the thesis by summarizing the findings together, and providing outlooks for future research direction.

Chapter 2

Literature review

2.1 Microblogging sites and popular contents

One of the consequences of the technological revolution is that a massive amount of data is provided by humans while interacting with digital devices on daily basis [3]. With the spreading of smart phones and other mobile devices, there is a growing tendency that people's communication gets shifted to the digital domain and they obtain information about "hot issues" happening across the world [61] through social media.

Microblogging is a modern communication tool by which users generate and share tremendous information on many aspects such as politics, economy, sports and entertainment [33]. Users of such services can follow others, establishing complex follower-followee networks, thus creating important social media platforms, where social connections are carriers of information [94]. The users interact with each other through various kinds of activities such as post, repost, comment, reply, mention or alike. User-generated posts on microblogging sites usually contain bits of information, or "memes" [13] that are brief text updates and/or micromedia such as photos, video or audio clips. A meme will be reposted if it is appealing to the other users. Reposting behaviors make the posts visible to the followers of the user who share the posts due to the follower-followee relationship. Through the process of virality, a large number of users can be reached with a particular meme, forming trends on microblogging platforms.

2.1.1 Twitter and trending list

Twitter is one of the world's most famous microblogging sites which offers worldwide service, with 217 million monetizable daily active users [106] as of the fourth quarter of 2021. Messages on microblogs may be identified using keywords called hashtag [1]. Users on Twitter sometimes tweet and retweet tweets containing a hashtag (# symbol) before a relevant keyword or phrase, for example #covid-19. Hashtags help categorization of tweets and increases visibility in Twitter search [101].

A "trend" on Twitter is a subject of interest that is being discussed at a greater rate than the others in this system. Some trends contain hashtags. By Twitter's definition, trending topics are immediately popular, rather than those have been popular for a while or on a daily basis [104]. The exact algorithm that extracts trending topics is unknown. A trending topic on Twitter consists of the topic itself as well as a stream of tweets which include that topic. Twitter trending list shows trending topics which are usually hot topics and latest events happening around the globe. Studying Twitter trending topics enables researchers to learn about people's opinions and sentiments about the current trends [39, 46], and make predictions for a wide range of outcomes such as the COVID-19 waves [28], crimes [17], financial market prices [88, 1], elections [85] and so on.

2.1.2 Sina Weibo and Hot Search List (HSL)

Sina Weibo, a Twitter-like microblogging system, is the most popular microblogging site in mainland China [131], where Twitter and other Western social media are blocked. Sina Weibo and Twitter are two major online social networking sites in the world, their users are mostly disjoint, barely overlap geographically, representing the online users of China and the rest of the world [41]. Sina Weibo has increasingly penetrated into the daily lives of ordinary Chinese individuals [127] and become one of the most popular media in information diffusion and communication [33] since it was launched in August 2009 [58]. Sina Weibo is a major vehicle of self expression especially for young Chinese people and a forum for social movements [44]. Although Sina Weibo is less known worldwide and has received less academic attention than Twitter, it has surpassed Twitter in terms of user size, with 248 million daily active users and 573 million monthly active users [97] as of the third quarter of 2021.

[1] A hashtag is a type of label or metadata tag used on social network and microblogging services and have been used to mark individual messages as relevant to a particular topic or "channel" [135].

As the counterpart of Twitter in China, Sina Weibo shares many similarities in size, structure, and influence with Twitter. Both networking platforms show directed user relationship, allow unlimited inbound followers and attract hundreds of millions of users. Sina Weibo works similarly to Twitter as users may follow others and have followers, they can post texts and pictures, add hashtags to them, react to others' posts, and repost them, either from personal computers or mobile devices. Weibo replaced the 140-character limit for users' posts to 2000 characters [20] since Februrary 2016. There are also differences with the two platforms. Instead of using one hashtag at the beginning like Twitter, topics on Weibo are confined in double hashtags one at the beginning and one at the end of the topic description, for example, #Pneumonia of unknown cause detected in Wuhan#. In this thesis, we use hashtag as the synonym of topic, we refer a hashtag on Sina Weibo as the content contained within the double #s.

HSL history and popularity criteria

The popularity of hashtags on Sina Weibo emerges as users participate in the search for them, in the discussion on them and in their spreading. Like Twitter, Sina Weibo also creates a ranking list of popular hashtags to inform users. Sina Weibo HSL is a section that displays the 50 most popular hashtags in real-time [127]. The range of topics covered by hashtags which occur on the HSL is very diverse, to name a few, social events, TV programs, celebrities, entertainment, health and politics. Accordingly, Sina Weibo HSL nowadays has become a popular tool for the Chinese netizen to look for information, follow hot topics, news events and celebrity gossip.

Sina Weibo HSL has gone through several changes. Back in 2014, a real-time list of hottest searches was launched on the client with an updating frequency of once every ten minutes [110], allowing users to see the latest hot information anytime, anywhere. In 2017, the updating frequency of the hashtags together with their ranks accelerated to every minute [110]. Until 2021 there used to be one or two advertisement rank positions included in the top 50 ranks[2].

[2] There used to be commercial hashtags that are paid for the positions on the HSL, which often occupy the third or/and fifth ranks and marked with "Recommendation (荐)". Since 2021, Weibo went through another reform that the commercial hashtags are randomly placed between ranks 3 and 4 or/and ranks 6 and 7, without a rank number associated in front of the advertisements, resulting in maximum 52 hashtags on the HSL, excluding the imposed positive energy recommendation position (hot search top) above all ranks which is often promotion for positive contents. Sina Weibo has also separated a new list specifically for

As the name of the ranking list indicates, the Hot Search List is based to a large extent on the search activities related to the hashtags. The search volume for a hashtag indicates its popularity and on Sina Weibo this quantity is the main underlying for the HSL. The search volume index of a hashtag is a comprehensive measure [84] which takes into account multiple dimensions such as the number of searches in Sina Weibo and the quality of the user accounts involved in the search, for the aim of preventing manipulated fake popularity [84]. However, the concrete algorithm of determining the ranking of the hashtags on the HSL has been unknown and that Weibo has been target of criticism for making financial gains. This can be understood as the HSL serves as an advertising tool to boost popularity, getting to the HSL not only informs users about popularity but also amplifies it a lot, which, in turn, may have severe financial consequences. As a response to the criticism, on 23 August 2021, Sina Weibo released what it called the rule of capturing the "hotness" H of a hashtag at a certain time [2]. The corresponding formula is as follows:

$$H = (S_H + D_H + R_H) \times I_H, \tag{2.1}$$

where S_H is search hotness, referring to the search volume, including manual input search and click-and-jump search, discussion hotness D_H is for the amount of discussion, including original posting and re-posting, reading hotness R_H represents the volume of readings in the spreading process of the hashtag, and interaction hotness I_H refers to the interaction rate of hot search results page.

While Sina Weibo emphasizes the objectiveness and fairness of HSL, it admits at the same time to "promote positive content" and that "official media reports shall prevail" in case of major negative social events [2] and the intervention in other cases (redundancy, serious inaccurate information as identified by government departments of content inducing severe conflicts).

2.2 Attention dynamics on microblogging sites

In our times of information deluge, the dynamics of public attention is of eminent importance from many aspects, including education, politics, marketing and governance. In the new media, the flow of information has dramatically accelerated, leading often to

entertainment and celebrity related hashtags, though in the normal HSL there are still entertainment related hashtags. The data in this study excludes those commercial hashtags.

rapidly changing public attention. At the same time these media provide unprecedented possibilities to study attention dynamics [117, 80] as they produce Big Data open for investigation. The microblogging service Twitter [102] is particularly suited to provide the basis for quantitative studies on the dynamics of public attention as the content of the messages is available [103]. Accordingly, Twitter data have been used to identify classes of dynamic collective attention [53], investigate party-related activity and to predict election outcomes [32]. Furthermore, Twitter data have served as the basis of modelling attention dynamics during pre-election time [49] or studying the relationship between public attention and social emotions [71].

2.2.1 Emergence of popularity

Popularity may be of different kinds: immediate popularity upon release, and ever-increasing popularity in succeeding periods [87]. One emergent phenomenon in a microblogging complex system is the emergence of popularity of a topic. When it comes to the definition of popularity on social media, researchers had various metrics by regarding popularity of an online content as the frequency of daily occurrence [130], the number of reposts/views at a time [10, 63], the cascade size [38], topic pervasiveness in a community [11], or straightforwardly the displayed list of popular items whose popularity metric is defined by corresponding social media platform providers.

Researchers on Twitter information diffusion have observed variation in the spread of online information across topics [79], shown the effects of the content [100], semantic characterization [54] and the co-occurrence of hashtags [73] on popularity. Approaches have been introduced to predict the popularity peaks of the new hashtags on Twitter from the perspectives of machine learning [64], and taking the context of Twitter social network [123]. Previous study on Twitter trending topics has shown retweets by other users are more important than the number of followers in determining trends [8]. Some studies modeled user behaviors to capture the emergence of Twitter trending topics based on characteristics of the retweet graphs [96]. A recent study on Twitter trending topics has investigated real-time Twitter Trends detection along with the ranking of the top terms [48].

2.2.2 Popularity competition

The ranking of trending contents on social media changes over time, following the rise and fall of public attention: Old trends vanish and new trends emerge. The rank of a hashtag on the list increases or decreases, reflecting the relative popularity of the topic to other topics at a certain instant of time. The amount of time a hashtag stays on the list indicates the degree of its ability to attract consistent public attention.

Research studies on Twitter trending list have analyzed the dynamics of trending topics through comprehensive statistical analysis from the aspects of lexical composition, trending time, trend re-occurrence [7], etc. There have been different factors identified, which contribute to the success of a topic, like novelty of the piece of information and the resonance level of the messages spread as well as the influence of certain members of the propagating network [78]. The evolution of Twitter trends is characterized by phases of burst, peak and fade [8] and the patterns of temporal evolution of popularity of hashtags have been ordered into six different categories [122].

Research on Sina Weibo popularity appeared soon after its launch. Researchers studying popularity in Sina Weibo have presented evolution analysis of trending topics [125], long-term variation of popularity [130], prediction of hot topics based on content quality and structural characteristics of early adopters [10], as well as bursty human activity patterns [118]. The attention of the researchers have turned gradually to the HSL. In essence, the HSL is a real-time hot topic detecting system, which provides a (supposedly) objective ranking on the hotness of the topics [127]. Within each topic, a vast number of messages keep evolving and mutating as the topic flows through the network. In this scenario, a topic is an incarnation of a meme, while the messages spreading along different threads are operational proxies to track its dynamics. Weibo system is generally considered an ideal laboratory for investigating information contagion, especially for Chinese content. Research on Weibo hot topics has focused on topic dynamics from the perspective of time, geography, demographics, emotion, retweeting, and correlation [35], on similarities and differences to Twitter [126, 124], emergence mechanisms [26], patterns of popularity evolution [51], prediction [134, 60, 136], social emotions and diffusion patterns [34] as well as impact of censorship [16].

The ranking dynamics on Weibo HSL enables us to investigate the dynamics of popular attention in such details which is not available on other platforms. We will focus on the temporal evolution of this ranking in order to study the attention dynamics. Recently, ranking dynamics has been studied widely from sports [12, 21, 66] to scientists, journals or companies [12]. There are stable rankings (word frequencies) with little or no changes in the ranks [67] and there are volatile ones with vivid dynamics (mentions of Twitter hashtags) [12]. Clearly, Weibo HSL belongs to the latter with rich dynamic properties. This ranking provides a proxy for the attention preferences of Weibo users enabling the quantification of the dynamics thereof, which reflects the changes in the attention due to events and measures.

2.2.3 Interventions on popularity

Trending topics on Twitter can be manipulated by concerted efforts by fans of celebrities or cultural phenomena. The economic and political relevance of popularity of items on online media is an incentive to the service providers to intervene into the trending lists. A linear influence model [121] was introduced to capture the network effect on endogenous diffusion of hashtags on Twitter trending list and demonstrate evidences of manipulation [132] on the observed dynamics. Certain trending topics on Twitter may be opportunistically targeted with desirable qualities by spammers [90]. Recent studies on Twitter trends have found likely presence of coordinated campaigns in AstroTurf version to influence and manipulate public opinion during the COVID crisis in Mexico [74].

In the case of Weibo, the exposure of hashtags on the HSL has a great promotional effect thus many are keen to be on the HSL or disappear from it in the case of a negative influence [89], resulting in strong competition and manipulation attempts. Studies reveal that hashtags from different topical categories differ in time length of prehistory (from birth till first appearance on the HSL) and the types of accounts involved in the propagation [26]. Celebrity and entertainment related hashtags are often associated with marketing accounts [133] which can be influenced by social capital [26]. Studies have also identified online censorship [16, 109] control practices and the possibility of algorithmic intervention [25, 27] on Sina Weibo. Recent findings indicate possibility of algorithmic intervention [25] from the platform provider towards COVID-related hashtags on the HSL during the COVID-19 pandemic. Research indicated that human editorial decisions were involved in the curation of Weibo trending topics with the aim of increasing user engagement [56] and that Weibo actively facilitates the production and spread of online contention to attract more users through a range of recommendation mechanisms built into the platform, including the trending topic list and channels such as Sina-owned official accounts [56].

2.2.4 Impact of COVID-19 on public attention

Public attention becomes a focal issue in times of crises like pandemics. As early as 2010, four years after it was launched, Twitter was shown to be an adequate, real-time content, sentiment, and public attention trend-tracking tool [18] and was used to study rapidly-evolving public sentiment with respect to the epidemic H1N1 [86]. The analysis of tweets enabled to quantify the difference between attention and fear and their distancedependence in the case of the Ebola epidemic [107]. For the present pandemic COVID-19, Twitter studies on public attention have occurred, focusing on the perception of policies by the public [128, 62], fighting COVID-19 misinformation [72], and the psychological impacts associated with social media exposure during the pandemic [37, 30].

The first COVID-19 epicenter was in China where the service of Twitter is blocked, whereas its local substitute Sina Weibo, is very popular [50]. Therefore, it is natural to use data from Weibo for similar purposes as was introduced earlier for Twitter in other countries. Scientists have already recognized that this microblogging service provides important insight into the function of the Chinese society [98, 69]. Recently, some studies have appeared dealing with the public attention towards COVID-19 on Sina Weibo. The attention of Chinese netizens to COVID-19 was measured by analyzing keyword frequency in retrieved COVID-related posts from randomly sampled Weibo users [120] or hashtags on Sina Weibo Hot Search List (HSL) [119], and was studied for identifying the sentiment and emotion trends of public attention [119, 57].

Social media search indices have shown correlations with the epidemic curve [83, 5, 45, 75, 55] and have been used for prediction of the transmission of infectious diseases, such as cholera [19], Ebola [59], influenza [5] and Middle East Respiratory Syndrome (MERS) [83]. For the recent COVID-19, online search trends such as Google Trends and Baidu Index have shown strong correlations with real-world cases and deaths [45, 75, 55] and were used to predict the number of new suspected or confirmed cases [75, 55]. In these studies, the searched keywords on social media are mainly the symptoms and the names of the disease. Nevertheless, there are various aspects of the pandemic that influence the society, including the measures and regulations given by the government, the scientific knowledge provided by healthcare experts, aspects related to frontline doctors and nurses, the donation behaviors between countries, the remote working new norm and so on.

Collective attention towards online news decays with time due to the fade of novelty and attractiveness in the competition with the other news [117]. Recent studies show a peak of collective attention towards COVID-19 in late January 2020 and a subsequent collapse, in terms of the dynamic behavior of words used on Twitter [6, 29].

Chapter 3

Emergence of hashtag popularity on Sina Weibo

3.1 Introduction

To understand the emergence of hashtag popularity in online social networking complex systems, we study the largest Chinese microblogging site Sina Weibo, which has a Hot Search List (HSL) showing in real time the ranking of the 50 most popular hashtags based on search activity. We investigate the prehistory of successful hashtags from 17 July 2020 to 17 September 2020 by mapping out the related interaction network preceding the selection to HSL. We have found that the circadian activity pattern has an impact on the time needed to get to the HSL. When analyzing this time we distinguish two extreme categories: a) “Born in Rome”, which means hashtags are mostly first created by super-hubs or reach super-hubs at an early stage during their propagation and thus gain immediate wide attention from the broad public, and b) “Sleeping Beauty”, meaning the hashtags gain little attention at the beginning and reach system-wide popularity after a considerable time lag. The evolution of the repost networks of successful hashtags before getting to the HSL show two types of growth patterns: “smooth” and “stepwise”. The former is usually dominated by a super-hub and the latter results from consecutive waves of contributions of smaller hubs. The repost networks of unsuccessful hashtags exhibit a simple evolution pattern.

Why and how does some online information become popular is an important question. While many studies have been focusing on the "why" part, putting efforts in identification of factors that lead to information popularity, less attention has been paid to what the patterns of the temporal network dynamics look like in the history of these contents before they achieve a high level of popularity measured in our case by getting to the HSL. Different routes to popularity have been observed for memes. Studies on information diffusion have found pervasive existence of two consecutive spikes of popularity in the diffusion of different information from different media during the whole lifetime of a meme's propagation [129]. Studies on tweet's popularity have shown tipping points [63] may emerge through the lifecycle. What makes the history of those popular hashtags specific? What mechanisms can be deduced from the network dynamics in the early stage? In our study, we focus particularly on the popularity emergence period, which is the prehistory *before* a hashtag becomes popular enough to appear on the HSL.

Our goal is to unfold the different routes of hashtags leading them to the HSL by studying the repost networks as well as their giant components during the time period from the first creation of the hashtags to their first appearance on the HSL. We investigate the influencing factors of the time needed for a successful hashtag to get to the HSL and identify the time of the day when the hashtag was born and the effect of huge hubs. The evolution of the repost network show either smooth or stepwise character which are also related to the above factors.

3.2 Materials and methods

3.2.1 Data description

We wrote a web scraper to crawl Sina Weibo HSL from 17 July 2020 to 17 September 2020, with a frequency of every 5 minutes. We extracted 10144 hashtags that have appeared on the HSL during this time period and traced back the original user-generated posts containing these hashtags during the time interval from birth till first appearance on the HSL. There are unavoidable problems related to the data. First, the crawling of the data was occasionally interrupted leading to loss of data. The estimated related data corruption is approximately 5%. Censorship is another source of information loss, one example is about the pop star Wu Yifan whose account was closed and all posts related to him were no longer available on Weibo ever since he was arrested by police due to several crimes [115]. Another source of data loss is due to a possibility provided by Sina Weibo, enabling users to choose privacy option, which hides their activities and makes it impossible to trace back the chain of reposts along that branch. The datasets supporting the conclusions of this study are available in the GitHub repository, https://github.com/cuihaosabrina/Emergence_Popularity_Sina_Weibo.

3.2.2 Network construction

In this section, we introduce how we construct our hashtag repost networks and the properties we study. The repost network (called retweet graph in the context of Twitter) is a standard tool to study the spreading of content on microblogging sites [96]. The temporal directed repost network consists of users as nodes and reposts as links between them, pointing in our definition towards the user who reposts. The repost networks of a hashtag contain all the origin nodes that posted this hashtag as well as the related reposts of the hashtag. Since all reposts have timestamps, the evolution of the repost network can be followed and can be traced back to the origins of the hashtag. The timestamp of the first post containing a hashtag is its birth time. We focus on the whole repost network and its largest connected component (LCC), which consists of the largest number of connected nodes in the whole hashtag repost network. When studying the component size, we disregard the directedness of the links. As the repost network evolves, the LCC identified at some stage of the evolution may be replaced by another more recently formed giant component at a later stage. In fact, this change happens often just before the hashtag almost reaches the HSL. We study the dynamics of the LCC in the repost network structure just before a hashtag reaches the HSL since it captures the most influential nodes and links during the process of the hashtag popularity emergence. For different hashtags, the growth rates of the whole network as well as the final LCC at different stages during prehistory period may have different growth characteristics as shown in Fig. 3.1. We compare, analyze, and categorize these patterns.

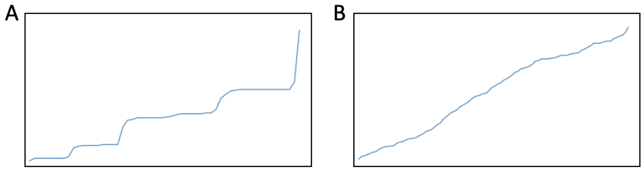

Figure 3.1: Shape examples of stepwise and smooth patterns of repost network cumulative link growth trajectories.

3.2.3 Classification of link growth trajectories

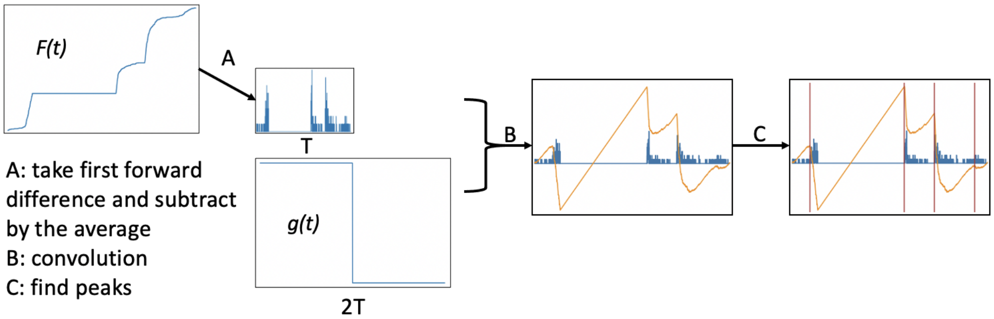
F(t)
A
T
A: take first forward difference and subtract by the average
B: convolution
C: find peaks
g(t)
2T
B
C

Figure 3.2: Example workflow of peak detection in repost network cumulative link growth trajectory $F(t)$. $F(t)$ is a discrete function, where $t \in \mathbb{Z}, 0 \leq t \leq T$, T is the total number of minutes in the prehistory. Step A: $\widetilde{f} = f'(t) - \bar{f}$, where $f'(t)$ is the first forward difference, $t \in \mathbb{N}, 1 \leq t \leq T$, and $\bar{f}$ is the average value of $f'(t)$. Step B: take the convolution $(\widetilde{f} * g)(t)$ where $g(t)$ is a step function. Step C: detect peaks by comparing with neighboring values in the convolved series.

We characterize the different repost network dynamics by studying growth patterns of the cumulative number of links $F(t)$ at a minute resolution. $F(t)$ is a discrete function, where $t \in \mathbb{Z}, 0 \leq t \leq T$, T is the total number of minutes in the prehistory. We use a classifier to distinguish between stepwise and smooth growth, which is based on the detection of local peaks in the derivative of the function $F(t)$. Since $F(t)$ is discrete, we obtain $f'(t), t \in \mathbb{N}$, $1 \leq t \leq T$ by taking the first forward difference of $F(t)$. Then we take the convolution of $\widetilde{f} = f'(t) - \bar{f}$ and $g(t)$, where $\bar{f}$ is the average value of $f'(t)$, and $g(t)$ is defined as follows

$$g(t) = \begin{cases} 1 & t \in \mathbb{N}, 1 \leq t \leq T \\ -1 & t \in \mathbb{N}, T+1 \leq t \leq 2T \end{cases}$$

In practice, we calculte the convolution $(\widetilde{f} * g)(t)$ using the convolve method in the numpy [42] Python module, with the mode parameter equals 'valid'. We find all local maxima by comparing with neighboring values in the convolved series $(\widetilde{f} * g)(t)$ using the peak detection function find_peaks in the scipy.signal [108] Python module. The principle of the classifier is demonstrated in Fig. 3.2. If there are more than two peaks identified and any of the time intervals between two consecutive peaks is greater than one hour, then we classify $F(t)$ as stepwise, otherwise smooth. The same classification procedure applies to the LCC. As for the repost network increment time series in Fig. 3.7E and Fig. 3.7F, the resize was done by using TimeSeriesResampler from tslearn [93] Python package, with the method of spline interpolation [113].

