"Born in Rome" or "Sleeping Beauty": Emergence of hashtag popularity on a microblogging site

Hao Cui and János Kertész

ABSTRACT

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.

Introduction

Microblogging sites are online social networking platforms where users interact with each other through activities such as post, repost, comment, reply, mention or like. 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, the number of reposts/views at a time, straightforwardly the displayed list of popular items by social media platform providers. 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. Among the flooding 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 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 and, at the same time, they trigger collective awareness 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 or from manipulation by companies and organizations.

It is of primary interest to understand why some hashtags get to the top rank list and others not. What makes the history of those popular hashtags specific? Understanding the emergence of these popular topics plays important role in marketing, governance and trend predictions in real world. Researchers on Twitter information diffusion have observed variation in the spread of online information across topics, shown the effects of the content, semantic characterization and the co-occurrence of hashtags on popularity. Approaches have been introduced to predict the popularity peaks of the new hashtags on Twitter from the perspectives of machine learning, and taking the context of Twitter social network. Previous study on Twitter trending topics has shown retweets by other users are more important than the number of followers in determining trends. Some studies modeled user behaviors to capture the emergence of Twitter trending topics based on characteristics of the retweet graphs. A recent study on Twitter trending topics has investigated real-time Twitter Trends detection along with the ranking of the top terms.

Twitter is a worldwide service, with 217 million monetizable daily active users as of the fourth quarter of 2021. Its Chinese counterpart, Sina Weibo, the most popular microblogging service in China, has 248 million daily active users and 573 million monthly active users as of the third quarter of 2021, who generate and propagate information in the whole Weibo system. Accordingly, Sina Weibo has become a popular tool for the Chinese public to look for information. The topics covered by hashtags which occur on the HSL are very diverse, to name a few, social events, TV programs, celebrities, entertainment, health and politics. During the time of COVID-19, the Weibo HSL played an important role in keeping people aware of the COVID-related news and updates. Since the HSL has an advertising effect on the hashtags to the public, it is natural that
celebrities or companies would desire a position on the HSL or disappear from the HSL in the case of a negative influence. Studies have also identified online censorship control practices and the possibility of algorithmic intervention on Sina Weibo.

Although Sina Weibo has received less academic attention than Twitter, while it has overtaken Twitter in terms of the number of users, research on its popularity appeared soon after its launch. The attention of the researchers studying popularity in Sina Weibo turned gradually to the HSL, earlier they have presented evolution analysis of trending topics, long-term variation of popularity, prediction of hot topics based on content quality and structural characteristics of early adopters, as well as bursty human activity patterns.

Here we study the dynamics of the popularity of hashtags on the microblogging site Sina Weibo and focus on their history before they get to 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.

The paper is organized as follows: In the next section we describe the results of our investigations, followed by a discussion. The paper terminates with a section on methods and data. A Supplementary Information with videos on the repost network evolution complements the paper.

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.

Role of birth time

![Figure 1. 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 observed circadian pattern shows that practically no new hashtags appear on the HSL in a certain time interval during the night, with an average size of 7.18 hours, and a standard deviation of 0.85 hours.](image)

According to its size, China should have five geographic time zones 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 whose geographical time zones are similar and the population.
accounts for 65.8 percent of the national population\textsuperscript{26}. 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. 1. Zooming into the figure as seen in the inset in Fig. 1, 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\textsuperscript{27} that the selection of hashtags to the HSL follows an automated procedure just based on a formula (see Section Methods).

![Figure 2.](image)

**Figure 2.** 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\textsuperscript{28}. (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) Histogram of the HSL time. Section 1⃝ represents the category "Born in Rome" and section 2⃝ represents "Sleeping Beauty". (D) Parameterized probability density functions of the HSL time by different time intervals of birth time of the day, using kernel density estimation (KDE)\textsuperscript{29}.

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 User Development Report\textsuperscript{28}, we show in Fig. 2A 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 2B 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 2B 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. 1 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. 2B, 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 2D describes the distribution of $t_{HSL}$ of the hashtags in Fig. 2B 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 2C shows the distribution of $t_{HSL}$, with a rapid decrease till 8 hours, followed by a slower and longer decrease afterwards.

**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 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, "央视新闻" ("CCTV News", 126M followers), "人民日报" ("People’s Daily", 145M followers), "头条新闻" ("Headline News", 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 shown in Fig. 2C, we consider the hashtags whose time needed from birth till HSL within 8 hours to be in the category "Born in Rome". We have also seen some hashtags with very few (re)posts prior to the HSL, for example, #美国从英法订购1亿剂新冠疫苗# (#US orders 100 million doses of coronavirus vaccine from UK and France#), 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. 2C, the hashtags in the "Sleeping Beauty" category are those whose $t_{HSL}$ is greater than 30 hours. When it comes to the hashtag content, as shown in Fig. 3, "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.

**Relation with repost network dynamics**

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 shows the ratio of different link growth pattern dynamics of the total network and the final giant component in the two categories (for examples see Fig. 5). As shown in Fig. 3, 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 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**

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. 4A, 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. 4D, in minute resolution. The network increment per minute shows a fast (exponential) decay and then a slower one as time goes on. In Fig. 4B and Fig. 4C, we show examples of hashtags
Figure 3. 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.

Figure 4. Comparison of repost network growth patterns of failed hashtags born in the super-hub "央视新闻" ("CCTV News") 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", #做10000多台手术老医生惜别手术台# (#Doing more than 10000 operations, the old doctor bid farewell to the operating table#). (B) An example of a hashtag from the category "Born in Rome", #乘风破浪的姐姐三公分组# (#Sisters Who Make Waves (a variety show) grouping of the third public performance#). (C) An example of a hashtag from the category "Sleeping Beauty", #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.
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 4E and Fig. 4F 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 4D. 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.