3.3 Results

We call those hashtags successful, which make it to the HSL. The prehistory of a successful hashtag prior to entering the HSL starts from the birth of the earliest post containing this hashtag and ends at the moment this hashtag first appears on the HSL. Does the birth time of a hashtag influence the time length of its prehistory? What are the patterns of the repost network dynamics and their relation with the time length of the prehistory? To answer these questions, we summarize the observed statistical patterns of the successful hashtags that have appeared on the HSL in the observation period.

3.3.1 Role of birth time

According to its size, China should have five geographic time zones [114] but it follows one single standard time, the Chinese (or Beijing) Standard Time. In principle, this could lead to the screening of any circadian pattern. However, Weibo users are densely distributed in the eastern and central regions of China [15] whose geographical time zones are similar and the population accounts for 65.8 percent of the national population [116]. The company Weibo Corporation has its headquarters in Beijing. In fact, we have detected clear circadian patterns.

The cumulative number of hashtags that have ever appeared on the HSL grows approximately linearly, as shown in Fig. 3.3. Zooming into the figure as seen in the inset in Fig. 3.3, a periodic pattern becomes visible indicating that practically no new hashtags appear on the HSL in certain time intervals during nights. The beginning and ending boundaries of the idle periods are sharp rather than gradual, leading to the suspicion of human control of the HSL and the controllers' working times follow a circadian pattern. This is in contrast to the claim of Sina Weibo [2] that the selection of hashtags to the HSL follows an automated procedure just based on a formula (see Eq. 2.1).

People creating the hashtags on Weibo are largely influenced by their circadian rhythm thus the number of launched hashtags shows according variations. Following the Weibo

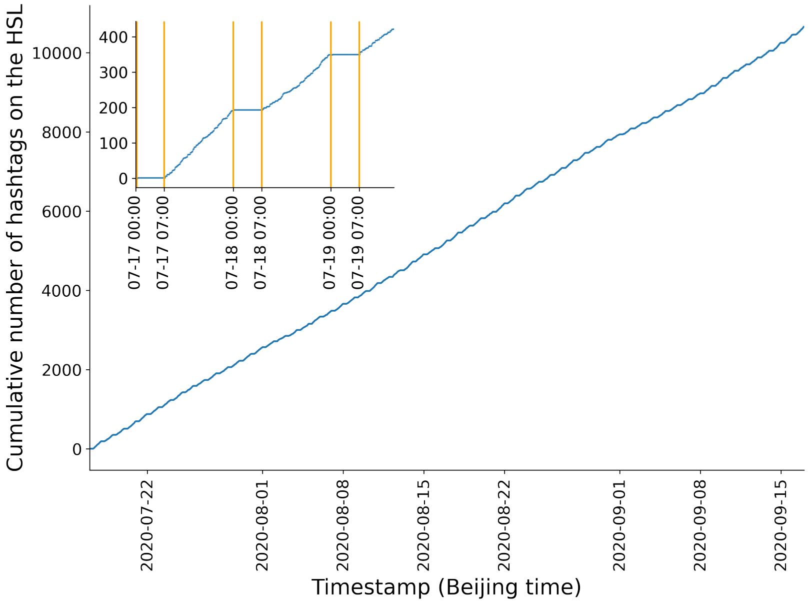

Figure 3.3: Growth of cumulative number of hashtags that have ever appeared on the Sina Weibo Hot Search List (HSL) with time, from 17 July 2020 to 17 September 2020. The inset enlarges the first two days and shows a clear circadian pattern as illustrated by the yellow lines. During the night time between midnight to approximately 7 am, practically no new hashtags appear on the HSL. The night time intervals are the horizontal parts in the figure, with an average size of 7.18 hours, and a standard deviation of 0.85 hours.

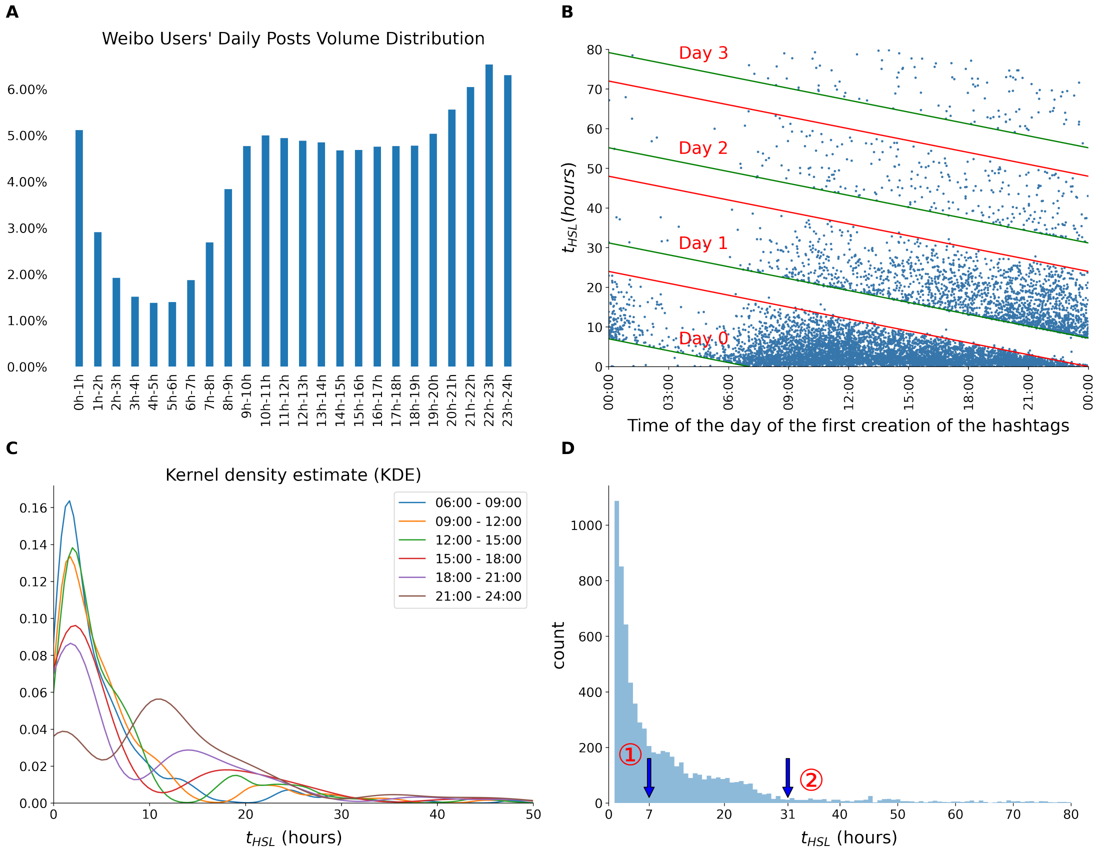

Figure 3.4: Statistics of 10144 hashtags that have appeared on Sina Weibo Hot Search List (HSL) from 17 July 2020 to 17 September 2020. (A) Distribution of Weibo users' daily posts volume according to Weibo User Development Report [14]. (B) Relationship between birth time of the day of the hashtags and the hours from birth to first appearance on the HSL, which we call the "HSL time" and denote as t_{HSL}. The vertical difference between two lines of the same color is 24 hours, the difference of a red line and a green line on y axis is 7.18 hours. All lines are parallel. (C) Parameterized probability density functions of the HSL time by different time intervals of birth time of the day, using kernel density estimation (KDE) [111], with the parameter bw = "scott" [82]. (D) Histogram of the HSL time. Section ① represents the category "Born in Rome" and section ② represents "Sleeping Beauty".

User Development Report [14], we show in Fig. 3.4A that the number of new user-generated posts gradually increases from around 5 am, reaching the first peak around noon followed by a small decrease from 1 pm – 2 pm, then a steady increase from 3 pm till the peak in the evening hours 10 pm – 11 pm, and then a final decay afterwards until 5 am. Figure 3.4B shows a scatter plot of hashtags' birth times of the day and the time length of the prehistories, which starts from the hashtag birth time until first appearance on the HSL, the "HSL time" denoted by t_{HSL}. Figure 3.4B shows the prehistory spanning for 4 days, separated by white stripes with vertical widths of 7.18 hours, which is the average of the night time periods in Fig. 3.3 when practically no new hashtags enter the HSL. A point within Day i section means that the corresponding hashtag enters the HSL after i days of its birth, i.e., Day 0 means it gets to the HSL within the same day as it was born. From the overall statistics in Fig. 3.4B, we could see that for the hashtags whose birth time of the day is in the morning, the time it takes to enter the HSL ranges from very immediate till around 10 hours in most cases. Hashtags born from midnight to 6 am enter only exceptionally the HSL; this stripe is practically empty, indicating an idle mode with some manually introduced special cases. Figure 3.4C describes the distribution of t_{HSL} of the hashtags in Fig. 3.4B parametrized on time intervals. For hashtags whose birth time of the day is after 9 pm, they will either get to the HSL in the same day within three hours, or they will show up after at least around seven hours. As the t_{HSL} gets longer, the hashtags that enter the HSL become fewer. Figure 3.4D shows the distribution of t_{HSL}, with a rapid decrease till 7 hours, followed by a slower and longer decrease afterwards.

3.3.2 Roads to success

Successful hashtags, which make it to the HSL, may have very different prehistories. We have seen that the time of the birth of the hashtag matters as it affects the time needed to get to the HSL. In general, we observed that there are hashtags, which get very fast to the HSL and others, which need rather long time. The hashtags belonging to the first group need very short time to get to the HSL - we call this group "Born in Rome". On the other hand, there is a group of hashtags which surpass a dormant period before discovered by a broader audience and get finally to the HSL - these are called "Sleeping Beauty". Furthermore, we are investigating the repost network during the prehistory, and explore the differences in its evolution and topology.

“Born in Rome”

As the proverb goes, “All roads lead to Rome”, so that those already born in Rome are more likely to succeed. The name suggests that these hashtags achieve success on the HSL easily as they usually immediately gain a huge attention wave or several attention waves shortly one after the other, reaching the HSL within a few hours. The attention wave-drivers are usually super-hubs or a crowd of smaller hubs. A super-hub is an influential node whose number of followers is huge and the positioning of the account is authoritative to the type of content it posts. To name a few such super-hubs, “CCTV News”(“央视新闻”, 126M followers), “People’s Daily” (“人民日报”, 145M followers), “Headline News” (“头条新闻”, 100M followers). Successful hashtags concerning accidents, crimes, natural disasters and other societal issues (called here “social”) are usually associated with the above mentioned super-hubs. For hashtags related to stars and entertainment, it is more often to see the contributions of series of smaller hubs to their success. Video examples of repost network evolution in the prehistory are available in the Supplementary Information. For this type of emergence mechanism of popularity, the time for a successful hashtag to enter the HSL is usually short. As Fig. 3.3 and Fig. 3.4B suggests, the closer the hashtags are born to midnight, the less likely they tend to appear on the HSL immediately, instead, they tend to show up after at least seven hours, making their prehistory longer. To factor out the influence of night hours, we consider hashtags whose time needed from birth till HSL within 7 hours to be in the category “Born in Rome”. As shown in Fig. 3.4D, at 7 hours we see the start of a shoulder which marks change in the shape of the count of hashtags versus waiting time. We have also seen some hashtags with very few (re)posts prior to the HSL, for example, #US orders 100 million doses of coronavirus vaccine from UK and France# (#美国从英法订购1亿剂新冠疫苗#), which could result from human intervention regarding international news.

“Sleeping Beauty”

We call another type of successful hashtags “Sleeping Beauty”, when the emergence mechanism results in relatively long time needed from birth till HSL. Hashtags in this category usually experience a low activity dormant period before being picked up by crucial influencing nodes. They might need several attention waves, and that the inter-wave time intervals can be long before a final trigger of significant popularity pushing them to the HSL. As marked in Fig. 3.4D, at around 31 hours the count of hashtags drops to a very low-level plateau. We use 31 hours (one day plus seven hours inactive night period) as the boundary for “Sleeping Beauty”, indicating that the delay is substantial and not due to the night break. When it comes to the hashtag content, as shown in Fig. 3.5, “Sleeping

Beauty" exhibits a higher proportion of the Stars and lower proportion of Social and International categories than "Born in Rome". See classification details in the Supplementary Information.

For the "Sleeping Beauty" category, as t_{HSL} increases, it is more likely to experience the "rebirth" of the same hashtag, so that the hashtags generated at a later time might not refer to the same event at the birth of the hashtag, though the hashtag itself remains unchanged. The examples are shown in the Supplementary Information. In order to avoid such cases, we restricted the "Sleeping Beauty" category to those with $t_{HSL} < 5$ days, resulting in altogether 571 hashtags in this category and crawled all their reposts. In addition, we produced an equal-sized random sample from the "Born in Rome" category. We crawled the number of followers of all users who participated in the posting behavior, with 69k users in total.

Relation with repost network dynamics

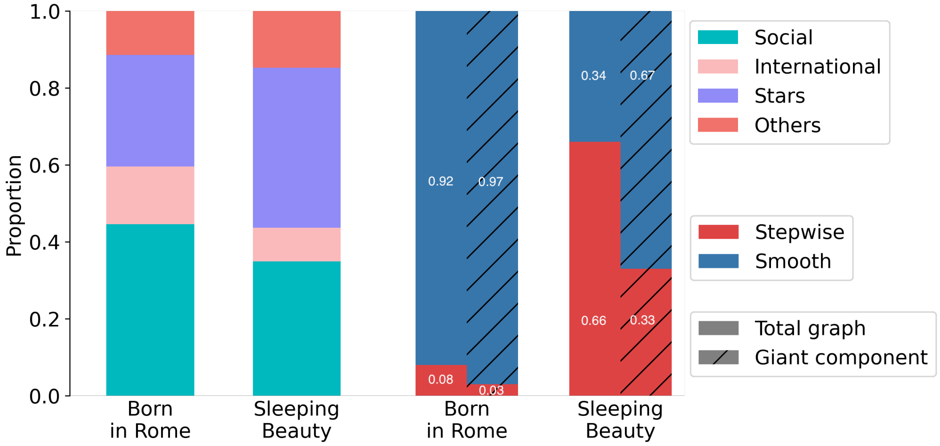

Figure 3.5: Distribution of hashtags from "Sleeping Beauty" category and the same size random sample from "Born in Rome" category by hashtag content and the shape of their link growth patterns, whether stepwise or smooth, of the whole repost network as well as the final giant component.

The hashtag repost network grows in time as we define it as the cumulative (or aggregate) network of the reposts of online users. Different repost networks vary in growth speeds and topological structures. We have studied the repost network dynamics of hashtags in the "Born in Rome" and "Sleeping Beauty" categories. Fig. 3.5 shows the ratio of

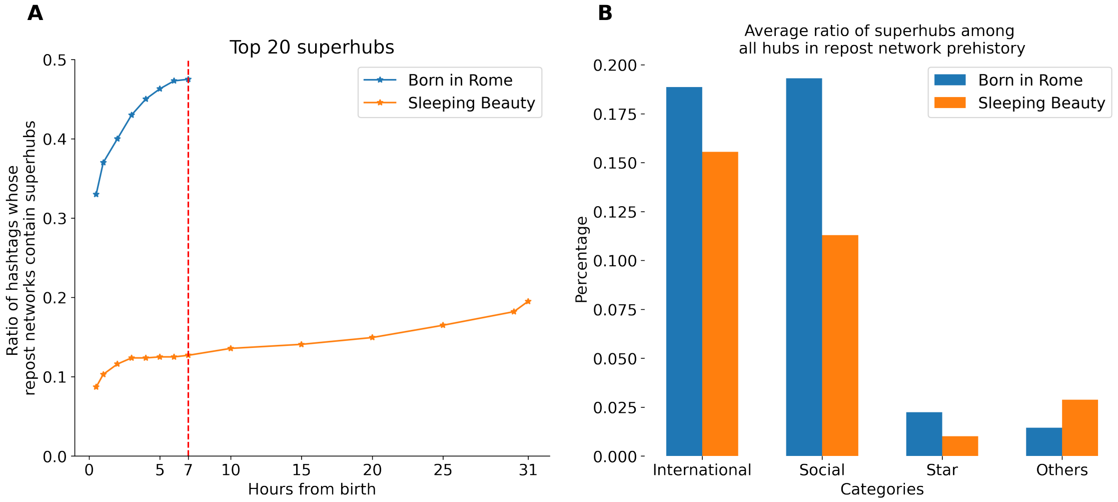

Figure 3.6: Role of super-hubs for hashtags in “Born in Rome” and “Sleeping Beauty” categories. (A) Ratio of hashtags whose repost networks in the prehistory contain at least one super-hub (top 20 largest hubs excluding celebrities). (B) Ratio of super-hubs among all hubs (nodes with top 10k largest degrees) in the prehistory.

different link growth pattern dynamics of the total network and the final giant component in the two categories (for examples see Fig. 3.1). As shown in Fig. 3.5, for the total repost network growth, the majority of “Sleeping beauty” have stepwise shape, meaning the necessity of several attention waves to gain the popularity to enter the HSL. As for the “Born in Rome” category, the majority hashtags have smooth shape in the repost network link growth, meaning that the power of the hub(s) at their early stage is enough to push the hashtags to reach system-wide popularity. The ratio of hashtags influenced by super-hubs (top 20 largest hubs excluding celebrities, see SI) in the prehistory is a function of time measured from the birth. As shown in Fig. 3.6A, this ratio starts at a higher value for “Born in Rome” category and increases rapidly, while for “Sleeping Beauty” category, it remains at a relatively low level during the whole prehistory period. As shown in Fig. 3.6B, super-hubs play a more important role in the categories of International and Social, while for the Star category, smaller hubs are dominant. The proportions of stepwise shape in the giant components of both categories are fewer than those of the total graph. This is reasonable since the formation starting time of the final giant component could be later than that of the whole repost network.

Failure and success

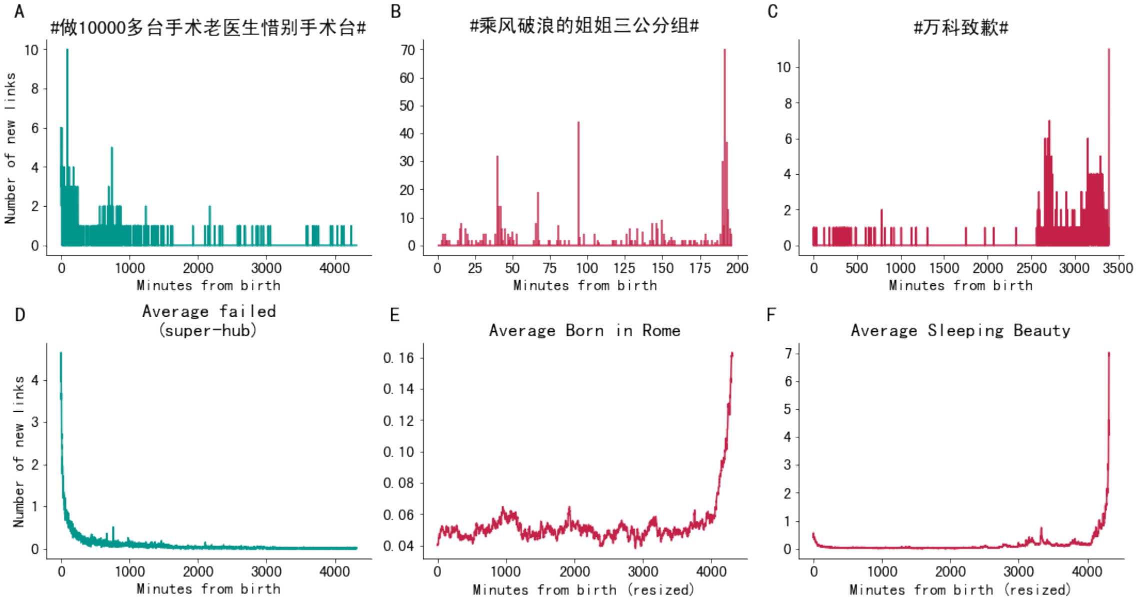

Figure 3.7: Comparison of repost network growth patterns of failed hashtags born in the super-hub “CCTV News” (“央视新闻”) as well as successful hashtags from the categories “Born in Rome” and “Sleeping Beauty”. Note the different time scales in the figures. (A) An example of a failed hashtag born in the super-hub “CCTV News”, #Doing more than 10000 operations, the old doctor bid farewell to the operating table# (#做10000多台手术老医生惜别手术台#). (B) An example of a hashtag from the category “Born in Rome”, #Sisters Who Make Waves (a variety show) grouping of the third public performance# (#乘风破浪的姐姐三公分组#). (C) An example of a hashtag from the category “Sleeping Beauty”, ,#Vanke apologizes# (#万科致歉#). (D) Average repost network growth pattern of 100 randomly selected failed hashtags from the super-hub “CCTV News” in late August 2020, lasting for three days (4320 minutes) from birth time. (E) Average repost network growth pattern of all “Born in Rome” sample hashtags, time length resized to 4320 for all hashtags. (F) Average repost network growth pattern of all “Sleeping Beauty” hashtags, time length resized to 4320 for all hashtags.

Hubs or super-hubs are needed for a hashtag to reach popularity, however, not all hashtags born in super-hubs are successful as many, in fact, the majority of them fail to land on the HSL. How does the growth pattern of the repost network of unsuccessful hashtags differ from those of successful ones? We took the super-hub #CCTV News# as an example and studied the repost network evolution of 100 randomly selected hashtags in late August 2020. One example is shown in Fig. 3.7A, the hashtag first attracted considerable attention, and then the attention decreased in a fluctuating manner and the temporary gains were not enough to compete with other hashtags for a position on the HSL. The averaged repost network growth pattern of the unsuccessful hashtags born in #CCTV News# is shown in Fig. 3.7D, in minute resolution. The network increment per minute shows a fast (exponential) decay and then a slower one as time goes on. In Fig. 3.7B and Fig. 3.7C, we show examples of hashtags from "Born in Rome" and "Sleeping Beauty" categories respectively. One or several attention waves are launched before the hashtags reach HSL, and the number of new links generally shows an increasing trend. Figure 3.7E and Fig. 3.7F show the averaged repost network growth patterns of "Born in Rome" and "Sleeping Beauty" hashtags respectively, with the time length resized to three days. The fast decay in the early time behavior of the averaged "Sleeping Beauty" curve is very similar to that of the unsuccessful ones, as shown in Fig 3.7D. The higher initial value for the unsuccessful hashtags is due to the fact that we selected the unsuccessful hashtags from those starting at the super-hub "CCTV News" which assured considerable early attention, while for the "Sleeping Beauty" hashtags we took all cases, irrespective of the popularity of the node where the hashtags were born. In contrast to the unsuccessful hashtags, "Sleeping Beauties" experience at a later stage a push in the attention dynamics due to getting picked up by a large hub which finally help them to get to the HSL.

3.4 Discussion

In this paper, we examined the emergence of hashtag popularity on the Chinese microblogging site Sina Weibo by analyzing prehistory of the repost network evolution of hashtags that finally get to the HSL. We have focused on the HSL time t_{HSL} and studied differences in the repost network dynamics of the whole network as well as of the final giant component for successful hashtags. Our studies have identified two extreme types of popularity emergence mechanisms for successful hashtags: That of the "Born in Rome" and of the "Sleeping Beauty", and pointed out the role of different hubs in the process. Compared with "Sleeping Beauty" hashtags, those in "Born in Rome" category tend to reach superhubs at an early stage of their spreading process, facilitating their success to the HSL. For "Sleeping Beauty" hashtags, instead of reaching super-hubs at an early stage, they are likely to gain several attention waves from smaller hubs, resulting often in a stepwise growth pattern of the repost network.

Previous studies on the emergence of popularity of entities in online text streams observed two patterns: a “bursty” one where content blasts into activity in public discourse without a precedent, and a “delayed” pattern which experiences a period of inactivity before resurfacing [40]. These patterns are similar to the “Born in Rome” and “Sleeping Beauty” discussed in this paper. As Fig. 3.5 shows, two thirds of sleeping beauty hashtags exhibit a stepwise shape of repost network growth, meaning that there are at least two peaks before reaching the HSL and there is considerable time delay between the peaks. In the case of two-peaks, the hashtag first experiences a hibernation period with low activity after birth, then a peak, then another hibernation, then a final peak and gets finally to the HSL. Our findings about sleeping beauties is in general in accordance with previous studies [129], though we consider only the prehistory phase of information diffusion. We also notice that in Fig. 3.5 around one third of sleeping beauty hashtags have smooth shape, meaning they only experience one long hibernation period before a peak that propel them to the HSL. For the final giant component growth pattern, the majority of hashtags have smooth shape, they experience one hibernation period before a final peak to reach HSL, implying the time period closer to the appearance on HSL is more important in determining the temporal popularity that makes the hashtags enter the HSL. For sleeping beauties, strategically locating or intervening the tipping point(s) could contribute to marketing efficiency and future popularity prediction [129], or to destroy the formation of trends in the case of misinformation.

When it comes to the content categories of the hashtags, we observe differences between Star and Social/International. For both “Born in Rome” and “Sleeping Beauty” hashtags, International and Social hashtags tend to have higher ratios of super-hubs than the Star hashtags. For the Star hashtags, non-super-hubs play a more important role in their popularity. In fact, it is often the case that the repost network evolution of hashtags related, e.g., to celebrities results from “collective efforts” and their popularity accounts for concerted influence of several smaller hubs which are usually marketing accounts. This emphasizes the importance of social capital in making hashtags related to stars popular enough to enter the HSL.

Though super-hubs are important in triggering hashtag popularity, by far not all hashtags created by the most prominent super-hubs make it to the HSL. The timing of the first creation of a hashtag is an important factor to its popularity evolution, since it influences the system-wide user attention level as well as the pool of the competing hashtags. From the statistics shown in Fig. 3.4A, the volume of user posts from midnight to 6 am is at most a factor of 2 lower than during a similar period in the rest of the day, while the proportion of successful hashtags in the same time period is significantly less as Fig. 3.4B suggests, indicating the disadvantage of hashtags born in that time period in achieving success. Sina

Weibo Hot Search List is of commercial value with an advertising effect on the hashtags by substantially increasing their visibility to the whole public. Understanding the mechanism of the emergence of hashtag popularity and the importance of timing, on the one hand, could contribute to marketing and maximizing the spreading efficiency by playing with these factors. On the other hand, it provides Weibo users with better knowledge to differentiate about the possible social capital influence in promoting certain contents, such as Star hashtags.

It is of great social value and importance to have hashtags on HSL that raise real public awareness and concern. Of course, even hashtags from super-hubs could fail to get to the list, let alone those from regular users with only a few followers. For hashtags in the latter category, it is hard to be successful. In fact, hashtags posted by normal users need to be (re)posted by influential ones to be promoted enough, leading to the necessity that influential nodes should get aware of their social responsibilities to participate in such situations where the voice of unprivileged people are unheard by the whole audience. More importantly, the prime responsibility is carried by the platform provider. It is challenging but important to use a fair algorithm to take into account of the opinion of the "invisible majority" (in terms of number of followers) and capture real hot topics that gain true attention from the public. In reality, there are signs that - in spite of the claims by Sina Weibo - the selection of the hashtags to the list is not entirely automated. One clear signature of this is the night break during which practically no new hashtags appear on HSL, however, time to time some do.

Although this study focuses on the emergence mechanism of the hashtag popularity specific to Weibo, we mention that our approach may shed some lights on more than just for this online microblogging complex system. Generally, for any successful cultural product, such as a song, a TV series, a best-seller book, etc., there is also a prehistory prior to the success when attention of the broad society is reached. During that prehistory, people interact with each other in relation to this product, for example, recommend, comment and consume. Such processes are in the focus of the study of innovations [77]. What interaction mechanisms lead to the success of a cultural product? Does the birth timing of this cultural product influence the time length it takes to achieve success? What are the differences in the popularity mechanisms between products from "Born in Rome" and "Sleeping Beauty" types if there are any? What are the components when forming the attention waves if there are several? Of course, the time scale for such products is very different from that of the hashtags, but the online digital "footprints" of the related social interactions could be very helpful in uncovering the important details of the processes.

3.5 Supplementary information

3.5.1 SI1 Evolution of repost networks before getting to the HSL

Here we give examples of hashtag repost networks and show in the movies how the repost networks of different categories of hashtags evolve before they get to the HSL [23].

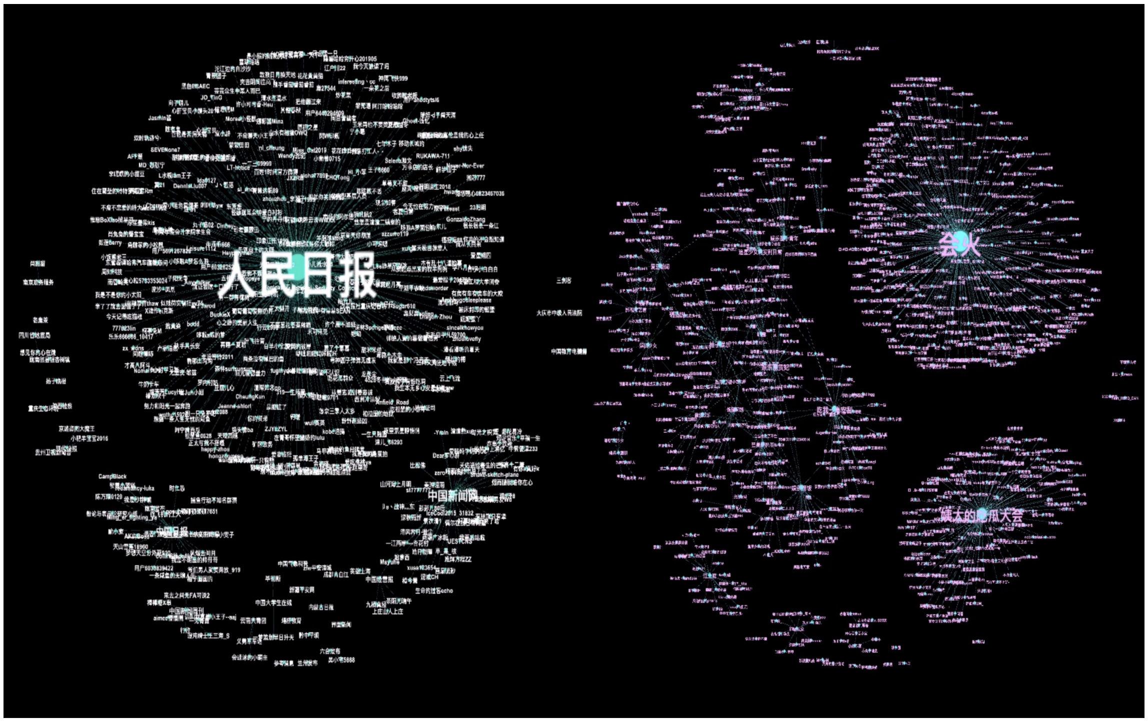

Figure 3.8: Examples of hashtag repost networks. Left: “Born in Rome”, right: “Sleeping Beauty”.

3.5.2 SI2 Hashtag categorization

We have classified the hashtags into the following categories based on human judgement: Social, Stars, International, and Others. The Social category consists of hashtags that are related to social accidents, crimes, natural disasters, and other events that are related to social life. The Stars category consists of movie/sports stars, singers, idols, celebrities as well as the TV programs/movies and events that they participate in. The International category consists of hashtags whose content is related to news outside of China. The Others category consists of the rest of the hashtags that fall into none of the above categories. Table 3.1 contains example hashtags and their translation.

3.5.3 SI3 Hashtag rebirth

As the prehistory length t_{HSL} increases, it is more likely to experience the hashtag "rebirth", that the hashtag posted after a few days might not refer to the same event as the birth of the hashtag, though the hashtag itself remains unchanged. We show some examples here.

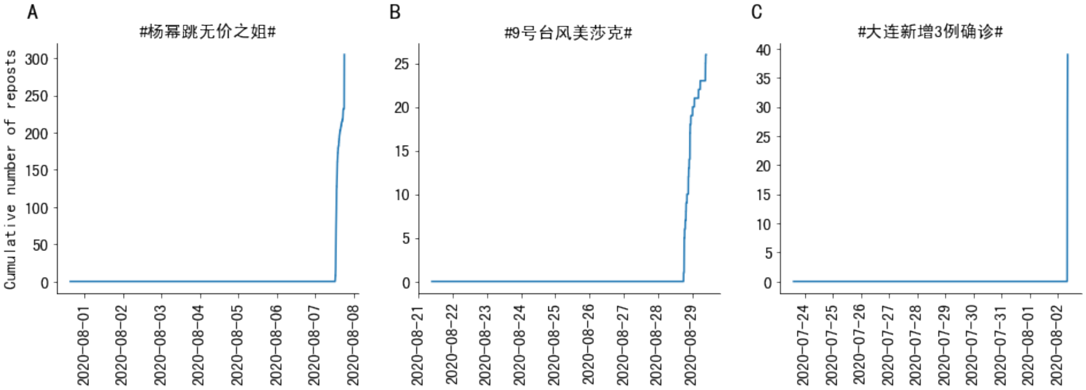

Figure 3.9: Examples of hashtag "rebirth". (A) #Yang Mi dances Priceless Sister# (#杨幂跳无价之姐#) (B) #Typhoon Maysak No. 9# (#9号台风美莎克#) (C) #3 new infected cases confirmed in Dalian# (#大连新增3例确诊#)

As shown in Fig. 3.9A, the hashtag #Yang Mi dances Priceless Sister# (#杨幂跳无价之姐#) first appeared on the HSL on 2020 August 7 at 17:57. It was born on 2020 July 31 15:24, with the content about the trailer of a TV show that the celebrity Yang Mi would dance Priceless Sister in the next episode, no reposts. The second post containing this hashtag was posted on 2020 August 7 at 12:09 when the TV show actually started. The hashtag at the birth and the hashtag created on August 7 refer to different sources. The success of the hashtag on the HSL was the result of the TV show on August 7 instead of the trailer one week ago.

As shown in Fig. 3.9B, the hashtag #Typhoon Maysak No. 9# (#9号台风美莎克#) first appeared on the HSL on 2020 August 29 at 09:54. It was born on August 21 at 09:25, with the content about the possibility that Typhoon Maysak might be coming soon. From 2020 August 21 15:42 on, no new (re)posts until 2020 August 28 when the Typhoon really formed. The posts containing the same hashtag at the later time refers to the real Typhoon rather than the warning at the beginning.

As shown in Fig. 3.9C, the hashtag #3 new infected cases confirmed in Dalian# (#大连新增3例确诊#) first appeared on the HSL on 2020 August 2 at 08:17. It was born on 2020 July 23 at 14:43. Though the hashtag remains the same, the content at the birth refers to the new infected cases at that time and the hashtag on 2020 August 2 refers to the new infected cases on August 1.

As the prehistory gets longer, it is more common to see the "rebirth" of the same hashtags referring to a different event from birth. In order to avoid such influences, when studying the properties of hashtags in the "Sleeping Beauty" category, we choose those hashtags whose $t_{HSL} < 5$ days.

3.5.4 SI4 Division of the categories "Rome" and "Beauty"

We choose hashtags whose $t_{HSL} < 7$ hours to be in the category of "Born in Rome", and 31 hours $< t_{HSL} < 5$ days to be in the category of "Sleeping Beauty", with the following reasons: (1) As shown in Fig. 3.3, and Fig. 3.4B, there is a period of around 7 hours during the night when almost no new hashtags appear on the HSL. If a hashtag was born close to midnight, then it is very likely that it will experience the 7-hour break and automatically not be in the category of "Born in Rome". Similarly, for "Sleeping Beauty", we use 24+7 = 31 as the boundary, indicating that the delay is substantial and not due to the night break.

(2) The limiting points 7 and 31 hours are shown in Fig. 3.4D, and mark changes in the shape of the count of hashtags vs waiting time. At 7 hours we see the start of a shoulder and at around 31 hours the count drops to a very low-level plateau. (3) The analyzed results in Fig. 3.5 using 7-31 and 8-30 are similar indicating that the somewhat arbitrary choice of the limits do not influence qualitatively the results, compare Fig. 3.5 with Fig. 3.10. The proportion of stepwise pattern in the Rome category is slightly higher than the 7-31 division, while the Beauty category remains the same. The division of content categories are also similar.

As for the intermediate category between Rome and Beauty, we took a random sample with size 400 from this category and the proportion of stepwise shape is 0.53, which is between the Born in Rome category and the Sleeping Beauty category, as expected.

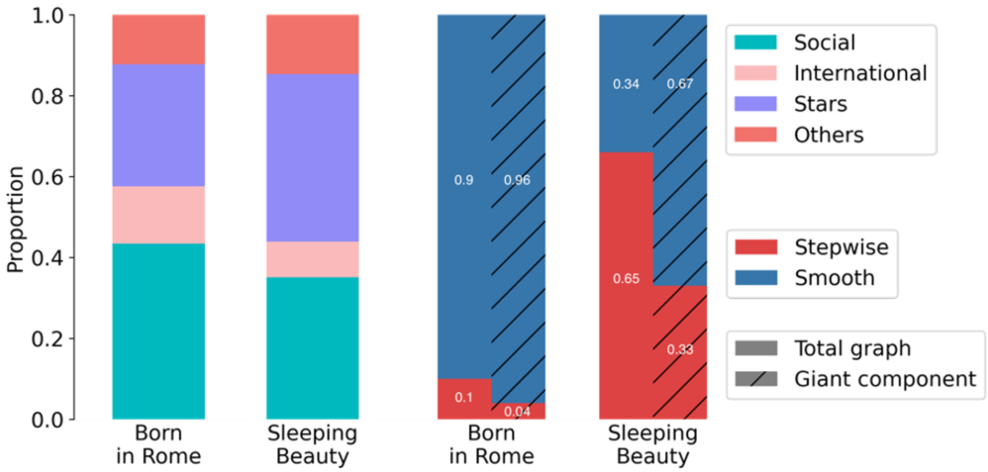

Figure 3.10: Distribution of hashtags in each categories using 8 hours and 31 hours as division choices of “Born in Rome” and “Sleeping Beauty”.

3.5.5 SI5 Distribution of sizes of nodes originating HSL hashtags

As shown in Fig. 3.11, the number of followers of the origin nodes involved in our dataset is ranked in decreasing order. The following table shows the top 20 largest nodes excluding celebrities, which we consider to be the super-hubs.

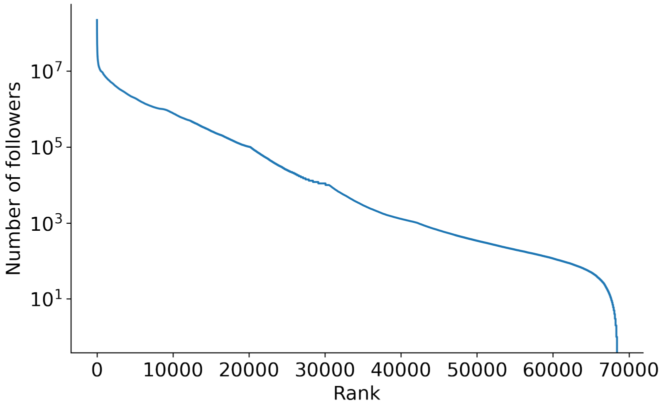

Figure 3.11: Distribution of number of followers of origin nodes. X axis linear scale, y axis logarithmic scale.

Table 3.1: Examples of content categorization.

Social	Stars	International	Others
#Pensions in 31 provinces have all risen# (#31省份养老金已全部上涨#)	#Liang Zhengxian Master of Space Management# (#梁正贤空间管理大师#)	#Employees in White House cafes diagnosed with covid# (#白宫内部咖啡厅员工确诊新冠#)	#If all creatures became cats# (#假如所有生物都变成猫 #)
#76-year-old man rescues 200-pound drowning man# (#76岁老人救起200多斤溺水者#)	#Song Dandan's 60th birthday party# (#宋丹丹60岁生日宴#)	#UAE to fully normalize relations with Israel# (#阿联酋将与以色列实现关系全面正常化#)	#My Youth Pain# (#我的青春疼痛 #)
#Shanxi police cracked down on intentional homicide 30 years ago# (#山西警方破获30年前故意杀人案#)	#Jay Chou's old photos from 20 years ago# (#周杰伦晒20年前旧照#)	#Melbourne, Australia will implement curfew# (#澳大利亚墨尔本将实施宵禁#)	#How to gracefully carry a quilt to school# (#如何优雅地背着被子去学校 #)
#Weibo will regulate the big appetite eating broadcast content# (#微博将整治大胃王吃播内容#)	#BonBon Girls' first EP # (#硬糖少女首张EP#)	#France said it would not ban Huawei's investment in 5G in France# (#法国表态不会禁止华为在法投资5G#)	#The differences in love between the teenage years and now# (#十几岁和现在恋爱的区别 #)
#A large area of mountain collapse occurred on National Highway 347# (#347国道发生大面积山体垮塌#)	#Luhan 6 days in a row shooting fight scene heat stroke# (#鹿晗连续6天拍打戏中暑#)	#U.S. sees largest wave of bankruptcies in a decade# (#美国现十年来最大破产潮#)	#Even the dog can play skateboard# (#连狗子都会玩滑板了 #)
#60 percent of the post-90s dare not speculate in stocks# (#6成90后不敢炒股#)	#Yiyang Xu process is more important than the result# (#徐艺洋过程比结果重要#)	#A tourist resort in Brazil only open to tourists who have been infected with covid# (#巴西一旅游胜地只对曾感染新冠游客开放#)	#Pineapple Mustard Fritter Shrimp# (#菠萝芥末油条虾 #)
#A kindergarten school bus collided with a truck in Shaanxi# (#陕西一幼儿园校车与货车相撞#)	#Street visit to encounter Jackson Wang# (#街访偶遇王嘉尔#)	#Explosion in the Philippines# (#菲律宾发生爆炸#)	#Hogwarts College makeup# (#霍格沃茨学院妆 #)

Table 3.2: Top 20 largest nodes (excluding celebrities) on Sina Weibo

Rank	Username	Number of followers
1	超话社区 (Super Talk Community)	222M
2	人民日报 (People's Daily)	148M
3	央视新闻 (CCTV News)	129M
4	新华社 (Xinhua News Agency)	108M
5	头条新闻 (Headline News)	102M
6	新华网 (Xinhuanet)	95.6M
7	人民网 (People's Net)	82.4M
8	新浪新闻 (Sina News)	77.5M
9	中国新闻网 (China News Net)	75.4M
10	中国日报 (China Daily)	64.9M
11	中国新闻周刊 (China News Weekly)	60.6M
12	微搞笑排行榜 (Weibo Funny List)	55.9M
13	新浪新闻客户端 (Sina News Client)	52.3M
14	环球资讯 (Global Information)	48.9M
15	每日经济新闻 (Daily Economic News)	48M
16	新京报 (Beijing News)	45.8M
17	NBA	42.8M
18	新浪娱乐 (Sina Entertainment)	41.9M
19	微博钱包福利 (Weibo Wallet Benefits)	40.5M
20	新闻晨报 (Morning News)	38.8M

Chapter 4

Popularity competition and identification of interventions on Sina Weibo

4.1 Introduction

Microblogging sites are important vehicles for the users to obtain information and shape public opinion thus they are arenas of continuous competition for popularity. Most popular topics are usually indicated on ranking lists. In this chapter, we investigate the public attention dynamics through the Hot Search List (HSL) of the Chinese microblog Sina Weibo, where trending hashtags are ranked based on a multi-dimensional search volume index. We characterize the rank dynamics by the time spent by hashtags on the list, the time of the day they appear there, the rank diversity, and by the ranking trajectories. We show how the circadian rhythm affects the popularity of hashtags, and observe categories of their rank trajectories by a machine learning classification algorithm. By analyzing patterns of the ranking dynamics we identify anomalies that are likely to result from the platform provider's intervention into the ranking, including the anchoring of hashtags to certain ranks on the HSL. We propose a simple model of ranking that explains the mechanism of this anchoring effect.

4.2 Materials and methods

4.2.1 Data description

Sina Weibo HSL contains the names of the hashtags, their ranks and the search volume hotness which is the base of the ranking (see Eq. (2.1)). We crawled the data from Sina Weibo HSL, with a frequency of $\Delta t = 5$ minutes from 22 May 2020 to 29 September 2020. Since the commercial advertisements randomly occupied the HSL at the third and the sixth ranks, in order to get a constant length of non-advertisement hashtags on the HSL at each timestamp, we removed all the advertisement hashtags which are labeled with "Recommendation (荐)", re-ranked the original HSL and took the top $L = 48$ hashtags for each timestamp, with L being the length of the list. Weibo was punished by the cyberspace authority of China to suspend the update of HSL for one week in June 2020 due to its interference with online communication [89], which causes a gap in the data (see Fig. 4.1). We then did our major analysis based on the data after the punishment. We took all the hashtags that have appeared on the HSL in a two month period from 17 July to 17 September 2020, and crawled all the posts containing these hashtags in their prehistory from birth till first appearance on the HSL. The datasets used in this research are available in a GitHub repository [24].

4.2.2 Ranking dynamics

Measures

A popular hashtag i enters the HSL at time t_i at rank $r_i(t_i)$ ($1 \leq r_i \leq L$) and disappears from it at time T_i. The rank of this hashtag changes with time producing a trajectory $r_i(t)$ until it disappears from the HSL. In order to capture the ranking characteristics of hashtags at different ranks, we use the measure rank diversity [66, 47]. Rank diversity $D(k)$ measures the number of different hashtags at rank $1 \leq k \leq L$ during a given period of time $t_{\min} \leq t \leq t_{\max}$:

$$D(k) = \sum_t \sum_i \delta(k, r_i(t))\phi_{i,k}(t), \tag{4.1}$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta and $\phi_i(t)$ is the indicator, which is 1 if hashtag i has not been at rank k until time t and 0 otherwise.

Rank diversity has been studied extensively. It is known that this quantity is characterized by profiles: For high ranks, their diversity have small values, while the behavior for lower ranks depends on whether the system is closed (only the rank changes but the items do not) or open (when items arrive on and leave from the list). In closed systems the dynamics at low ranks is also suppressed leading to low values of D and a maximum at intermediate ranks, while in open systems rank diversity grows monotonously, as it has been repeatedly observed in empirical data and demonstrated by simple diffusive models [66, 47, 65]. An open system can be considered as a part of a very large closed system.

The duration d_i of a popular hashtag on the HSL measures the time over which it is able to attract consistent public attention:

$$d_i = T_i - t_i. \tag{4.2}$$

The highest rank $r_i^{\min}$ of a hashtag measures its maximum relative ability to attract public attention during its whole lifetime on the HSL:

$$r_i^{\min} = \min_{t \in [t_i, T_i]} r_i(t). \tag{4.3}$$

Categorization of rank trajectories

The rank trajectory $r_i(t)$ is uniquely defined for $\forall$ hashtag i. Some hashtags have short lifetime on the HSL, others can attract popularity for a longer period of time; some go rapidly to high ranks, others never reach that level. Are there similarities between different shapes of the trajectories and can they be ordered into categories? Here we use machine learning techniques to find characteristic patterns in these rank trajectories. In order to deal with rank time series of different lengths, we use Dynamic Time Warping (DTW) [68] as a similarity measure between two time series. DTW computes the best possible alignment between two time series. Then we use k-means clustering to find clusters of characteristic shapes. The computation was done using python tslearn package [93].

4.3 Results

4.3.1 Circadian patterns

Human actions are largely influenced by the circadian rhythm and so are online activities. Figure 4.1A shows the increment of the number of hashtags per $\Delta t = 5$ minutes interval clearly demonstrating the cyclic structure during the observation period from 22 May 2020 to 29 September 2020, except for a short interruption in June 2020. Similarly in Fig. 4.1B, the median search volume index of hashtags on the HSL at a timestamp rises and decays in a periodic fashion. The missing of data for one-week in June 2020 is observed in both Fig. 4.1A and Fig. 4.1B, which results from the suspension of HSL by the cyberspace authority of China due to Weibo's interference with online communication [89].

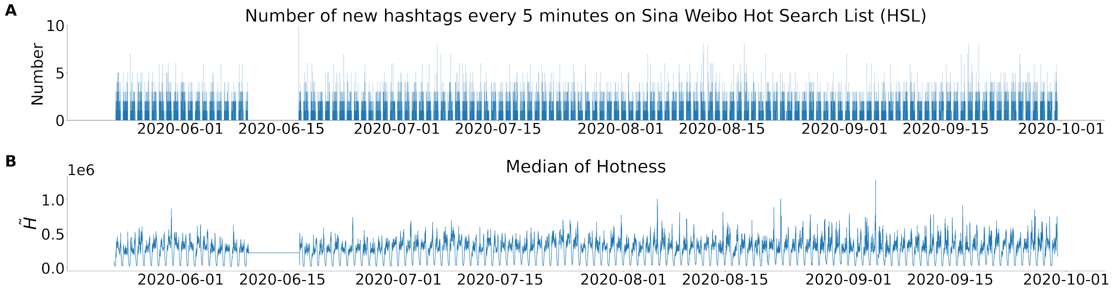

Figure 4.1: Circadian patterns of the Sina Weibo Hot Search List (HSL). (A) Increment of number of new hashtags per $\Delta t = 5$ minutes on the HSL during the observation period from 22 May 2020 to 29 September 2020. (B) Time series of the median of search volume index of all hashtags on the HSL at a timestamp, advertisement rank positions excluded. $\tilde{H}$ represents the median value hotness H of hashtags on Sina Weibo HSL at a timestamp. In both (A) and (B) the one-week gap due to the suspension of HSL by the cyberspace authority of China is visible.

4.3.2 Rank trajectory clustering

A successful hashtag i stays on the HSL between the time instants t_i, when it appears on the list, until T_i, when it finally disappears from it defining the duration $d_i = T_i - t_i$. Some hashtags stay on the list for very short time ($d_i < 10$ minutes), while some others stay for many hours. The rank of a hashtag i follows a trajectory $r_i(t)$. Some hashtags' trajectories go first up and then down, some go up and down and up again, there are also cases that hashtag's trajectory goes up and then it disappears. Also, the speed change of the trajectories is variable, resulting in a multitude of shapes of rank trajectories.

The duration distribution of hashtags in the observation period is shown in Fig. 4.2A. We observe a sharp peak for hashtags with short duration and two less pronounced peaks, where the latter are characterized by similar patterns of the ranking trajectories. The vertical red line at the local minimum of 1 hour separates the duration distribution into two sections, section 1 and 2, respectively. The individual rows in Fig. 4.2 correspond to the clustering of the rank trajectories on each of the separated sections: Section 1 (B,C,D) and Section 2 (E,F,G). Even for hashtags with short duration on the HSL (Section 1) it is worth categorizing the rank trajectories. In most cases the rank does not change much during the lifetime d_i (see Figs. 4.2B and C) and remains low, however, as shown in Fig. 4.2D, some ranks of the hashtags exhibit a clear directional motion: they go to lower rank numbers, i.e., to higher ranks and disappear from there. For the more expected rank trajectories shown from Fig. 4.2E to Fig. 4.2G, we also see some recognizable differences. Rank trajectories in Fig. 4.2E first go up and quickly go down after hitting the top, without staying at a certain rank for a long time. Rank trajectories in Fig. 4.2F first go up, stay stable around the highest ranks with little fluctuation for a long time and then go down. Rank trajectories in Fig. 4.2G first go up, with more fluctuations but never surpass the previous peak, then stay stable for a long time and finally go down. In the next Section we will show how the rank trajectory shapes are related to the time of the day the hashtags first appear on the HSL.

4.3.3 Anomaly detection

The dynamics of popularity as captured in the HSL should be sensitive to the actual trends and reflect the users' overall activity patterns. The individual rank trajectories show fluctuations but after averaging one would expect smooth behaviors. However, when studying the characteristics of the hashtags' rank dynamics on HSL, like the rank diversity or duration distribution we bumped into strange behaviors which we interpret as indications of interventions by the service provider.

4.3.4 Duration

Figure 4.3A is the d_i vs $t_i (\mathrm{mod}\ 24\mathrm{h})$ scatter plot, i.e., it shows the durations of the hashtags vs the times of the day when they first appeared on the HSL, with each point representing a hashtag. Hashtags tend to appear on the HSL starting from around 7 a.m. till midnight. We can see clear shapes of lower-left and upper-right triangles, separated by a stripe in the middle with a low number of points inside. The lower boundary of the upper-right triangle is very sharp, while the upper boundary of the lower-left triangle is less so. There are data points within the stripe, but the density is much less compared with the data points inside the triangles and also if we compare it to the users' overall activity pattern (see SI). The vertical distance between the triangle boundary lines is approximately 7 hours. The existence of these triangles suggests that the hashtags, which enter the HSL after 15 p.m. tend to either disappear from the HSL on the same day or stay on the HSL during the night and disappear after 7 a.m. the next day. This is presumably related to

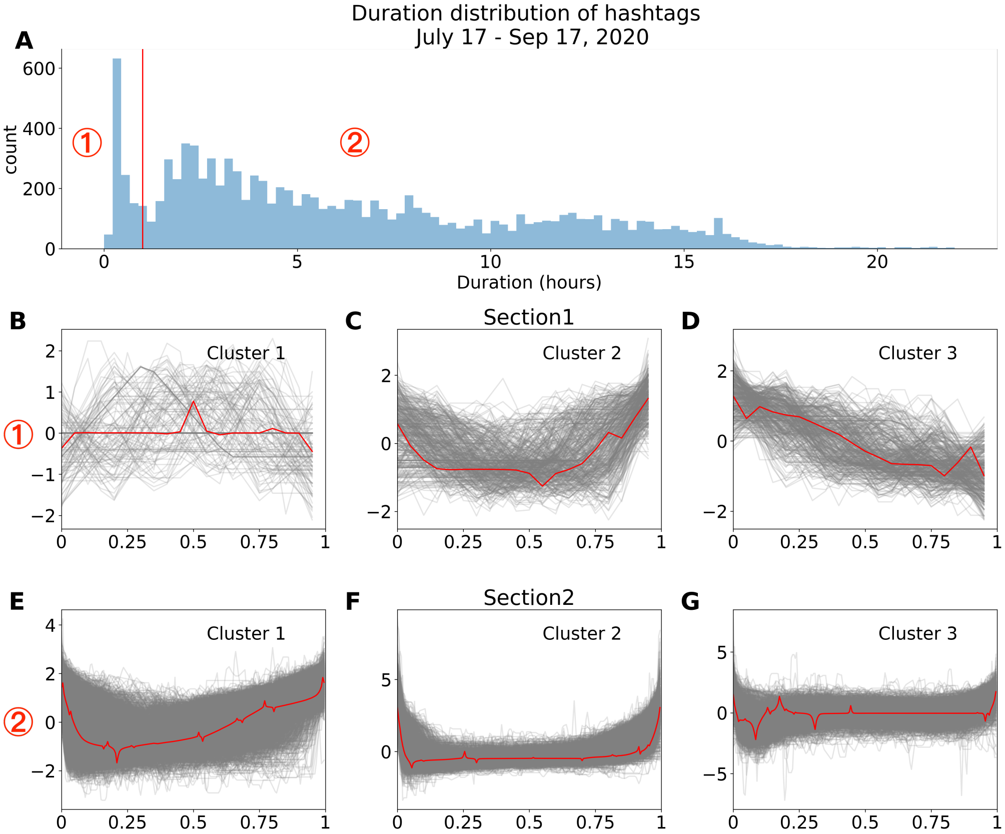

Figure 4.2: Clustering patterns of hashtag rank trajectories on the Sina Weibo HSL. (A) Distribution of hashtag duration on the HSL, divided into two sections based on local minima at 1 hour. Results of k-means clustering with 3 clusters in each section for time series data are shown, metric is dtw (dynamic time warping) distance, y-axis is normalized to the mean and the standard deviation and the x-axis by d_i. (B), (C), (D) correspond to duration interval from 0 to 1 hour (Section 1). (E), (F), (G) correspond to duration interval larger than 1 hour (Section 2). Red curves depict clustering centers in each category.

Sina Weibo working mode, already pointed out in previous studies [26], namely that Sina Weibo practically stops working between midnight and 7 a.m. If the ranking was automated following the formula Eq. 2.1, the changes from day to night should not be that sharp and the circadian pattern should follow more or less that of the people's activity.

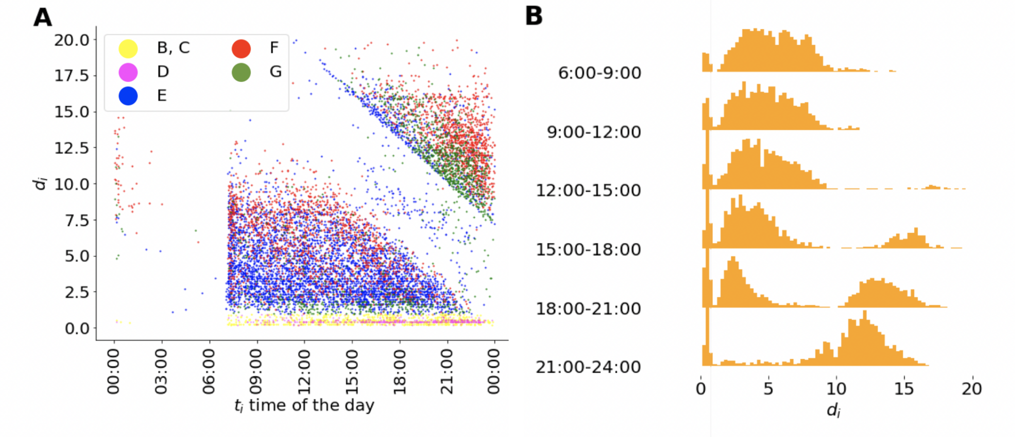

Figure 4.3: Relationship between hashtags' duration on the HSL and the time t_i. (A) Scatter plot of hashtags' duration on the HSL and the time of the day they first appear on the HSL. Each point is a hashtag, colored by the category it is clustered in Fig. 4.2. (B) Distribution of hashtags' duration on the HSL according to different time intervals during the day of first appearance on HSL.

Figure 4.3B shows the duration distribution of the hashtags as a decomposition of Fig. 4.2A by binned starting values of the times of the day. For each time interval, the observed distribution is trimodal. As the start time of the day $t_i (\mathrm{mod}\ 24\mathrm{h})$ goes on, the density of hashtags in the third mode is increasing. In Fig. 4.3A we see a low-density area at around 1 hour duration between the blue and the yellow dots, which corresponds to the minimum between sections 1 and 2 in Fig. 4.2A. Accordingly, in the duration distribution plot shown in Fig. 4.3B, a peak is observed for hashtags with duration shorter than 1 hour. Within this stripe there is an accumulation of pink dots corresponding to trajectories of category D, with a unique shape, namely starting at low rank and ending at a high one within a short period of time. The fact that the hashtags disappear from the HSL during their rising trend toward higher ranks might be related to platform interventions. In most other cases the more expected shape is observed, namely starting and ending from low rank and having in between some higher rank.

How are the shapes of the rank trajectories related to the time of the day the hashtags first appear on the HSL? Recall the Weibo working mode, if a hashtag's stay on the HSL is influenced by the night break, then it will automatically have a little-fluctuation period of at least seven hours, resulting in a rank trajectory shape similar to Fig. 4.2F or the last part of Fig. 4.2G, which we color in red and green respectively in Fig. 4.3A. Hashtags in Fig. 4.2F are born closer to midnight and further away from the hypotenuse of the upper-right triangle in Fig. 4.3A. This is reasonable since hashtags entering HSL close to midnight are likely exposed to the stay on the HSL during the night break. Hashtags with shape in Fig. 4.2G, however, are close to the hypotenuse boundary of the upper-right triangle in Fig. 4.3A. One possible explanation is that although these hashtags' attention level is already in decreasing trend, their stay on the HSL are prolonged by the night break, so that when the next day begins, they are replaced by new hashtags and leave the list. The majority of hashtags with shape shown in Fig. 4.2E are of shorter duration, located in the dense area of the lower-left triangle colored in blue in Fig. 4.3A. The separation of the red and blue areas in Fig. 4.3A lower-left triangle tells that hashtags which quickly go down after reaching their highest ranks on the HSL lack the ability to consistently attract public attention to maintain their positions on the list. In contrast, hashtags maintain relatively stable ranks (Fig. 4.2F) stay longer times on the HSL, as Fig. 4.3A lower-left red area suggests.

4.3.5 Ranking

As mentioned in Section 4.2, spontaneously evolving ranking dynamics have typical rank diversity patterns [66, 47, 65]. After sufficiently averaging the rank diversity, the curve shape is smoothened and depends on whether the system is open or closed, as shown in SI.

Figure 4.4A shows the distribution of the enter-ranks $r_i(t_i)$ and leave-ranks $r_i(T_i)$ on the HSL. The majority of hashtags do not land on the HSL from the very bottom of the ranking list, instead they tend to enter at ranks 44 - 46 while they tend to leave from the bottom ranks. Figure 4.4B shows the scatter plot of the highest rank of the hashtags and their duration on the HSL. The duration exhibits a bimodal pattern with a sudden jump at rank 16, and then it decreases. Figure 4.4C shows the relationship between the hashtags' enter-ranks on the HSL and their corresponding duration on the HSL. The popularity of a hashtag is reflected in its rank position and the duration it stays on the HSL. Hashtags at higher ranks are more stable and stay for longer hours on the HSL as shown by the rank diversity in Fig. 4.6A, thus it is strange for hashtags entering HSL at a high rank only stays for short duration.

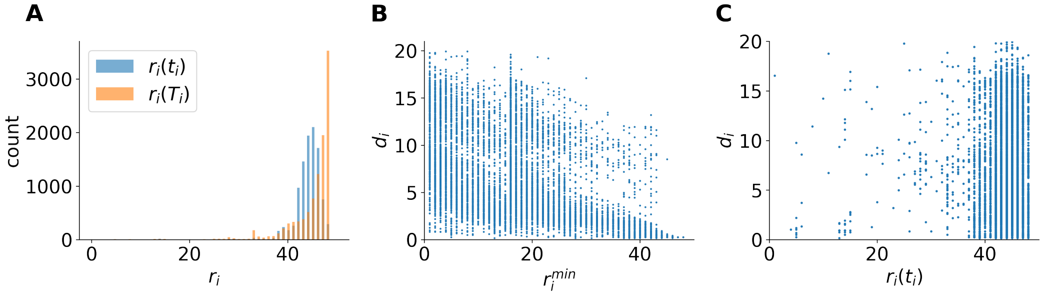

Figure 4.4: Ranking dynamics characterization of hashtags on the Sina Weibo HSL from 17 July 2020 to 17 Sep 2020. (A) Distribution of $r_i(t_i)$ and $r_i(T_i)$. (B) Scatter plot of r_i^{min} and d_i. (C) Scatter plot of $r_i(t_i)$ and d_i.

Figure 4.6A shows the rank diversity of the hashtags on the HSL broken down to daytime and nighttime. The difference between the behavior during the night and day is apparent: The former is more likely to the closed systems' characteristics with reduced activity while the latter is closer to the open systems' features although the trend around rank 44 turns down, probably due to the fact that the hashtags' enter-ranks $r_i(t_i)$ is shifted to the left as shown in Fig. 4.4A. There are apparent anomalies in these figures at certain ranks where rank diversity values are dropping systematically as compared to what is expected from the assumption of a smooth curve for these quantities. We think that the anomalies are due to interventions by the service provider, which anchors some of the hashtags on the HSL at specific ranking positions (see Section 4.3.7 Anchor effect).

4.3.6 Ranking dynamics in relation to prehistory

Before the hashtags gain enough popularity and land on the HSL, they go through different propagation routes during their prehistory. The time length of the prehistories t_{HSL} differ for different hashtags [26]. Some hashtags get to the HSL in very short time after birth, while others take longer. Figure 4.5 shows the relationship between t_{HSL} of the hashtags, the ranks they enter the HSL $r_i(t_i)$, and the duration d_i of their stay on the HSL. As shown in Fig. 4.5A, in accordance with Fig. 4.4A, the majority of hashtags enter the HSL at a low rank peaking around 45. Some hashtags enter the HSL at higher ranks, however, as the prehistory gets longer, the chance the hashtag enters the HSL from a high rank is less likely. As for the properties of hashtag duration on the HSL shown in Fig. 4.5B and Fig. 4.5C, the duration against prehistory length exhibits bimodal distribution. As the prehistory length increases, the first peak drops and the second peak rises. The bimodality similar to results shown in Fig. 4.3, is influenced by the Weibo circadian working mode.

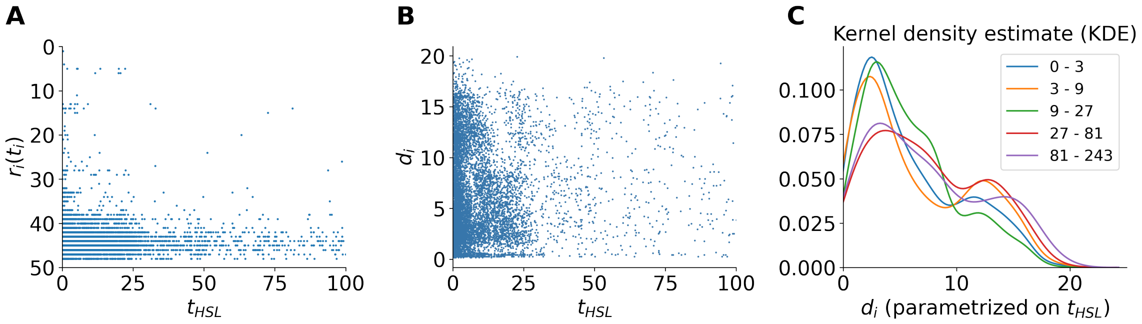

Figure 4.5: Prehistory length t_{HSL}, enter-ranks $r_i(t_i)$, and duration d_i of hashtags on the Sina Weibo HSL. (A) The relationship between the hashtags' prehistory time length and the ranks they first enter on the HSL. (B) The relationship between the hashtags' prehistory time length and the duration they stay on the HSL. (C) Parameterized probability density function of the hashtag duration on the HSL by prehistory time length, using kernel density estimation (KDE) [95], with the parameter bw = "scott" [82].

4.3.7 Anchor effect

In this section, we propose a ranking model with anchoring to simulate the dynamics of the hashtag ranking anomalies on the HSL.

Let us take a system of N elements (for the hashtags), each element has a random initial score of values within $(0, 1)$, and rank these elements from top to bottom based on the scores. Let r_i and s_i denote the rank and the score of the i-th element, respectively. The scores change in time and that causes the rank movement of the elements. We will choose a simple dynamics: An element is randomly selected, 1 is added to its score and the ranking is changed if necessary. The idea of the anchor is the following: Set an anchor at position A. For hashtags whose $r < A$, it is difficult to go down the ranking list; for hashtags whose $r > A$, it is difficult to go up (note that high rank means low r value). The anchor represents a barrier characterized by an increment δ. Let $\varphi(r_i) = i$ denote the selection of the element at a given rank at an instant of time.

The procedure of ranking at each step is shown below. Randomly pick one element j and $s_j^{\text{new}} = s_j + 1$. There are three possibilities:
(a) $r_j < A$. Update the top $A - 1$ ranks, no change of the anchor element.

(b) $r_j = A$. If $k = \varphi(A-1)$ and $s_j^{\text{new}} > s_k + \delta$, update the top A rank. Otherwise, no change of ranks.
(c) $r_j > A$. If $\ell = \varphi(A)$ and $s_j^{\text{new}} > s_\ell + \delta$, old anchor rank drops to $A+1$, update the top $A+1$ ranks. Otherwise, update ranks lower than A, no change of the anchor element.

We simulate a system with 500 elements and take the top $L = 48$ ranks to approximate an open system.

The rank diversity of a non-intervened system has parabola-shape (see Supplementary Information). The intervention produces a deep valley at the anchor position, very similar to those observed in the measured curves in Fig. 4.6, which shows the comparison between the real data and our model with anchoring. Figure 4.6A shows the average rank diversity of the observation period, separately for day and night. The measures during the day from 7 a.m to 24 p.m has larger value than during the night from 24 p.m to 7 a.m. and the night behavior is closer to that of a closed system, according to the suppressed activity during that period. At certain positions (ranks 16, 28, and 33) there are large drops in the values of the function, indicating intervention by "anchoring" hashtags at these specific ranks. The simple model of anchoring reproduces qualitatively the effect.

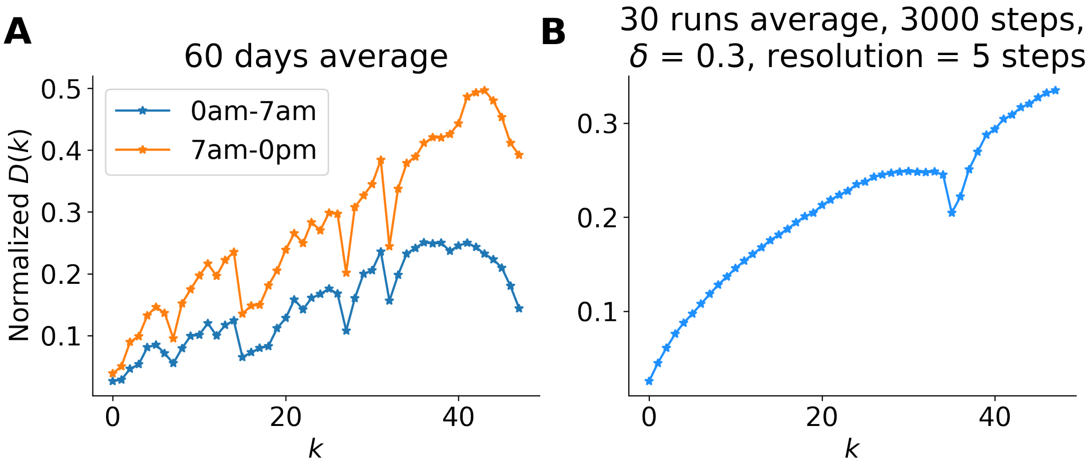

Figure 4.6: Rank dynamics comparison between empirical data and a ranking model with anchoring.

With this model we support the assumption that the observed anomalies in the ranking functions are due to intervention.

4.4 Discussion

Public attention is precious and it is nowadays largely dependent on online social media, therefore it is of great interest to understand the dynamics governing popularity on such platforms. Considerable effort has been devoted to this task on Twitter [32, 7, 8, 117] and some results are also known on Sina Weibo [35, 126, 26]. In order to attract attention, people, companies, and political actors are tempted to make use of hidden manipulations besides well known tools of direct advertisements or propaganda [132, 90, 9, 25, 26]. Thus popularity can emerge spontaneously via collective attention from online users who are genuinely interested in a topic and form trends, quantified and captured by the algorithm of the platform, or trends emerge from intervention by the platform provider motivated by financial or other interests. (It should be noted that "collective attention" may also be influenced, e.g., by spamming [90] or coordinated campaigns [126, 74].)

In this chapter, we studied the attention dynamics of trending hashtags on the Sina Weibo Hot Search List by using various measures of ranking dynamics, like entering and leaving ranks as well as duration of hashtags on HSL, rank diversity, and categories of rank trajectories. The aim of the identification of regularities in the ranking dynamics was twofold: First, contribution to the quantitative characterization of the dynamics of public attention in order to better understand its mechanism, and second, finding signatures of interventions by the service provider.

The duration of the hashtags on the HSL in relation to the time of the day they enter the list shows trimodality (Fig. 4.3). This is related to the fact that the appearance of hashtags on the HSL have circadian patterns (Fig. 4.1A). On the one hand, the pattern is caused by the circadian rhythm of the users whose activities depend on the time of the day (see SI), on the other hand it is imposed by the apparent working mode of Sina Weibo, which reduces the night-time flow of new hashtags to the HSL almost to zero level. The night break is reflected in the very low number of points in the stripe separating the two triangles in Fig. 4.3A and in the particularly sharp upper boundary of this stripe. This seven hour gap has been shown to influence the prehistory of the successful hashtags [26] by contributing to the difference between shorter and longer prehistories and it creates a link between the behavior of the hashtags on the HSL and their prehistory (Fig. 4.5).

The observation of the extreme daily pattern is already an example that we are able to identify interventions by the service provider through anomalous behavior of the ranking dynamics. Obviously, the ranking is not automated following a plain formula like Eq. 2.1 but depends on human control. Even more importantly, we show an anchoring effect at some rank positions on the HSL, where rank diversity is suppressed as compared to the expected smooth behavior of this quantity. Using a simple ranking model we show how anchoring at some rank positions changes rank diversity. A further observation indicating intervention is that some hashtags on the HSL appear at high ranks and disappear in short time (Fig. 4.4C). Similarly, there are many hashtags that just stay on the HSL for short time which is shown in the first peak in Fig. 4.3B. The fact that the peak is separated from the rest of the distribution is also likely to be related to intervention.

Sina Weibo is the microblogging site with world-wide the largest number of active users, who are overwhelmingly Chinese speakers. While we believe that alone the size of Sina Weibo justifies focused study, we know that most of our results are idiosyncratic. However, this is true only in a narrow sense as our results provide general lessons. We demonstrated that studying the ranking dynamics in popularity lists is worth for several reasons. First, we uncovered relationships between ranking dynamics and the circadian pattern of user activity, also establishing a link to the prehistory of items getting to the ranking list. Moreover, we identified different trajectory categories on the list, which characterize different dynamic patterns of popularity. Finally, and most importantly, we showed, how pinpointing anomalies in ranking statistics can be used to identify interventions by the service provider. As service providers have financial interests and may be under political pressure, the objectivity of the ranking lists and their truth content can be questioned. Similarly to the fight against fake news, the fight against manipulation of public attention is in the interest of the society and it also needs the tools of detecting interventions.

4.5 Supplementary information

4.5.1 SI1 Daily posts volume on Weibo

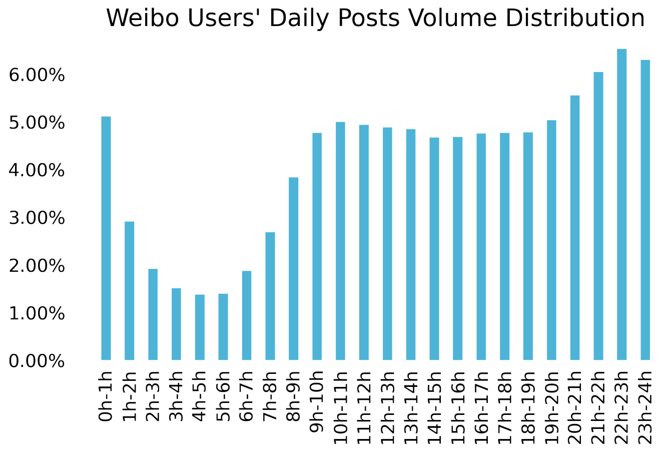

Figure 4.7: Distribution of Weibo users' daily posts volume according to Weibo User Development Report [14].

4.5.2 SI2 Examples of rank trajectories

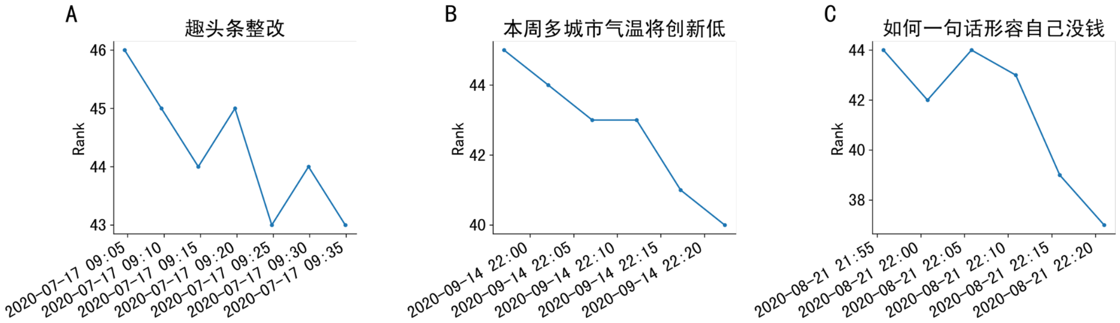

Figure 4.8: Examples of rank trajectories in Fig. 4.2D. (A) #Qutoutiao rectification# (#趣头条整改#) (B) (#Temperatures set to hit record lows in many cities this week#) #本周多城市气温将创新低# (C) #How to describe yourself having no money in one sentence# (#如何一句话形容自己没钱#)

4.5.3 SI3 Rank diversity in a model closed system

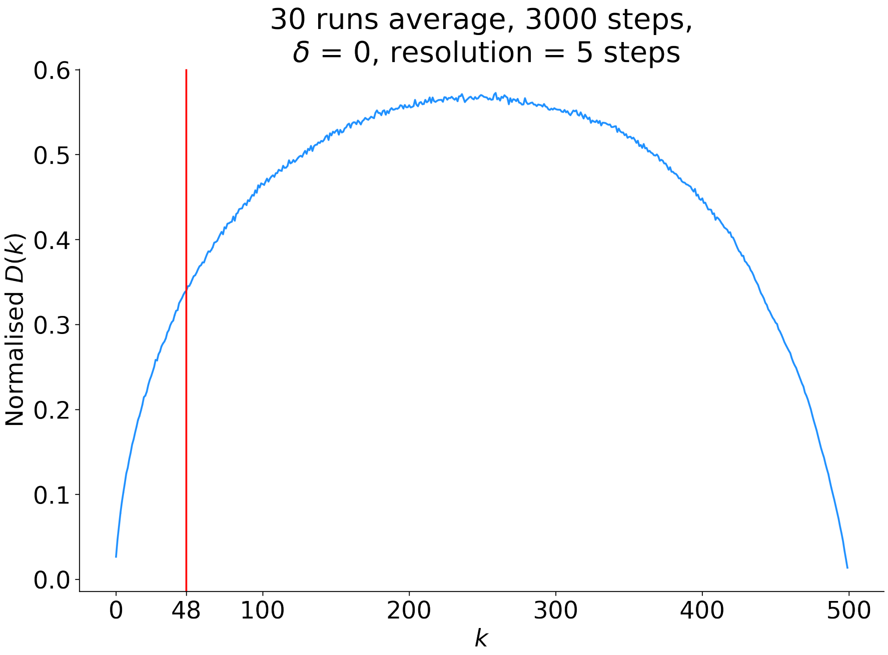

Figure 4.9: Parabola shaped normalized rank diversity in a model closed system of size 500. The top L = 48 can be considered as an open system.

Chapter 5

Attention dynamics on Sina Weibo during the COVID-19 pandemic

5.1 Introduction

Understanding attention dynamics on social media during pandemics could help governments minimize the effects. In this chapter, we focus on how COVID-19 has influenced the attention dynamics on the biggest Chinese microblogging website Sina Weibo during the first four months of the pandemic. We study the real-time Hot Search List (HSL), which provides the ranking of the most popular 50 hashtags based on the amount of Sina Weibo searches. We show how the specific events, measures and developments during the epidemic affected the emergence of different kinds of hashtags and the ranking on the HSL. We also pose the so far less studied question about the temporal correlation patterns between different COVID-related topic categories with the changing world situations of COVID-19. A significant increase of COVID-19 related hashtags started to occur on HSL around January 20, 2020, when the transmission of the disease between humans was announced. Then very rapidly a situation was reached where COVID-related hashtags occupied 30-70% of the HSL, however, with changing content. We give an analysis of how the hashtag topics changed during the investigated time span and conclude that there are three periods separated by February 12 and March 12. In period 1, we see strong topical correlations and clustering of hashtags; in period 2, the correlations are weakened, without clustering pattern; in period 3, we see a potential of clustering while not as strong as in period 1. We further explore the dynamics of HSL by measuring the ranking dynamics and the lifetimes of hashtags on the list. This way we can obtain information about the decay of attention, which is important for decisions about the temporal placement of governmental measures to achieve permanent awareness. Furthermore, our observations indicate abnormally higher rank diversity in the top 15 ranks on HSL due to the COVID-19 related hashtags, revealing the possibility of algorithmic intervention from the platform provider.

The HSL of Sina Weibo during the period of COVID-19 provides rich data about public attention and its dynamics in China. Based on that data, we have been able to track the evolution of public attention in different periods during the pandemic, follow how the attention of the population shifted from one group of topics to another and study the changing correlation patterns of different COVID-related topic categories with the evolving COVID-19 situation in the world. During our studies we have discovered signatures of the possible algorithmic control on this social media platform. Understanding the dynamics of public attention on social media promotes instant and effective communications among governments, health experts and the public, helping the government to monitor public opinion, maintain the stability of the society as well as develop and deliver more effective measures to minimize the effects of the pandemic.

5.2 Materials and methods

5.2.1 Data description

We took data from Weibo HSL to study attention dynamics as it captures vibrant real-time change of public attention. Due to the random existence of one or two commercial advertisements at the third and the sixth ranks, in order to get a constant length of non-advertisement hashtags on the HSL at each timestamp, we removed all the hashtags labeled with "荐", re-ranked the original HSL and took the top 48 hashtags for each timestamp. All the HSL we mentioned later in this chapter mean the re-ranked HSL with 48 ranks. We directly downloaded the data from the HSL with a frequency of every 5 minutes from December 16, 2019 to April 17, 2020. There are in total 26022 hashtags and 9120 of them are related to the aspects of COVID-19. To relate social media contents with real-life pandemic situation in Mainland China, we collected the daily number of infections, deaths, and recoveries from the official website of National Health Commission of China [70]. In the following subsection we explain how we identified the different categories of hashtags. The datasets supporting the conclusions of this chapter are available in a GitHub repository [22].

5.2.2 Topical classification and correlations

Fig. 5.1 shows the number of daily infections, deaths and recoveries in Mainland China. The number of daily infections and deaths have a sharp peak on February 12 due to the adoption of new diagnostic criteria [92]. The decreasing trend of daily infections since the peak turned to increasing after March 13, as a result of the rising number of imported coronavirus cases from abroad [81]. We will argue that there are three periods to be distinguished after the outburst of COVID-19 around January 19, separated by the maximum and local minimum of the daily number of infections on February 12 and March 12, respectively.

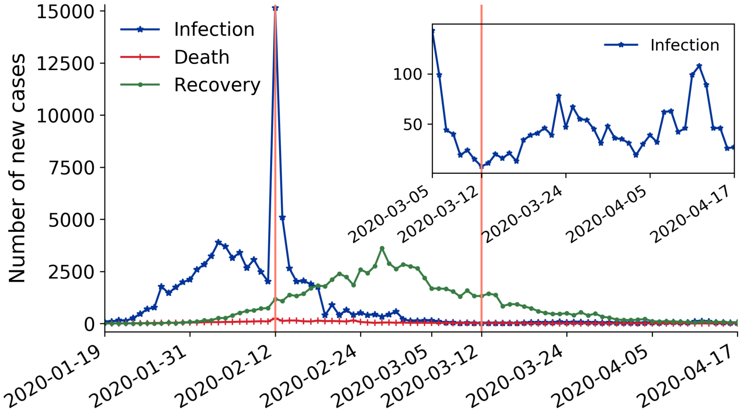

Figure 5.1: COVID-19 daily infection, death and recovery in Mainland China. The inset enlarges the tail of the infection curve. Three periods after the outbreak on January 19 are separated by the highlighted peak and local minimum.

To study the public attention towards COVID-related information, we first extracted hashtags which encompass all aspects of COVID-19 and classified them based on geographic regions and the exposure order under the pandemic into three categories: Mainland China, East Asia outside of Mainland China and Other Countries outside of East Asia. With a focus on COVID-hashtags related to Mainland China, we manually classified them based on semantic meanings into the following seven disjoint sub-categories. The Bad News category comprises hashtags on confirmed infections and deaths in different regions of

Mainland China as well as shortages of essential supplies. The Good News category consists of news on cases of recovery, sufficiency of supplies, and decrease in daily infections or deaths. The Regulations category consists of authority announcements of national, regional, institutional laws, rules and regulations associated with public behaviors and concerns during the pandemic. The Life Influence category contains hashtags that reflect the pandemic influence on the aspects of citizen lives. The Front Lines category includes hashtags related to the lives of front line workers (mainly doctors and nurses) and their interactions with patients in hospitals. The Science category incorporates scientific understandings of the virus properties, vaccine development, and ways for public protection given by authoritative doctors. The Supports category takes into account hashtags on worldwide donations and emotional supports. For ambiguous cases which contain information of more than one category, our classifications were based on the focus of the main subject. Due to the syntactic-semantic complexity of Chinese language, the classifications were made by two independent annotators. Final consensus was reached in case of disagreement. The Mainland China sub-categories are summarized in Table 5.1 together with examples. The full list of COVID-related hashtags is available in the dataset, which we have made public [22].

To further understand how the Mainland China related COVID-hashtags are correlated with each other and with the daily number of infections/deaths/recoveries in the three separated time periods, we measured the Pearson's correlations between the seven series of daily number of new hashtags in each of the sub-categories defined above, together with the three series of daily number of infections/deaths/recoveries. The correlation of these ten time series are calculated using the percentage change between the current and the prior element instead of the actual value in order to reduce the effect of the trend which can cause spurious correlations. For time series category $X = \{X_{t_i} : t_i \in T, i = 1, 2, ...n\}$ and category $Y = \{Y_{t_i} : t_i \in T, i = 1, 2, ...n\}$, where T is the time index set, the Pearson's correlation is calculated using the percentage change series $\tilde{X} = \{\frac{X_{t_{i+1}} - X_{t_i}}{X_{t_i}}, t_i \in T, i = 1, 2, ...n\}$ and $\tilde{Y} = \{\frac{Y_{t_{i+1}} - Y_{t_i}}{Y_{t_i}}, t_i \in T, i = 1, 2, ...n\}$.

5.2.3 Attention dynamics

One natural measure of social media attention towards a topic category is the quantity of the related hashtags. The growing pattern of the cumulative number of hashtags on the HSL with time reflects the dynamics of the public attention. We separately measured the growth of the cumulative number of all hashtags and all COVID-related hashtags that ever appeared on the HSL in our observation period. To understand how much COVID-

Table 5.1: Mainland China COVID-hashtag details. A summary of the example hashtags in each sub-category of Mainland China category and the number of hashtags in different time periods.

Category	Examples	Period 1	Period 2	Period 3
Bad News	#全国累计确诊新冠肺炎66492例# (#National cumulative confirmed COVID-19 cases reach 66492#) #黑龙江聚集性疫情共48起发病194人# (#Heilongjiang in total 48 clustered epidemic 194 infected cases#) #武汉多家医院物资紧张# (#Wuhan many hospitals supplies in shortage#)	451	193	272
Good News	#火神山医院累计治愈患者破千# (#Huoshenshan Hospital has cured over a thousand patients#) #7省区现有确诊病例清零# (#7 provinces current infected cases down to zero#) #疫情形势出现3个积极变化# (#Epidemic situation shows 3 positive changes#)	145	257	121
Regulations	#上海地铁不戴口罩不得进站# (#Shanghai metro station not allowed to enter without wearing a mask#) #疫情影响严重的地区可增发生活补助# (#Additional living allowances can be issued in areas severely affected by the epidemic#) #非疫情严重国家进京者居家观察14天# (#Home observation for 14 days for visitors to enter Beijing from non-severe epidemic countries#)	318	325	633
Life Influence	#武汉市民江滩唱起国歌# (#Wuhan citizens sing national anthem at the River Beach#) #疫情期间点外卖指南# (#Guide to ordering takeout during the epidemic#) #一季度民航业亏损398亿# (#Civil aviation industry suffered a loss of 39.8 billion in the first quarter#)	310	371	649
Front Lines	#钟南山等专家连线武汉ICU团队# (#Zhong Nanshan and other experts connected to the Wuhan ICU team#) #护士握手呼唤79岁新冠患者# (#Nurse shakes hands and calls 79-year-old COVID-19 patient#) #方舱医院收治第一批患者现场# (#Site of Fangcang shelter hospital taking the first batch of patients#)	251	329	347
Science	#各年龄段人群普遍易感新冠病毒# (#People of all ages are generally susceptible to coronavirus#) #口罩的正确使用方法# (#The correct use of masks#) #如何区分感冒流感和新冠肺炎# (#How to distinguish between flu and COVID-19#)	180	170	123
Supports	#汶川村民自发支援武汉100吨蔬菜# (#Wenchuan villagers spontaneously support Wuhan 100 tons of vegetables#) #欧盟对华运送12吨急需物资# (#EU sends 12 tons of urgently needed supplies to China#) #武汉给援汉医疗队全员的感谢信# (#Thank you letter from Wuhan to all members of the medical aid team#)	151	144	116

information occupies the HSL at each timestamp, we constructed the historical ratio trajectory of the COVID-related hashtags on the HSL since the first COVID-hashtag #武汉发现不明原因肺炎# (#Pneumonia of unknown cause detected in Wuhan#) appeared on December 31, 2019.

Lifetime duration

The lifetime duration of a hashtag on the HSL indicates the ability of obtaining persistent attention from the public. We quantified the duration (continuous existence on the HSL) of a hashtag with τ:

$$\tau = \tau_1 - \tau_0,$$

where τ_0 is the timestamp of the first and τ_1 is the timestamp of the last appearance of a hashtag on the HSL.

We compared the duration of the hashtags across various categories and different time scopes. We compared the duration of the hashtags before the outbreak on January 19, all COVID-related hashtags, and non-COVID hashtags after the outbreak. To ensure complete life cycles of the hashtags, we took all hashtags whose first arrivals on the HSL are between December 19, 2019 and January 18, 2020 as the sample for hashtags before the pandemic, which includes 6161 in total. Similarly, we took all COVID-hashtags whose first arrivals are no later than April 14, with a total number of 8808. For the non-COVID hashtags after the outbreak, we took a random sample of all non-COVID hashtags with the same size as the COVID sample. Hashtags that reappeared after disappearing from the HSL were excluded from our calculation. To understand the overall attention variation towards COVID-hashtags with time, we investigated the daily value of their cumulative average duration. We denote D_j as the cumulative average of duration from December 31, 2019 (day 0) until day j. D_j is calculated as follows:

$$D_j = \frac{1}{|S(j)|} \sum_{i=0}^{j} \sum_{\alpha \in S(j)} d_i^\alpha \tag{5.1}$$

where d_i^α is the duration of hashtag α whose first appearance was on day i. $S(j)$ is the set of all the hashtags whose first appearance is in the interval $[0, j]$.

Ranking

The changes in the ranking patterns of the hashtags at different time periods reflect the general public attention dynamics. Rank diversity $d(k)$ [66] is defined as the number of distinct elements in a complex system that occupy the rank k at some point during a given length of time. Rank diversity is known to give characteristic profiles for different types of systems; e.g., in open systems (where only the top part of the competing items is ranked) behaves differently from closed systems (where all the items are ranked) [66, 67]. In this chapter, we use rank diversity to measure the number of different hashtags occupying a given rank on the HSL over a given length of time, and thus obtain overall information on the total dynamical trend of the hashtags on the HSL. We normalize the rank diversity value by the total number of unique hashtags that have appeared on the HSL in a given time interval. We compared the rank diversity in the 48 ranks on the HSL before the outbreak and during the different periods after the outbreak, with and without COVID-19 hashtags.

The public attention towards a hashtag can also be indicated by its highest rank during the lifetime on the HSL. The highest rank of a hashtag reveals its highest ability and achievement when competing for attention with the other hashtags. We studied the highest rank distribution of the classified COVID-hashtags and compared the results with the hashtags before the outbreak as well as the non-COVID hashtags after the outbreak (SI). To understand the overall highest rank variation towards COVID-hashtags with time, we investigated the daily value trajectory of their cumulative average highest rank. We denote H_j as the cumulative average of highest rank from December 31 (day 0), 2019 until day j. H_j is calculated as follows:

$$H_j = \frac{1}{|S(j)|} \sum_{i=0}^{j} \sum_{\alpha \in S(j)} h_i^{\alpha} \tag{5.2}$$

where h_i^{α} is the highest rank of hashtag α whose first appearance was on day i. $S(j)$ is the set of all the hashtags whose first appearance was in the interval $[0, j]$.

5.3 Results

Statistics and categorization of hashtags

The cumulative number of new hashtags on HSL grows approximately linearly (see Fig. 5.2 (A)), indicating a nearly constant attention capacity and need for news of the users.

Closer inspection tells, however, that the rate of new hashtags decreases between January 10 and February 12 followed by an increased rate until March 28 after which the original slope of 225 ± 4 new hashtags/day sets in. We attribute this change in the slope to the effect of COVID-related hashtags.

The first COVID-related hashtag appeared on the HSL on December 31, 2019, followed by only a few ones in the following week. As the first death case occurred on January 11, second one occurred on HSL on January 16 and more infected cases detected in other cities in China as well as in the surrounding Asian countries, rumours and scared emotions about the unknown pneumonia were permeating in the society and the number of daily COVID-related hashtags started to increase rapidly on January 19. On January 20, Chinese authorities announced to the public that the new coronavirus is transmissible between humans.

From our point of view the period until January 19 can be considered as pre-COVID. During that time at most three COVID-related hashtags per day have occurred on the HSL and the cumulative number of different hashtags on HSL has grown approximately linearly with an unaltered slope (see Fig. 5.2 (A)). Around January 19 the number of COVID-related hashtags started growing and, at the same time, the overall growth of the total number of hashtags slightly decreased, indicating that the new COVID-related hashtags stay longer on HSL as compared with those before the outbreak. This results in a decrease of the total number of new hashtags per unit time on HSL. After January 19, a rapid increase can be observed in the number of COVID-hashtags (see the inset of Fig. 5.2 (A)). This has, finally, also an effect on the total cumulative number of hashtags resulting in an increased slope in Fig. 5.2 (A).

Fig. 5.2 (B) shows the cumulative number of geographically categorized COVID-hashtags with Mainland China, East Asia outside of Mainland China, and Other Countries outside of East Asia as categories. The Mainland China category starts to rise rapidly from January 19, reaches a peak in the following week, and then gradually drops with a few rebounds. The second peak and the decline of Mainland China category is intertwined with the trajectory of the Other Countries category in mid-March. The East Asia category remains at a relatively low level throughout the pandemic.

COVID-19 was first observed in east Asia, with Mainland China being the hardest-stricken region, followed by places with growing infections such as South Korea, Diamond Princess cruise ship and Japan. The epicenter of COVID-19 later shifted to Europe and the rest of the world as the situation mitigated in east Asia. The results depicted in Fig. 5.2 (B) follow these events closely, confirming the role of the real-time HSL on Weibo as a reflection

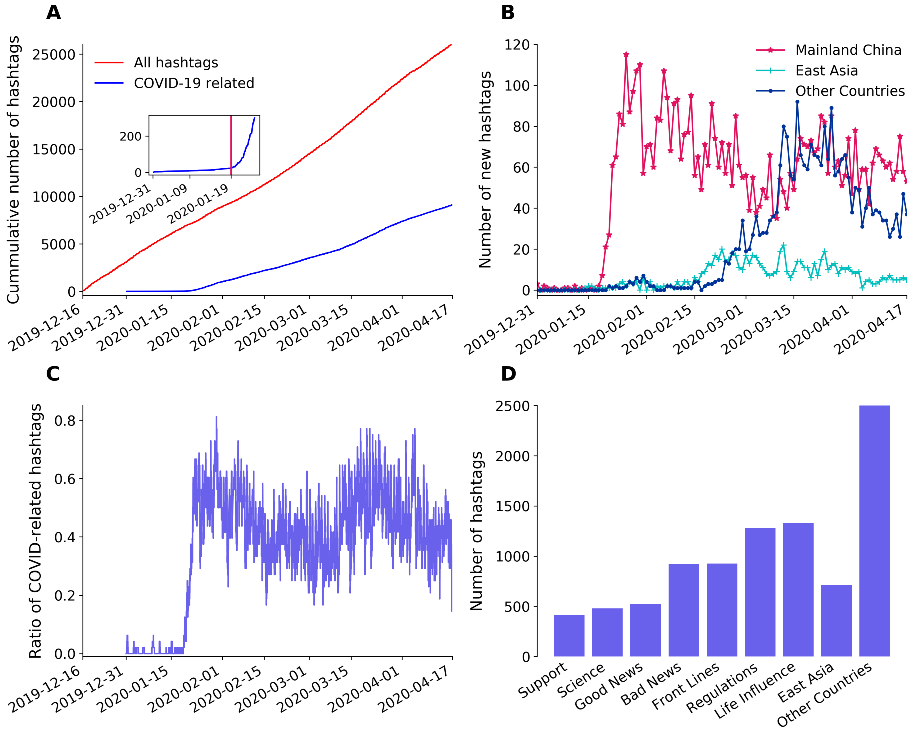

Figure 5.2: Overview of COVID-hashtags on Weibo re-ranked Hot Search List (HSL) throughout the pandemic. (A) Cumulative number of all hashtags and all COVID-hashtags with time. The inset indicates rapid increase in COVID-related hashtags starting from January 19 marked by a vertical red line. (B) Daily new COVID-hashtags on Mainland China, East Asia outside of Mainland China and Other Countries outside of East Asia. (C) Ratio of COVID-hashtags on the HSL at each timestamp. (D) Distribution of all COVID-hashtags by categories.

of the real world. Unsurprisingly, the upward and downward trend periods of Mainland China and Other Countries coincide with Fig. 5.2 (C), where the ratio of COVID-related hashtags on the HSL at each timestamp is displayed. The swift third peak on April 4 in Fig. 5.2 (C) is due to the national Qingming Festival (also known as the Tomb-Sweeping Day), where the victims who died in the COVID-19 pandemic were mourned. The dynamics of the COVID-related hashtags on the HSL demonstrates vibrant generations of newly created COVID-19 hashtags about the relevant up-to-date events around the world. Fig. 5.2 (D) shows the distribution of the hashtags in the sub-categories of Mainland China category along with East Asia and Other Countries. Among the seven sub-categories that belong to Mainland China, Support, Science, and Good News have relatively fewer hashtags, compared with Front lines, Life Influence, Bad News, and Regulations.

5.3.1 Periodization and correlations

Figure 5.3 illustrates the attention dynamics of the sub-categories of Mainland China by showing the quantity variations in Fig. 5.3 (A) (C) (E), paired with their correlation matrices with daily infections, deaths, and recoveries in Fig. 5.3 (B) (D) (F). As noted above, we have identified three periods in the investigated time interval: The first period is January 19 – February 11, separated by the huge peak in Fig. 5.1 from the the second one (February 12 – March 12). The third period (March 13 – April 17) is separated from period 2 by the second vertical line where the number of new infections has a local minimum (Fig. 5.1 inset).

In Table 5.1, we show the number of hashtags related to Mainland China in the different categories for the three periods. In Fig. 5.3, we show that the daily emergence of the categorized COVID-hashtags is dominated in the first two periods by Bad News, with increasing and decreasing trends in period 1 and period 2, respectively. In period 3, the categories Regulations, Life Influence, and Front Lines receive more attention as compared to the rest of the categories. Here the consistently high values in Regulations and Life Influence could result from the worsening world pandemic situation along with the rise of the imported infected cases in Mainland China, necessitating the establishment of measures to handle it. The categories of the Mainland China COVID-hashtags move with the number of infections and deaths in the world.

The patterns of the Pearson's correlation matrix of the ten time series reflect temporal structure with the three periods. Fig. 5.3 (B) shows a positive correlation block structure. There are strong correlations between New Death, Regulations, Science, and Bad News (upper left block) as well as between Supports, Good news and Front Lines (lower right

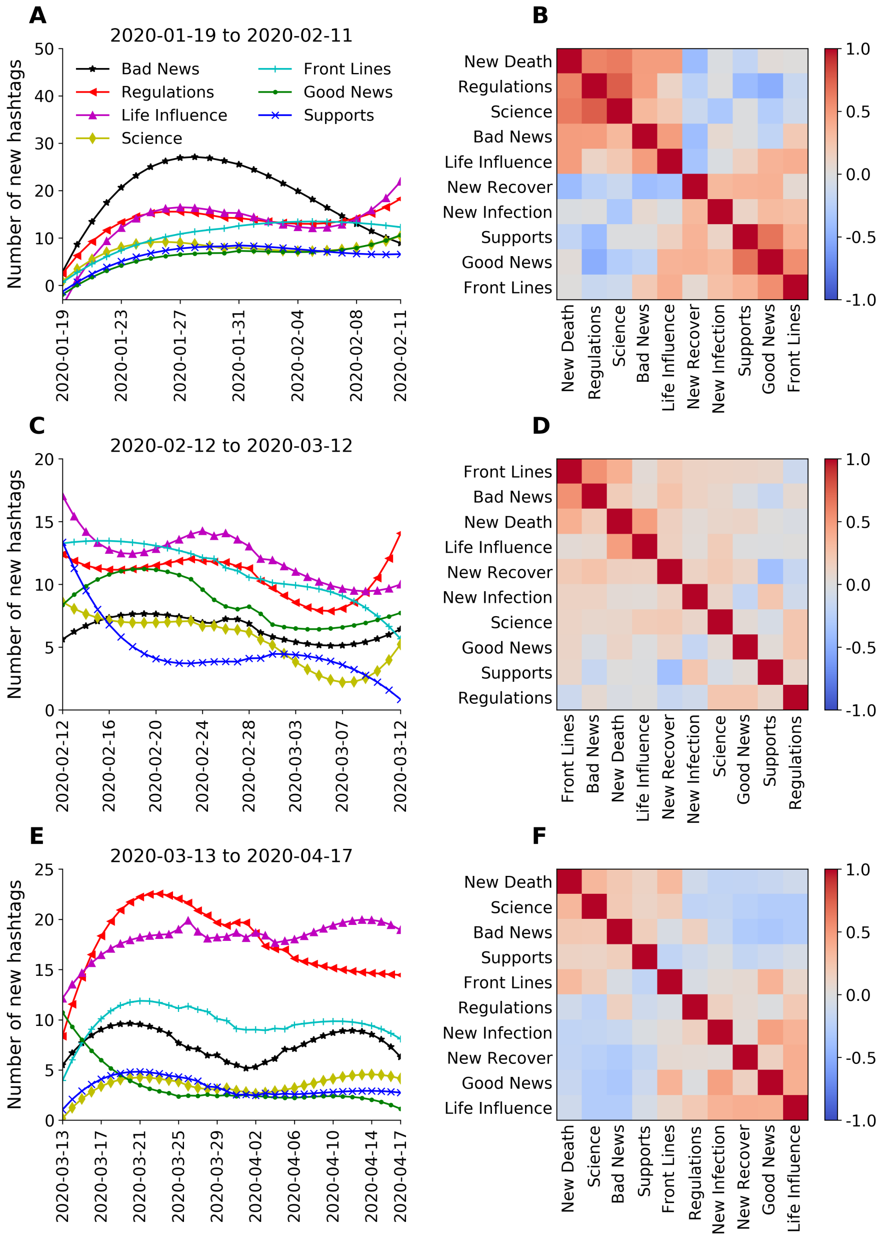

Figure 5.3: Time series of daily new hashtags (smoothened by a Savitzky-Golay filter [112] with polynomial order 3) from the sub-categories of Mainland China COVID-hashtags and their correlation matrices with daily new infections, deaths, and recoveries, in the three periods after the outbreak.

block) and there is considerable anti-correlation between the two blocks. Fig. 5.3 (D) (period 2) exhibits much weaker correlations, in fact, very few elements of the matrix reach values beyond the noise level (see SI). Exceptions are new strong correlations between New Death and Front Lines, as well as Bad News and Front Lines. In the third period (Fig. 5.3 (F)) the block structure gets again more pronounced, though not as pronounced as in the first period. Note that the categories had to be rearranged in order to achieve this structure. The major change is that Supports/Front Lines and Life influence/Regulations have exchanged positions. In period 1, the Bad News (mainly infections and deaths) of domestic cases in Mainland China were flooding, this lead to the urgent establishment of regulations, which caused life influences. In period 3, the domestic situation was under control, therefore, the Bad News in Mainland China were mainly caused by the worsening international situation (infections/deaths and Chinese coming back from abroad). Then the Regulations and corresponding Life Influences towards these issues were not anymore strongly associated with domestic deaths.

5.3.2 Rank diversity and hashtag dynamics

What is the effect of COVID-19 on the ranking dynamics? Fig. 5.4 shows a comparison of the rank diversity in the top 48 ranks taking non-COVID and COVID hashtags in different periods. Striking differences are observed between the rank diversity plots before and after the outbreak. As Fig. 5.4 (A) suggests, the rank diversity plot before the outbreak was approximately linear with moderate fluctuations. A clear gap emerges in the rank diversity after rank 15 in Fig. 5.4 (B) during the COVID period. We recognize resemblances in the rank diversity plots before the outbreak and after the outbreak considering only non-COVID hashtags, except for the strange drops at ranks 29 and 34 in Fig. 5.4 (C). Comparing Fig. 5.4 (D) with Fig. 5.4 (B), the gap after rank 15 is larger in the rank diversity plot considering only COVID-related hashtags. The rank diversity plots for hashtags in period 1 surpass period 2 and period 3 with both non-COVID and COVID hashtags as depicted in Fig. 5.4 (C) and (D), while the difference is much higher in the latter case.

Fig. 5.4 gives evidence that the COVID-hashtags cause the gap in the rank diversity plot after the outbreak. Taking the normalized rank diversity plot before the outbreak as a reference, a higher normalized rank diversity at a certain rank position represents a higher number of unique occurrences within the observation period, so that the COVID-related hashtags in the top 15 ranks change faster (with higher frequency) than normal. One possible explanation is that the COVID-hashtags kept emerging with higher frequency than before the outbreak and people payed much attention to these new hashtags. Addition-

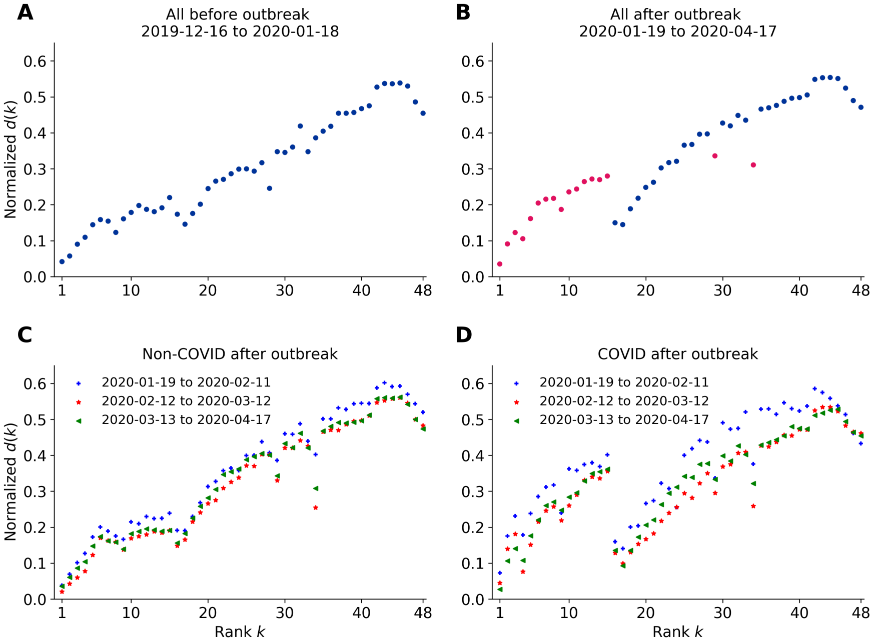

Figure 5.4: Rank diversity $d(k)$ of the 48 ranks on the HSL before and after COVID-19 outbreak. (A) Rank diversity taking all hashtags in our observation period before the outbreak, approximately linear except for the head and tail parts, with small fluctuations. (B) Rank diversity taking all hashtags after the outbreak, with strange points colored in red. A large gap occurs after the top 15th rank. (C) Rank diversity taking all non-COVID hashtags in the three periods after the outbreak, strong resemblances with (A). (D) Rank diversity taking all COVID-hashtags in the three periods after the outbreak. The result in period 1 is higher than period 2 and period 3, revealing a more dynamic change of the hashtags appeared on the HSL. The gap after rank 15 is more severe compared to (B).

ally, when the flooding hashtags contained similar information such as the new infections and deaths in different cities or provinces of China, the public interest towards individual hashtags could drop quickly, resulting in a higher number of unique hashtags at certain ranks in unit time on HSL. This effect of higher rank diversity for higher ranks seems to be amplified by the algorithm leading to the observed gap.

Strange drops of rank diversity at ranks 29 and 34 can also be seen on our plots in Fig. 5.4. As provided in SI, there are hashtags that stay at the ranks 29 and 34 for an unusually long time and then disappear from the HSL, indicating algorithmic intervention from Weibo. As one of the most popular and influential social media in China, Weibo might shoulder the responsibility during the global public health emergency to keep people informed about related news in China and around the globe, by means of changing the algorithm towards COVID-hashtags to promote crucial news and keep them updating in the top 15 positions and leave the list at rank 29 or 34. Our methods are sensitive enough to demonstrate this type of interventions. Therefore our observations reflect a combination of both spontaneous attention dynamics from the public and the controlled effects from Sina Weibo.

Rank diversity captures attention dynamics from the point of view of the overall dynamical rank movements of the hashtags on the HSL. It is interesting to follow the dynamics also from the aspect of the individual hashtags. The average highest rank of a category of hashtags on a given day is characteristic to the attention paid to that category. (Note, of course, that getting to the HSL expresses already considerable attention.) Similarly, the average duration is another measure of attention. However, in the latter case it should be mentioned that short duration can be caused by decaying attention to the general topic (in this case the hashtag is likely to be replaced by another from a different topic) or because of the heavy stream of new hashtags of the same topic.

How do the average highest rank and average duration accumulate with time? As Fig. 5.5 (A) shows, the cumulative average highest rank, H_j is initially at a top rank, indicating that the first few hashtags about the unknown pneumonia received a huge amount of attention from the public. As more COVID-related hashtags occurred, H_j becomes lower, with a rapid change at the beginning and a slower change later, separated by around January 30. This is due to the rapidly increasing number of COVID-related hashtags and the limited number of ranks on HSL. In Fig. 5.5 (B), the first peak of the cumulative average duration, D_j is on January 8, when the hashtag that eight patients infected by the unknown pneumonia recovered from hospital occurred. Then the D_j decreases first and then increases again, reaching the second peak on January 22, after which the increasing daily new hashtags with short durations started to play a greater role than the few hashtags with long durations.

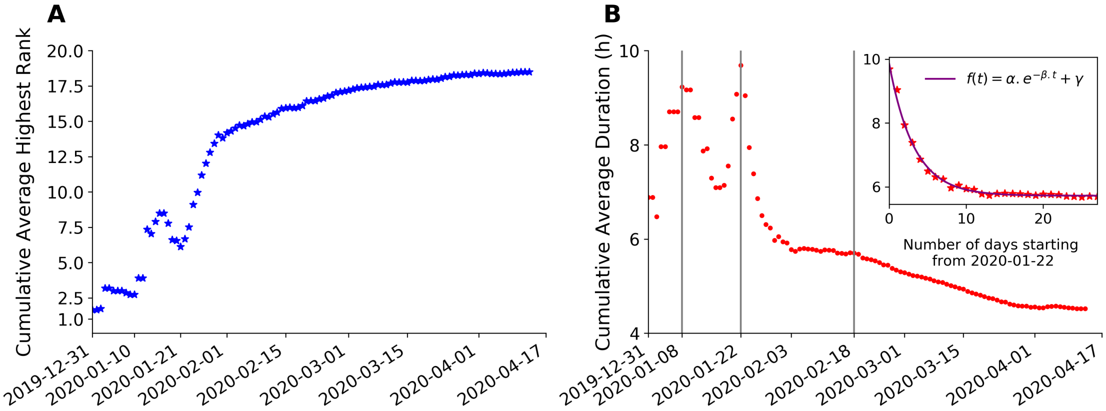

Figure 5.5: Attention decay. (A) Cumulative average highest rank of COVID-hashtags whose first appearance was in the time interval since December 31, 2019. (B) Cumulative average duration (hours). The inset shows a three-parameter exponential fit ($\alpha = 4.13\text{h}, \beta = 0.31\text{h}/\text{day}, \gamma = 5.72\text{h}$) for the cumulative average duration decay after January 22, 2020.

The fast decay of D_j in the period between January 22 and February 18 (see the inset in Fig. 5.5 (B)) was fitted by an exponential function:

$$f(t) = \alpha e^{-\beta.t} + \gamma, \tag{5.3}$$

with $\alpha = 4.13\text{h}, \beta = 0.31\text{h}/\text{day}, \gamma = 5.72\text{h}$. On February 18, hashtags of positive changes in the COVID situation started to appear on HSL. After that, the D_j exhibits a slower and longer decay.

5.4 Discussion

In this work, we have studied the public attention dynamics on the real-time Hot Search List (HSL) of the biggest Chinese microblogging website Sina Weibo under the influence of the COVID-19 pandemic. On the one hand, such study contributes to the understanding of the dynamics of public attention on social media and how it reflects the dynamics of the public thoughts and behaviors. On the other hand, identifying the online attention dynamics patterns and their relationship to events and measures during pandemic may contribute to its efficient management.

In order to follow the dynamics of public attention we have introduced sub-categories of COVID-19-related hashtags. Our results show diversification of the public attention after the outbreak on January 19, 2020 as indicated by changing frequencies of such hashtags in the different sub-categories. Moreover, the pattern of correlations with the real-world events and measures vary in three identified periods during the investigated time span. We conclude that at the beginning the dominant driving force of the public attention was the infection and death situation in Mainland China, with mainly domestic cases, while the international situation and the imported cases influenced the attention later. Our observations point toward the complexity of the attention patterns indicating that several components should be taken into account if such data are used to the prediction of the epidemic curve [75, 55].

Furthermore, we have shown that the cumulative average duration follows exponential decay immediately after the attention peak in the pandemic, but a slower decay for a longer time. The exponential decay suggests that the speed of governmental response is crucial in the early pandemic phase. This exponential decay sets a scale for the governments within which it should take quick actions and publish crucial measures and regulations to control the pandemic, healthcare experts should deliver scientific knowledge to inform the public how to protect themselves efficiently. The attention toward COVID-hashtags decayed as the circumstances in China got better. Nevertheless, the attention was influenced by the world pandemic situation which kept changing, hence the decay of public attention on the Chinese social media Weibo has become less clear cut. In any case, targeted and timely stimulus should be given to keep the attention and awareness of the public throughout the pandemic to prevent future waves of COVID-19.

In this chapter we have made the first step to relate the ranking dynamics of hot topics on social media with the public attention dynamics. We have provided a novel approach to study and quantify the attention dynamics taking advantage of the real-time Hot Search List (HSL) on Weibo. The rank diversity in the top 15 ranks containing COVID-hashtags are higher than normal. This could result from the spontaneous preference from the public towards COVID-related information. More likely, it is due to an algorithmic intervention towards COVID-hashtags from the platform provider. Sina Weibo may intentionally promote the COVID-related important information to make sure people will get aware of them. In this sense, such an algorithmic intervention can be useful for the public. The empirical rank diversity could be a combined influence of both the Weibo algorithm and the spontaneous public preference. This observation shows the possibility that rank diversity could be an adequate tool to investigate further the important aspect of algorithmic intervention in social media data.

Besides exploring the attention dynamics on the Chinese social media Sina Weibo, we also studied the cumulative growth of all topics and all the COVID-topics on Twitter trending list in the United States. As is shown in the supplementary information Fig. 5.6A, the cumulative number of all the Twitter trending topics in the United States was almost perfectly linear from January 1, 2020 to April 16, 2020. The time period that the cumulative number of all COVID-topics on Twitter trending list increases is in accordance with the rising period of the number of hashtags in the Other Countries category in Fig. 5.2B. The similarity of results on Sina Weibo HSL and Twitter trending list is a reflection that both platforms are influenced similarly by the major events worldwide during the COVID-19 pandemic. Though having more daily new topics on Twitter trending list than Weibo HSL, the number of COVID-topics on Twitter is much fewer. The topics on Twitter are generally shorter and have broader meaning, for example, #QuarantineLife, while the hashtags on Weibo are more detailed, for example, #小区窗台演唱会庆祝解除隔离# (#Community window concert to celebrate the lifting of quarantine#), contributing to the rich number of diverse hashtags on Weibo. It should be emphasized that both for Sina Weibo and Twitter the lists are produced by unknown algorithms and in the case of Sina Weibo we have been able to pinpoint direct interventions from the side of the provider into the ranking. However, the detailedness of Weibo HSL, its fixed length and the fact that HSL is the same for ordinary users seem to make Weibo HSL more suitable to study attention dynamics through ranking than Twitter, as Twitter trending lists are without fixed length and can be personalized.

5.5 Supplementary information

5.5.1 SI1 Twitter trending COVID-topics in the United States

Sina Weibo is the largest microblogging site in China, where Twitter, the worldwide most popular service of this kind does not operate. It is a natural idea to try to compare our observations made on Sina Weibo with Twitter attention dynamics. Unfortunately, there is no comparable statistics on Twitter to the HSL. Instead, Twitter has the service to inform about most retweeted hashtags during the last 24 hours updated on the minute basis and broken down to countries [99]. We have chosen to study the US tweets.

Categorization of tweets has been widely investigated [137, 52], including recent attempts to analyze the impact of COVID-related topics [91] on Twitter by analyzing the sentiments to 10 words related to COVID. Twitter even created a “COVID-19 stream” [105] to promote this type of research. In spite of these, a direct comparison of our results on Sina Weibo with Twitter is hindered by a number of factors, including the different characters of the listings, the different roles hashtags play in these services and the differences due to the scripts. Nevertheless, we tried to capture at least the overall trends (see Fig. 5.6).

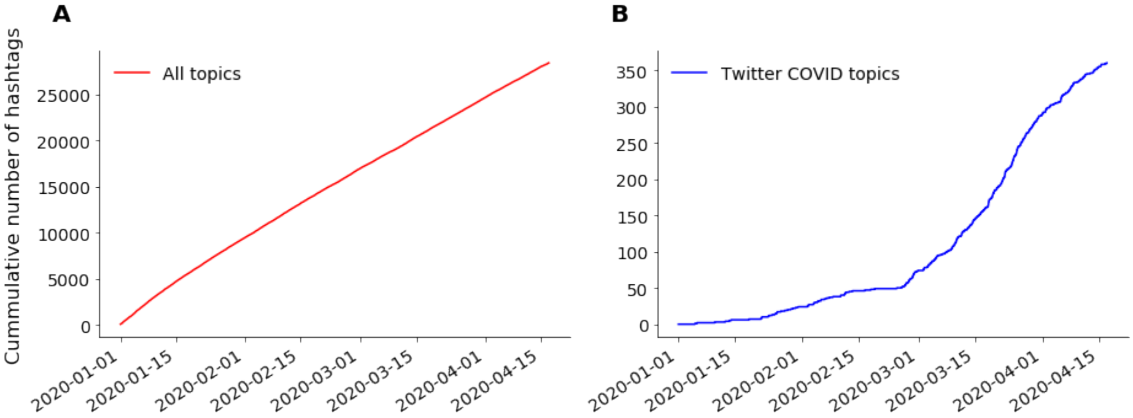

Figure 5.6: Overview of the cumulative number of topics during the observation period on Twitter trending list in the United States from January 1, 2020 to April 16, 2020. (A) Cumulative growth of all topics. (B) Cumulative growth of COVID-related topics.

Fig. 5.6 (A) shows the cumulative number of all the Twitter trending topics in the United States is almost perfectly linear. As Fig. 5.6 (B) shows, the COVID-topics on Twitter trending list first grows very slowly at the beginning phase, and then starts to increase dramatically from late February 2020. The rate of COVID-related topics is, however, much smaller in the Twitter list than on that of the Sina Weibo.

5.5.2 SI2 Significance of correlations

To understand how the categories of time series of daily new hashtags move together and whether there are blocks of categories that co-move, we presented the correlation matrices plot between the ten time series in the three periods after the outbreak. In order to get information about the significance of the correlations we apply a null model, which is created by shuffling the times of the individual values, thus smearing out the correlations.

Due to the finiteness of the time series, there will be non-zero background noise level denoted by Z in the null model, defining the background to which measured real correlations can be compared. Z is calculated by correlating 500 shuffled time series for each of the 10 categories. We observed that all the pairs have similar standard deviations between around 0.16 to 0.2. We take a uniform value $Z = 0.2$.

In Fig. 5.7 we show correlations where only those C_{ij} correlation matrix elements are presented for which $Z < |C_{ij}|$. The figure shows the different Mainland China topical categories and their thresholded correlations in the three pandemic phases. In Fig. 5.7 (B) most of the correlations are beyond the threshold, while in Fig. 5.7 (D) very few are beyond the threshold. In Fig. 5.7 (F), though some values at the upper left and lower right corners are beyond the threshold, they are much weaker than in Fig. 5.7 (B).

5.5.3 SI3 Categorized Sina Weibo hashtags and properties

We showed in Fig. 5.4 that the gap between the top 15 ranks and the rest of the ranks in the rank diversity plot after the outbreak is caused by the COVID-hashtags. In order to further understand the properties of COVID-hashtags and how they influenced the HSL hashtag dynamics, we compared the highest rank and duration distribution of different COVID-categories with the non-COVID hashtags before and after the outbreak.

Fig. 5.8 shows a detailed comparison of the highest rank and duration of the categorized Mainland China COVID-hashtags on Weibo Hot Search List (HSL), before and after the COVID-19 outbreak. As Fig. 5.8 (A) shows, most of the categories have a median of highest rank close to 15. Science category and Bad News category are generally higher ranked than other categories. The median highest rank of the non-COVID hashtags after the outbreak is the same with that of the hashtags before the outbreak (rank 19), while the median highest rank of the COVID-hashtags is higher than both (rank 16). Fig. 5.8 (B) shows the lifetime duration of the different categories. The median duration of most of the categories is less than 3.5 hours. Science category has the highest duration among all categories. Non-COVID hashtags after the outbreak (3.95 hours) and hashtags before the outbreak (3.80 hours) have similar duration distributions. The COVID-hashtags generally have shorter duration (3.21 hours) than non-COVID hashtags.

5.5.4 SI4 Hashtag rank trajectory examples

In this chapter, we have seen strange drops in the rank diversity plot at the ranks 29 and 34 after the outbreak, this implies that the number of unique hashtags occurred at these ranks

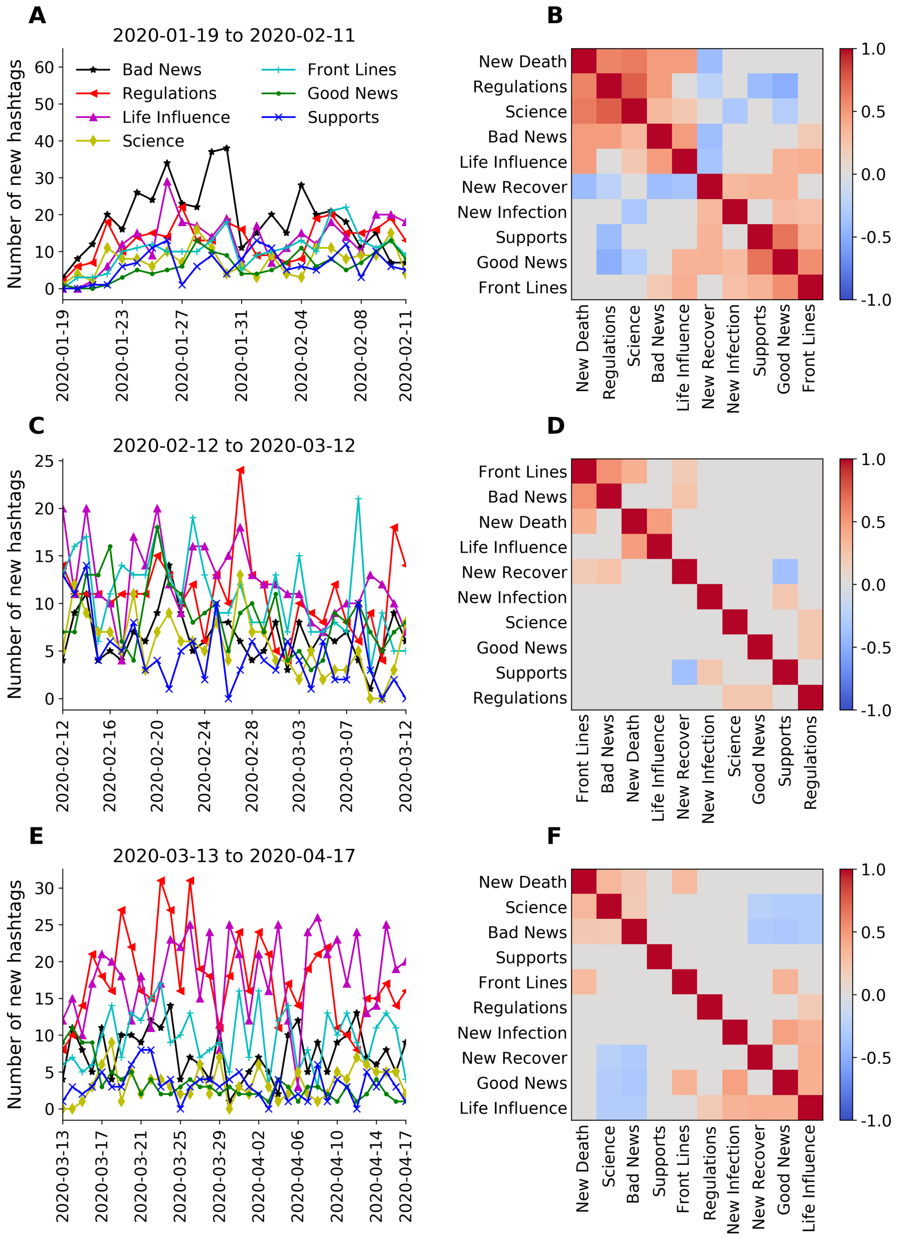

Figure 5.7: Time series of daily new hashtags (unsmoothened) from the sub-categories of Mainland China COVID-hashtags and their correlation matrices with daily new infections, deaths, and recoveries, in the three periods after the outbreak. Correlations lower than 0.2 are considered as insignificant and are converted to zero.

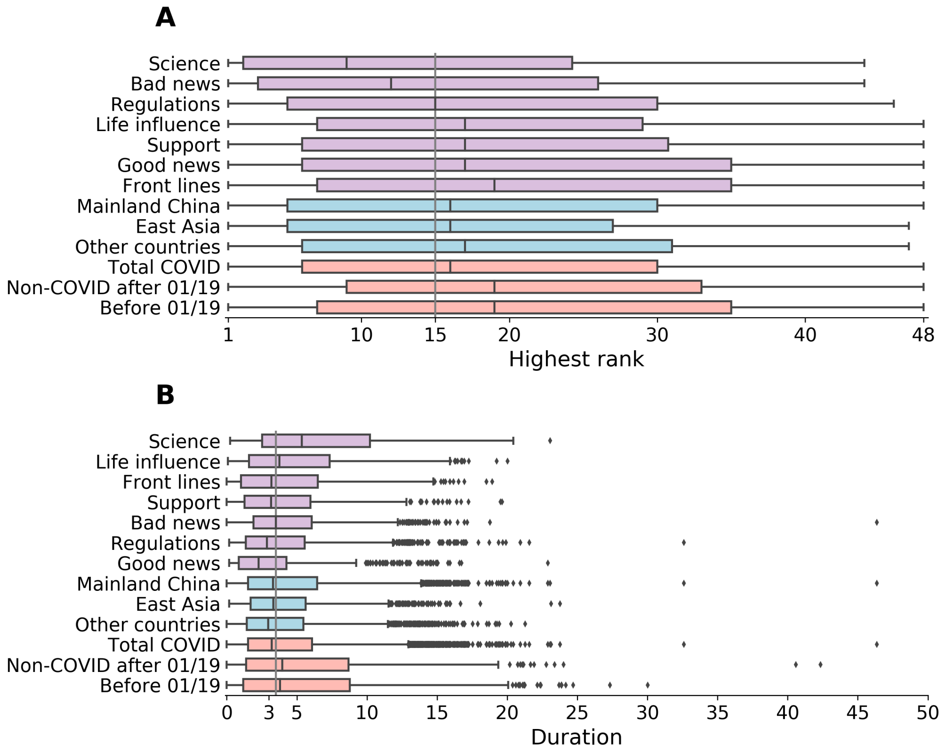

Figure 5.8: Boxplots of the highest rank and duration of the different categories. In both plots, the purple categories are sub-categories of Mainland China category, which is colored in blue. The blue categories are sub-categories of Total COVID category, which is colored in orange. The dots in (B) are the outlier hashtags with long duration, e.g, the #疫情地图# (#Infection Map#) in the Bad News category and the #武汉再发现居家确诊病人将问责# (#(District Party Secretary) Will be accountable if home-confirmed patients found again in Wuhan#) in the Regulations category.

in a given time interval is smaller than usual, so that there should be hashtags staying there for unusually long time. Here we present examples of normal and abnormal hashtag rank trajectory plots, and verify there are hashtags that stay at certain ranks such as rank 29 and 34 on the HSL for a strangely long time without any fluctuation.

Fig. 5.9 shows examples of abnormal and normal rank trajectory plots of COVID-related hashtags on Weibo HSL. In Fig. 5.9 (A), (C), (E), the ranks of the hashtags stay strangely long time at ranks 29 and 34, and then disappear from the HSL. Fig. 5.9 (B), (D), (F) show relatively natural fluctuations in the rank trajectory plots. The example hashtags and their translations are shown in Table 5.2. The abnormal rank plots are likely due to the algorithmic intervention from Sina Weibo.

Table 5.2: Chinese original and translations of example hashtags in Fig. 5.9.

Example Hashtags	Translation
#企业复工要为职工配发口罩#	#Enterprises must distribute masks to employees when they return to work#
#中国不会出现大规模通货膨胀#	#China will not see massive inflation#
#联合国秘书长呼吁全球共同向新冠宣战#	#The UN Secretary-General calls on the world to declare war on COVID-19#
#武汉封城#	#Wuhan lockdown#
#一图看懂新型冠状病毒肺炎#	#A picture to understand the new coronavirus pneumonia#
#疫情地图#	#Infection map#

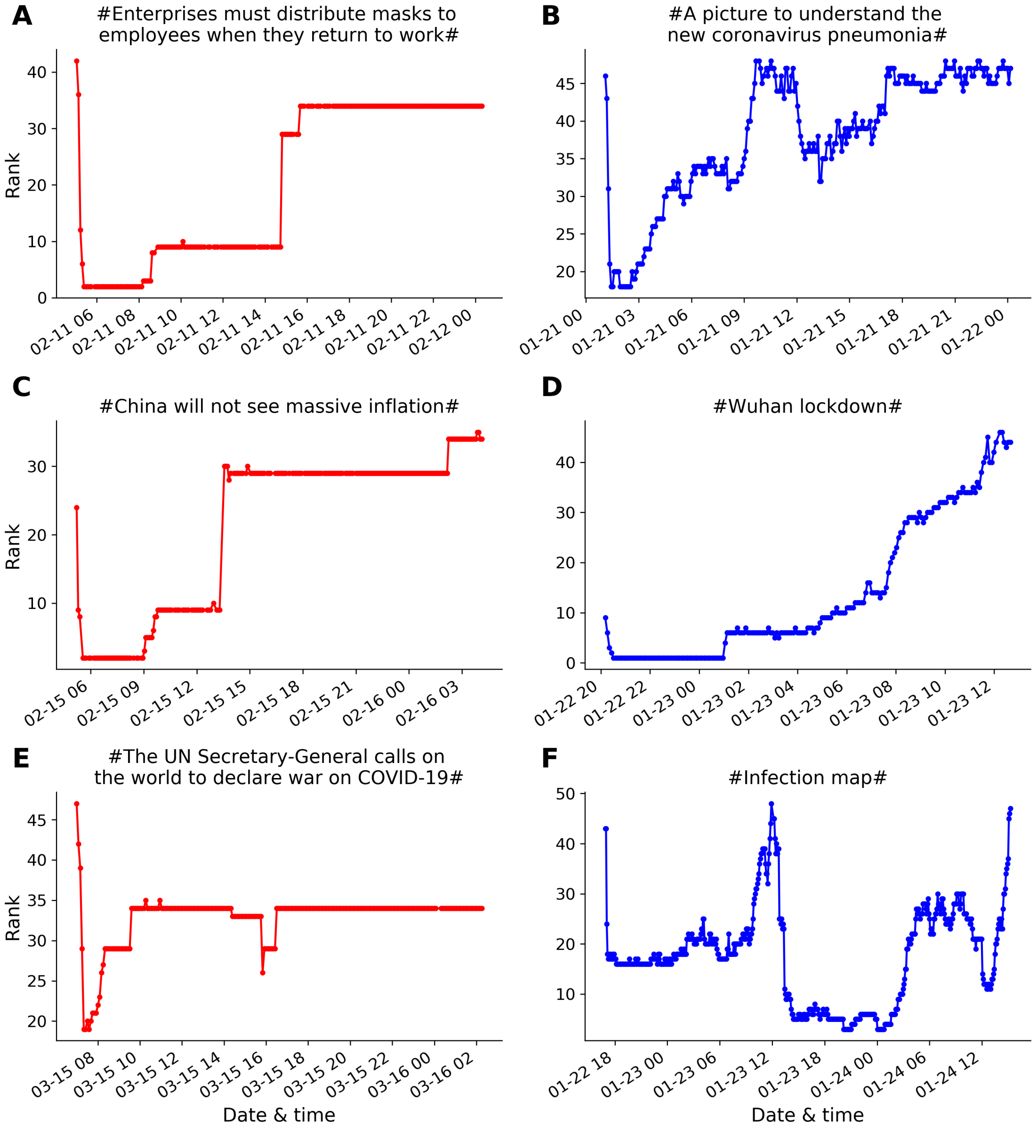

Figure 5.9: Examples of rank trajectory plots of COVID-related hashtags. (A), (C), (E) Abnormal rank trajectory plots. (B), (D), (F) Normal rank trajectory plots.

Chapter 6

Conclusions

6.1 Summary

The development of digital technology has changed our lives in many ways, resulting, among others, in a shift from the struggle for information to the struggle with the information deluge. Being exposed to so many effects, the attention of people has therefore become valuable and there is a never-ending competition for it from the side of marketing, politics, and governance. Attention of people at societal level results in popularity, sometimes lasting very short in the spirit of Andy Warhol, however, having an impact on sales figures and political elections. Understanding the mechanisms leading to popularity and influencing its evolution is therefore both a scientific challenge and important for applications. New media provide a promising testing ground for such research.

This thesis studies the attention dynamics on the Chinese microblogging site Sina Weibo from the aspects of popularity emergence, competition, and the influence of exogenous factors, in this case the COVID-19 pandemic. By using the Hot Search List of Sina Weibo and crawling data directly, we focused on questions related to how the hashtags become popular and the patterns of their ranking dynamics on the HSL. Special attention was paid to signatures of interventions by the service provider influencing the mechanism of getting to the HSL and the ranking of displayed hashtags. This thesis brings insight into popularity dynamics on microblogging sites, and it also gives tools to identify the aforementioned interventions.

After providing a condensed literature overview, I described the emergence mechanism of the popular hashtags on Sina Weibo in chapter 3. I showed the routes leading to the popularity of the successful hashtags by analyzing the characteristics of the evolution dynamics of their repost networks in the prehistory before their appearance on the HSL. I identified two extreme categories “Born in Rome” and “Sleeping Beauty” based on how fast the hashtags reach the HSL after birth. I then related the repost network growth patterns to the two categories as well as the role of the hubs. I found that the birth timing of the hashtags is crucial in influencing whether they will be popular enough to appear on the HSL. Understanding the emergence mechanism of hashtag popularity and the importance of timing, on the one hand, could contribute to marketing and maximizing the spreading efficiency by playing with these factors. On the other hand, it provides Weibo users with better knowledge to differentiate about the possible social capital influence in promoting certain contents, such as Star hashtags.

For the dynamics of already popular hashtags, I relate measures of ranking for those successful hashtags on the HSL as a proxy of the attention dynamics which is a novel approach in quantifying attention dynamics on social media. In chapter 4, I focus on the dynamics of popular hashtags after their appearance on the Sina Weibo HSL by studying the patterns of their rank trajectories, dynamics of duration and ranking. I pinpoint signatures of anomalies from the observations and propose a ranking model with anchoring effect to simulate the interventions on the ranking list by the platform provider. Our results indicate the interventions of the platform provider in the ranking of the hashtags on the HSL. As service providers have financial interests and may be under political pressure, the objectivity of the ranking lists and the truth of their content can be questioned. Similarly to the fight against fake news, the fight against the manipulation of public attention is in the interest of society, and it also needs the tools to detect interventions.

The outbreak of COVID-19 pandemic provided us a chance to study the influence of exogenous events on the popularity list of Sina Weibo. In chapter 5, by following the time evolution of the ranking, I could identify three different periods of the beginning phase of the pandemic and detect how the attention of the people changed during them as well as how the decay of the attention related to COVID topics changed depending on the period. I applied a novel approach to study and quantify the attention dynamics taking advantage of the Weibo HSL and made a comparison of the hashtag ranking dynamics before and after the outbreak. I discovered anomalies attributed to the COVID-related hashtags on the HSL measured by rank diversity, which revealed the possibility of interventions by the service provider. This observation shows the possibility that rank diversity could be an adequate tool to investigate further the important aspect of algorithmic intervention in social media data. Our work shows the importance of taking into account of platform interventions when doing research using data from social media.

6.2 Limitations and outlook

Discovering regularities of attention dynamics is of great interest for several reasons, as mentioned above. Online platforms provide an abundance of data for such studies, however, this data should be handled with care. Important problems like the relationship between information spreading and success, prehistory of the outburst of popularity and ranking dynamics could be studied. However, many factors could be involved in the spreading and popularity of the hashtags. The fact that some hashtags become trends in such systems can not only be due to natural selection by the internet audience, but result from algorithmic recommendation, manipulations from the Internet water army, government-driven campaigns or combinations of joint factors. It is a challenging task to separate the different effects, whether it is spontaneous, algorithmic or governmental.

The existence of various interventions blurred the regularities of the properties of the microblogging system, hindering the understanding of the natural processes purely originating from users. On the other hand, the observation of irregularities make it possible for us to identify interventions and be more aware of how the system is being manipulated. One of our results based on the detected anomalies in the ranking process is that on Sina Weibo the intervention by the service provider is apparently rather strong, despite of the claims of using an objective formula for the HSL. A very interesting future research topic could be to study deeper the different kinds of interventions. In order to do so, one would need longer observation period, more data, and novel approaches including advanced text analysis and machine learning methods. Also, the extension of the investigation to other media, like online news sites could be helpful. An ideal goal would be to uncover the hidden algorithm (or set of rules) used for interventions.

This thesis focuses on the Chinese microblogging site Sina Weibo, and it is important to see whether the results or a part of them can be generally applied. Clearly, some results are idiosyncratic to this specific platform. For instance, the night-time break with stopping hashtags to appear on the HSL seems to be specific to Sina Weibo, closely linked to the manual control probably used often for interventions. Every platform has its own algorithm and characteristics. Nevertheless, the relationship between the prehistory and the success as well as the ranking dynamics could have general aspects. Most importantly, the approach used in this thesis should have a broad applicability. This includes the tracing of the repost networks and relating them to the success of the hashtags or the use of detecting irregularities in the ranking characteristics to identify interventions or other anomalies. It would be very interesting to conduct a comparative study of diverse media platforms and identify the overall rules leading to popularity together with the specific characteristics of the individual platforms.

The approach used here to analyze the popularity dynamics of hashtags can be applied more generally. The success of cultural products such as popular songs, and best-selling books also has a prehistory period before they achieve a high level of popularity. During this time, people interact with each other by commenting, recommending, liking and so on, and there are also rankings available for the most popular items. The analysis presented in this thesis could apply to such products, and it would be interesting to see if there are universalities or similarities with hashtags on microblogging sites in their dynamics. Such a study could contribute to understanding the difference between endogenous and exogenous effects in the development of popularity, which is an important question in marketing.

Bibliography

[1] Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra. Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, 1(3):1, 2018.

[2] Weibo Administrator. Weibo hot search regulation rules. *Sina Weibo* https://weibo.com/1934183965/KuKyPkp8Y?type=repost, 2021.

[3] Morales A.J., Borondo J., Losada J.C., and Benito R.M. Efficiency of human activity on information spreading on twitter. *Social Networks*, 39:1–11, 10 2014.

[4] Cevat Giray Aksoy, Michael Ganslmeier, and Panu Poutvaara. Public attention and policy responses to covid-19 pandemic. *MedRxiv*, 2020.

[5] Ali Alessa and Miad Faezipour. A review of influenza detection and prediction through social networking sites. *Theoretical Biology and Medical Modelling*, 15(2), 2018.

[6] Thayer Alshaabi, Michael V Arnold, Joshua R Minot, Jane Lydia Adams, David Rushing Dewhurst, Andrew J Reagan, Roby Muhamad, Christopher M Danforth, and Peter Sheridan Dodds. How the world's collective attention is being paid to a pandemic: Covid-19 related n-gram time series for 24 languages on twitter. *Plos one*, 16(1):e0244476, 2021.

[7] Issa Annamoradnejad and Jafar Habibi. A comprehensive analysis of twitter trending topics. In *2019 5th International Conference on Web Research (ICWR)*, pages 22–27. IEEE, 2019.

[8] Sitaram Asur, Bernardo A Huberman, Gabor Szabo, and Chunyan Wang. Trends in social media: Persistence and decay. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):434–437, 2011.

[9] David Bamman, Brendan O'Connor, and Noah Smith. Censorship and deletion practices in chinese social media. *First Monday*, 2012.

[10] Peng Bao, Hua-Wei Shen, Junming Huang, and Xue-Qi Cheng. Popularity prediction in microblogging network: a case study on sina weibo. In *Proceedings of the 22nd international conference on world wide web*, pages 177–178, 2013.

[11] Lachlan Birdsey, Claudia Szabo, and Yong Meng Teo. Twitter knows: understanding the emergence of topics in social networks. In *2015 Winter Simulation Conference (WSC)*, pages 4009–4020. IEEE, 2015.

[12] Nicholas Blumm, Gourab Ghoshal, Zalán Forró, Maximilian Schich, Ginestra Bianconi, Jean-Philippe Bouchaud, and Albert-László Barabási. Dynamics of ranking processes in complex systems. *Physical Review Letters*, 109:128701, 2012.

[13] Francesco Bonchi, Carlos Castillo, and Dino Ienco. Meme ranking to maximize posts virality in microblogging platforms. *Journal of Intelligent Information Systems*, 40(2):211–239, Apr 2013.

[14] Sina Weibo Data Center. 2015 weibo user development report. *Weibo Report.* https://data.weibo.com/report/reportDetail?id=333, 2016.

[15] Sina Weibo Data Center. Weibo 2020 user development report. *Weibo Report.* https://data.weibo.com/report/reportDetail?id=456, 2021.

[16] Le Chen, Chi Zhang, and Christo Wilson. Tweeting under pressure: analyzing trending topics and evolving word choice on sina weibo. In *Proceedings of the first ACM conference on Online social networks*, pages 89–100, 2013.

[17] Xinyu Chen, Youngwoon Cho, and Suk Young Jang. Crime prediction using twitter sentiment and weather. In *2015 systems and information engineering design symposium*, pages 63–68. IEEE, 2015.

[18] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118, 2010.

[19] Rumi Chunara, Jason R. Andrews, and John S. Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American Society of Tropical Medicine and Hygiene*, 86(1):39–45, 2012.

[20] Groden Claire. http://fortune.com/2016/01/21/weibo-character-limit/.

[21] Regino Criado, Esther García, Francisco Pedroche, and Miguel Romance. A new method for comparing rankings through complex networks: Model and analysis of competitiveness of major european soccer leagues. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(4):043114, 2013.

[22] Hao Cui. Attention dynamics sina weibo covid19. https://github.com/cuihaosabrina/Attention_Dynamics_Sina_Weibo_COVID19. Accessed Sep. 18, 2022.

[23] Hao Cui. Evolution of repost networks before getting to the hsl. https://drive.google.com/drive/folders/1JULm8eNswUSOY4PC-cjTtvP4ocJzvaM-. Accessed Sep. 18, 2022.

[24] Hao Cui. Sina weibo interventions. https://github.com/cuihaosabrina/Sina_Weibo_Interventions. Accessed Sep. 18, 2022.

[25] Hao Cui and János Kertész. Attention dynamics on the chinese social media sina weibo during the covid-19 pandemic. *EPJ data science*, 10(1):8, 2021.

[26] Hao Cui and János Kertész. Born in rome or sleeping beauty: Emergence of hashtag popularity on a microblogging site. *arXiv preprint arXiv:2203.14802*, 2022.

[27] Hao Cui and János Kertész. Competition for popularity and identification of interventions on a chinese microblogging site. *arXiv preprint arXiv:2208.10176*, 2022.

[28] Sourav Das and Anup Kumar Kolya. Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on covid-19 by deep convolutional neural network. *Evolutionary Intelligence*, pages 1–22, 2021.

[29] David Rushing Dewhurst, Thayer Alshaabi, Michael V Arnold, Joshua R Minot, Christopher M Danforth, and Peter Sheridan Dodds. Divergent modes of online collective attention to the covid-19 pandemic are associated with future caseload variance. *arXiv preprint arXiv:2004.03516*, 2020.

[30] Souvik Dubey, Payel Biswas, Ritwik Ghosh, Subhankar Chatterjee, Mahua Jana Dubey, Subham Chatterjee, Durjoy Lahiri, and Carl J. Lavie. Psychosocial impact of covid-19. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(5):779 – 788, 2020.

[31] Joel Dyer and Blas Kolic. Public risk perception and emotion on twitter during the covid-19 pandemic. *Applied Network Science*, 5(1):1–32, 2020.

[32] Young-Ho Eom, Michelangelo Puliga, Jasmina Smailović, Igor Mozetič, and Guido Caldarelli. Twitter-based analysis of the dynamics of collective attention to political parties. *PloS one*, 10(7):e0131184, 2015.

[33] P. Fan, P. Li, Z. Jiang, W. Li, and H. Wang. Measurement and analysis of topology and information propagation on sina-microblog. In *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics*, pages 396–401, July 2011.

[34] Rui Fan, Jichang Zhao, Yan Chen, and Ke Xu. Anger is more influential than joy: Sentiment correlation in weibo. *PloS one*, 9(10):e110184, 2014.

[35] Rui Fan, Jichang Zhao, and Ke Xu. Topic dynamics in weibo: a comprehensive study. *Social Network Analysis and Mining*, 5(1):1–15, 2015.

[36] Karen Freberg, Kristin Graham, Karen McGaughey, and Laura A Freberg. Who are the social media influencers? a study of public perceptions of personality. *Public relations review*, 37(1):90–92, 2011.

[37] Junling Gao, Pinpin Zheng, Yingnan Jia, Hao Chen, Yimeng Mao, Suhong Chen, Yi Wang, Hua Fu, and Junming Dai. Mental health problems and social media exposure during covid-19 outbreak. *PloS one*, 15(4):e0231924, 2020.

[38] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2016.

[39] Balakrishnan Gokulakrishnan, Pavalanathan Priyanthan, Thiruchittampalam Ragavan, Nadarajah Prasath, and AShehan Perera. Opinion mining and sentiment analysis on a twitter data stream. In *International conference on advances in ICT for emerging regions (ICTer2012)*, pages 182–188. IEEE, 2012.

[40] David Graus, Daan Odijk, and Maarten de Rijke. The birth of collective memories: Analyzing emerging entities in text streams. *Journal of the Association for Information Science and Technology*, 69(6):773–786, 2018.

[41] W. Han, X. Zhu, Z. Zhu, W. Chen, W. Zheng, and J. Lu. Weibo, and a tale of two worlds. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 121–128, Aug 2015.

[42] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren

Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[43] Annie Hellweg. Social media sites of politicians influence their perception by constituents. *The Elon Journal of Undergraduate Research in Communications*, 2(1):22–36, 2011.

[44] D. Hewitt. Weibo brings changes to china. *BBC News* https://www.bbc.com/news/magazine-18773111, 2012.

[45] Thomas S Higgins, Arthur W Wu, Dhruv Sharma, Elisa A Illing, Kolin Rubel, and Jonathan Y Ting. Correlations of online search engine trends with coronavirus disease (covid-19) incidence: Infodemiology study. *JMIR Public Health Surveill*, 6(2):e19702, May 2020.

[46] Tao Hu, Siqin Wang, Wei Luo, Mengxi Zhang, Xiao Huang, Yingwei Yan, Regina Liu, Kelly Ly, Viraj Kacker, Bing She, et al. Revealing public opinion towards covid-19 vaccines with twitter data in the united states: spatiotemporal perspective. *Journal of Medical Internet Research*, page e30854, 2021.

[47] Gerardo Iñiguez, Carlos Pineda, Carlos Gershenson, and Albert-László Barabási. Dynamics of ranking. *Nature communications*, 13(1):1–7, 2022.

[48] Hikmat Ullah Khan, Shumaila Nasir, Kishwar Nasim, Danial Shabbir, and Ahsan Mahmood. Twitter trends: A ranking algorithm analysis on real time data. *Expert Systems with Applications*, 164:113990, 2021.

[49] J. Ko, H.W. Kwon, H.S. Kim, K. Lee, and M.Y. Choi. Model for twitter dynamics: Public attention and time series of tweeting. *Physica A*, 404:141–149, 2014.

[50] Manya Koetse. An introduction to Sina Weibo: Background and status quo. https://www.whatsonweibo.com/sinaweibo/. Accessed December 2, 2020.

[51] Qingchao Kong, Wenji Mao, Guandan Chen, and Daniel Zeng. Exploring trends and patterns of popularity stage evolution in social media. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 50(10):3817–3827, 2018.

[52] K. Lee, D. Palsetia, R. Narayanan, M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. In M. Spiliopoulou and et al., editors, *Proceedings of the 11th IEEE International Conference on Data Mining Workshops, Vancouver, Canada*, pages 251–258. IEEE Computer Society, Los Alamitos, 2011.

[53] J. Lehmann, B. Gonçalves, J.J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in Twitter. In *Proceedings of the 21st international conference on World Wide Web (WWW)*, pages 251–260, 2007.

[54] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 251–260, 2012.

[55] Cuilian Li, Li Jia Chen, Xueyu Chen, Mingzhi Zhang, Chi Pui Pang, and Haoyu Chen. Retrospective analysis of the possibility of predicting the covid-19 outbreak from internet searches and social media data, china, 2020. *Euro Surveill*, 25(10), 2020.

[56] Mengying Li. Promote diligently and censor politely: how sina weibo intervenes in online activism in china. *Information, Communication & Society*, pages 1–16, 2021.

[57] Xiaoya Li, Mingxin Zhou, Jiawei Wu, Arianna Yuan, Fei Wu, and Jiwei Li. Analyzing covid-19 on online social media: Trends, sentiments and emotions. *arXiv preprint arXiv:2005.14464*, 2020.

[58] H. Liang, G. Lu, and N. Xu. Analyzing user influence of microblog. In *2012 IEEE Fifth International Conference on Advanced Computational Intelligence (ICACI)*, pages 15–22, Oct 2012.

[59] Kui Liu, Li Li, Tao Jiang, Bin Chen, Zhenggang Jiang, Zhengting Wang, Yongdi Chen, Jianmin Jiang, and Hua Gu. Chinese public attention to the outbreak of ebola in west africa: Evidence from the online big data platform. *Int J Environ Res Public Health*, 13(8):780, 2016.

[60] Tieying Liu, Yang Zhong, and Kai Chen. Interdisciplinary study on popularity prediction of social classified hot online events in china. *Telematics and Informatics*, 34(3):755–764, 2017.

[61] Wang Liu, Li Xin, Liao Liangchuang, and Liu Li. A momentum theory for hot topic life-cycle: A case study of hot hashtag emerging in twitter. In *International Journal of Computers Communications and Control*, volume 11, pages 734–746, 10 2016.

[62] Christian E Lopez, Malolan Vasu, and Caleb Gallemore. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset. *arXiv preprint arXiv:2003.10359*, 2020.

[63] Haixin Ma, Weining Qian, Fan Xia, Xiaofeng He, Jun Xu, and Aoying Zhou. Towards modeling popularity of microblogs. *Frontiers of Computer Science*, 7(2):171–184, 2013.

[64] Zongyang Ma, Aixin Sun, and Gao Cong. On predicting the popularity of newly emerging hashtags in t witter. *Journal of the American Society for Information Science and Technology*, 64(7):1399–1410, 2013.

[65] José A Morales, Ewan Colman, Sergio Sánchez, Fernanda Sánchez-Puig, Carlos Pineda, Gerardo Iñiguez, Germinal Cocho, Jorge Flores, and Carlos Gershenson. Rank dynamics of word usage at multiple scales. *Frontiers in Physics*, page 45, 2018.

[66] José A Morales, Sergio Sánchez, Jorge Flores, Carlos Pineda, Carlos Gershenson, Germinal Cocho, Jerónimo Zizumbo, Rosalío F Rodríguez, and Gerardo Iñiguez. Generic temporal features of performance rankings in sports and games. *EPJ Data Science*, 5:1–16, 2016.

[67] José A. Morales, Ewan Colman, Sergio Sánchez, Fernanda Sánchez-Puig, Carlos Pineda, Gerardo Iñiguez, Germinal Cocho, Jorge Flores, and Carlos Gershenson. Rank dynamics of word usage at multiple scales. *Frontiers in Physics*, 6:45, 2018.

[68] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[69] Joyce Y. M. Nip and King wa Fu. Networked framing between source posts and their reposts: an analysis of public opinion on China's microblogs. *Information, Communication and Society*, 19:1127–1149, 2016.

[70] National Health Commission of People's Republic of China. National health commission of people's republic of china. http://www.nhc.gov.cn/xcs/xxgzbd/gzbd_index.shtml. Accessed December 2, 2020.

[71] Tai-Quan Peng, Guodao Sun, and Yingcai Wu. Interplay between public attention and public emotion toward multiple social issues on twitter. *PloS one*, 12(1):e0167896, 2017.

[72] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7):770–780, 2020.

[73] Nargis Pervin, Tuan Quang Phan, Anindya Datta, Hideaki Takeda, and Fujio Toriumi. Hashtag popularity on twitter: Analyzing co-occurrence of multiple hashtags. In *International Conference on Social Computing and Social Media*, pages 169–182. Springer, 2015.

[74] CA Piña-García and A Espinoza. Coordinated campaigns on twitter during the coronavirus health crisis in mexico. *Tapuya: Latin American Science, Technology and Society*, page 2035935, 2022.

[75] Lei Qin, Qiang Sun, Yidan Wang, Ke-Fei Wu, Mingchih Chen, Ben-Chang Shia, and Szu-Yuan Wu. Prediction of number of cases of 2019 novel coronavirus (covid-19) using social media search index. *Int J Environ Res Public Health*, 17(7):2365, 2020.

[76] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252, 2011.

[77] Everett M Rogers. *Diffusion of Innovations*. The Free Press, New York, 2010.

[78] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 18–33. Springer, 2011.

[79] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704, 2011.

[80] W. Russell Neuman, L. Guggenheim, S. Mo Jang, and S. Y. Bae. The dynamics of public attention: Agenda-setting theory meets big data. *Journal of Communication*, 64:193–214, 2014.

[81] Islamuddin Sajid. China reports 99 new virus cases, majority imported. https://www.aa.com.tr/en/asia-pacific/china-reports-99-new-virus-cases-majority-imported/1801667. Accessed December 2, 2020.

[82] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[83] Dong-Woo Seo and Soo-Yong Shin. Methods using social media and search queries to predict infectious disease outbreaks. *Healthc Inform Res*, 23(4):343–348, 2017.

[84] Weibo Customer Service. Common questions on the rules of real-time hot-search-list, hot-message-list and hot-topic-list. https://www.weibo.com/ttarticle/p/show?id=2309404007731978739654. Accessed July 25, 2022.

[85] Parul Sharma and Teng-Sheng Moh. Prediction of indian election using sentiment analysis on hindi twitter. In *2016 IEEE international conference on big data (big data)*, pages 1966–1971. IEEE, 2016.

[86] Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. The use of twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS one*, 6:e19467, 2011.

[87] Sitabhra Sinha and Raj Kumar Pan. How a hit is born: The emergence of popularity from the dynamics of collective choice. *Econophysics and Sociophysics: Trends and Perspectives*, 2:417–447, 2006.

[88] Michał Skuza and Andrzej Romanowski. Sentiment analysis of twitter data within big data distributed environment for stock prediction. In *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1349–1354. IEEE, 2015.

[89] Reuters Staff. China punishes microblog platform weibo for interfering with communication, 2020.

[90] Grant Stafford and Louis Lei Yu. An evaluation of the effect of spam on twitter trending topics. In *2013 International Conference on Social Computing*, pages 373–378. IEEE, 2013.

[91] Sarah Tam and Debbie Hahm. Exploring coronavirus twitter trends. https://towardsdatascience.com/coronavirus-twitter-trends-d32fed5a027e. Accessed August 8, 2020., 2020. Towards Data Science, Inc.

[92] Weizhen Tan and Holly Ellyatt. China confirms 15152 new coronavirus cases, 254 additional deaths. https://www.cnbc.com/2020/02/13/coronavirus-latest-updates-china-hubei.html. Accessed December 2, 2020.

[93] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar,

and Eli Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6, 2020.

[94] Io Taxidou and Peter M. Fischer. Online analysis of information diffusion in twitter. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 1313–1318, New York, NY, USA, 2014. ACM.

[95] George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265, 1992.

[96] Marijn ten Thij, Tanneke Ouboter, Daniël Worm, Nelly Litvak, Hans van den Berg, and Sandjai Bhulai. Modelling of trends in twitter using retweet graph dynamics. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 132–147. Springer, 2014.

[97] Lai Lin Thomala. Number of monthly active users of sina weibo from 1st quarter of 2018 to 3rd quarter of 2021. *statista* https://www.statista.com/statistics/795303/china-mau-of-sina-weibo/, 2021.

[98] Jingrong Tong and Landong Zuo. Weibo communication and government legitimacy in China: a computer-assisted analysis of Weibo messages on two 'mass incidents'. *Information, Communication and Society*, 17:66–85, 2014.

[99] Trends24. Twitter trend. https://trends24.in/about. Accessed August 8, 2020.

[100] Oren Tsur and Ari Rappoport. What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643–652, 2012.

[101] Twitter. How to use hashtags. https://help.twitter.com/en/using-twitter/how-to-use-hashtags#:~:text=People%20use%20the%20hashtag%20symbol,included%20anywhere%20in%20a%20Tweet.. Accessed Sep. 2, 2022.

[102] Twitter. Twitter micoroblog and social network service. https://about.twitter.com/. Accessed December 2, 2020.

[103] Twitter. Twitter: Research and experiments. https://help.twitter.com/en/rules-and-policies#research-and-experiments. Accessed December 2, 2020.

[104] Twitter. Twitter trends faq. https://help.twitter.com/en/using-twitter/twitter-trending-faqs. Accessed Sep. 2, 2022.

[105] Twitter. Covid-19 stream. https://developer.twitter.com/en/docs/labs/covid19-stream/overview. Accessed August 7, 2020., 2020. Twitter, Inc.

[106] TwitterIR. Q4 and fiscal year 2021 letter to shareholders. *Twitter* https://s22.q4cdn.com/826641620/files/doc_financials/2021/q4/Final-Q4'21-Shareholder-letter.pdf, 2022.

[107] Liza G. G. van Lent, Hande Sungur, Florian A. Kunneman, Bob van de Velde, and Enny Das. Too far to care? measuring public attention and fear for ebola using twitter. *J. Med. Internet Res.*, 19:e193, 2017.

[108] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[109] Juha Antero Vuori and Lauri Paltemaa. The lexicon of fear: Chinese internet control practice in sina weibo microblog censorship. *Surveillance & society*, 13(3/4):400–421, 2015.

[110] Weibo Administrator. Weibo hot search regulation rules, 2021.

[111] Wikipedia. Kernel density estimation. *Wikipedia* https://en.wikipedia.org/wiki/Kernel_density_estimation.

[112] Wikipedia. Savitzky–golay filter. https://en.wikipedia.org/wiki/Savitzky-Golay_filter. Accessed December 2, 2020.

[113] Wikipedia. Spline interpolation. *Wikipedia* https://en.wikipedia.org/wiki/Spline_interpolation.

[114] Wikipedia. Time in china. *Wikipedia* https://en.wikipedia.org/wiki/Time_in_China.

[115] Wikipedia. Kris wu sex scandal. *Wikipedia* https://en.wikipedia.org/wiki/Kris_Wu_sex_scandal, 2021.

[116] Wikipedia. Main data of the seventh national population census. *National Bureau of Statistics of China* http://www.stats.gov.cn/english/PressRelease/202105/t20210510_1817185.html, 2021.

[117] Fang Wu and Bernardo A Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

[118] Lianren Wu, Jiayin Qi, Nan Shi, Jinjie Li, and Qiang Yan. Revealing the relationship of topics popularity and bursty human activity patterns in social temporal networks. *Physica A: Statistical Mechanics and its Applications*, 588:126568, 2022.

[119] Zhao Y, Cheng S, Yu X, and Xu H. Chinese public's attention to the covid-19 epidemic on social media: Observational descriptive study. *Journal of Medical Internet Research*, 22:e18825, 2020.

[120] Zhu Y, Fu KW, Grépin KA, Liang H, and Fung IC. Limited early warnings and public attention to coronavirus disease 2019 in china, january-february, 2020: A longitudinal cohort of randomly sampled weibo users. *Disaster Med Public Health Prep*, 1-4, 2020.

[121] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608. IEEE, 2010.

[122] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186, 2011.

[123] Hai Yu, Ying Hu, and Peng Shi. A prediction method of peak time popularity based on twitter hashtags. *IEEE Access*, 8:61453–61461, 2020.

[124] Louis Yu, Sitaram Asur, and Bernardo A Huberman. What trends in chinese social media. *arXiv preprint arXiv:1107.3522*, 2011.

[125] Louis Lei Yu, Sitaram Asur, and Bernardo A Huberman. Artificial inflation: the real story of trends and trend-setters in sina weibo. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 514–519. IEEE, 2012.

[126] Louis Lei Yu, Sitaram Asur, and Bernardo A Huberman. Trend dynamics and attention in chinese social media. *American Behavioral Scientist*, 59(9):1142–1156, 2015.

[127] S. Yuan, Z. Tao, T. Zhu, and S. Bai. Realtime online hot topics prediction in sina weibo for news earlier report. In *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, pages 599–605, March 2017.

[128] Emma Zavarrone, Maria Gabriella Grassia, Marina Marino, Rasanna Cataldo, Rocco Mazza, and Nicola Canestrari. Co. me. ta–covid-19 media textual analysis. a dashboard for media monitoring. *arXiv preprint arXiv:2004.07742*, 2020.

[129] Leihan Zhang, Ke Xu, and Jichang Zhao. Sleeping beauties in meme diffusion. *Scientometrics*, 112(1):383–402, 2017.

[130] Leihan Zhang, Jichang Zhao, and Ke Xu. Who creates trends in online social media: The crowd or opinion leaders? *Journal of Computer-Mediated Communication*, 21(1):1–16, 2016.

[131] Yanshuang Zhang. Microblogging and its implications to chinese civil society and the urban public sphere: A case study of sina weibo. *PhD Thesis at the University of Queensland*, 2016.

[132] Yubao Zhang, Xin Ruan, Haining Wang, Hui Wang, and Su He. Twitter trends manipulation: a first look inside the security of twitter trending. *IEEE Transactions on Information Forensics and Security*, 12(1):144–156, 2016.

[133] Zizhu Zhang, Bing Li, Weiliang Zhao, and Jian Yang. A study on the retweeting behaviour of marketing microblogs with high retweets in sina weibo. In *2015 Third International Conference on Advanced Cloud and Big Data*, pages 20–27. IEEE, 2015.

[134] Juanjuan Zhao, Weili Wu, Xiaolong Zhang, Yan Qiang, Tao Liu, and Lidong Wu. A short-term trend prediction model of topic over sina weibo dataset. *Journal of Combinatorial Optimization*, 28(3):613–625, 2014.

[135] Z. Zhao, J. Sun, L. Yao, X. Wang, J. Chu, H. Liu, and G. Yu. Modeling chinese microblogs with five ws for topic hashtags extraction. *Tsinghua Science and Technology*, 22(2):135–148, April 2017.

[136] Yang Zhou, Lei Zhang, Xiaoqian Liu, Zhen Zhang, Shuotian Bai, and Tingshao Zhu. Predicting the trends of social events on chinese social media. *Cyberpsychology, Behavior, and Social Networking*, 20(9):533–539, 2017.

[137] A. Zubiaga, D. Spina, V. Fresno, and R. Martínez. Classifying trending topics: A typology of conversation triggers on twitter. In B. Berendt, A. de Vries, and W. Fan, editors, *Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, UK*. ACM, New York, 2011.