Discussion

To analyze the emergence of hashtag popularity on the Chinese microblogging website Sina Weibo, we have studied the 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 the final giant component for successful hashtags. We have pointed out the role of hubs in the process and identified two extreme types of popularity emergence mechanisms for successful hashtags: That of the "Born in Room" and of the "Sleeping Beauty".

Our studies indicate that in order to become popular enough to get to the HSL, a hashtag should be posted by a super-hub or several smaller hubs together. These are, of course, not sufficient conditions. 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. 2A, 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. 2B suggests, indicating the disadvantage of hashtags born in that time period in achieving success.

Sina Weibo Hot Search List has an advertising effect on the hashtags by substantially increasing their visibility to the whole public. In many cases, the repost network evolution of hashtags related, e.g., to celebrities fall into a pattern of "collective efforts" whose popularity accounts for several smaller hubs working together. These smaller hubs are usually marketing nodes, emphasizing the importance of social capital in making hashtags related to stars popular enough to enter the HSL. Of course, even hashtags from super-hubs could fail to get to the list, let alone those from normal users with only a few followers. For hashtags of 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. For those hashtags originating from super-hubs but failed, the format or the content as well as the emotions that arouse to the online users could also matter. There are also cases where a post contains multiple hashtags related to the content of the post. As time goes on, the hashtags’ propagation trajectories split, one of the hashtags become more popular and the others vanish.

Understanding the mechanism of the emergence of hashtag popularity and the importance of timing could contribute to marketing and maximizing the spreading efficiency by playing with these factors. More importantly, 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. The prime responsibility is carried by the platform provider. It should make the algorithm fair to capture real hot topics that gain true attention from the public. In fact 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.

We mention that our approach has a broader applicability than just for the 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. 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.
Methods

Hot Search List (HSL)
Sina Weibo is the most popular microblogging site in mainland China, where Twitter and other Western online social media are blocked. It 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. Sina Weibo is a major vehicle of self expression especially for young Chinese people and a forum for social movements.

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 other microblogging sites, Sina Weibo also creates a real time ranking of the 50 most popular hashtags to inform users. As the name of the ranking list indicates, the Hot Search List (HSL) is based to a large extent on the search activity related to the hashtags, however, the concrete algorithm has been unknown and has been target of criticism. The latter can be understood as getting to the HSL not only informs users about popularity but also boosts 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. The corresponding formula is as follows:

$$H = (S_H + D_H + R_H) \times I_H,$$

where $S_H$ is search hotness, referring to the search volume, including manual input search and click-and-jump search, $D_H$ refers to the amount of discussion, including original posting and re-posting, $R_H$ is the volume of readings in the spreading process of the hashtag, and $I_H$ 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 and the intervention in other cases (redundancy, serious inaccurate information as identified by government departments of content inducing severe conflicts). In fact, indication of intervention have been noticed on the HSL by statistical analysis.

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. The temporal directed repost network consists of users as nodes and reposts as links between them, pointing towards the user who reposts. Since all reposts have timestamps, the evolution of the repost network can be followed and can be traced back to the source(s) of the hashtag. The timestamp of the first post containing a hashtag is the birth time of that hashtag. We focus on the whole repost network and the largest connected component (LCC), which consists of the largest number of connected nodes in the whole hashtag repost network. When studying the network size growth, we disregard the directedness of the links. As the repost network evolves, the giant component 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. We compare, analyze, and categorize these patterns.

![Shape examples of stepwise and smooth patterns of repost network cumulative link growth trajectories.](image)

**Figure 5.** Shape examples of stepwise and smooth patterns of repost network cumulative link growth trajectories.

Classification of link growth trajectories
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...
Figure 6. 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:** $\tilde{f} = f'(t) - \bar{f}$, 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 $(\tilde{f} * g)(t)$ where $g(t)$ 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 calculate the convolution $(\tilde{f} * g)(t)$ using the convolve method in the numpy\textsuperscript{33} Python module, with the mode parameter equals ‘valid’. We find all local maxima by comparing with neighboring values in the convolved series $(\tilde{f} * g)(t)$ using the peak detection function find_peaks in the scipy.signal\textsuperscript{34} Python module. The principle of the classifier is demonstrated in Fig.6. 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. 4E and Fig. 4F, the resize was done by using TimeSeriesResampler from tslearn\textsuperscript{35} Python package, with the method of spline interpolation\textsuperscript{36}.

**Data and availability**

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. 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 584 hashtags in this category and crawled all their reposts. In addition, we produced an equal-sized random sample from the "Born in Rome" category. 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 long available on Weibo ever since he was arrested by police due to several crimes\textsuperscript{37}. 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 article are available in the github repository, https://github.com/cuihaosabrina/Emergence_Popularity_Sina_Weibo.

**References**

1. Zhang, L., Zhao, J. & Xu, K. Who creates trends in online social media: The crowd or opinion leaders? J. Comput. Commun. 21, 1–16 (2016).
2. Bao, P., Shen, H.-W., Huang, J. & Cheng, X.-Q. Popularity prediction in microblogging network: a case study on sina weibo. In *Proceedings of the 22nd international conference on world wide web*, 177–178 (2013).

3. Ma, H. *et al.* Towards modeling popularity of microblogs. *Front. Comput. Sci.* 7, 171–184 (2013).

4. Annamoradnejad, I. & Habibi, J. A comprehensive analysis of twitter trending topics. In *International Conference on Web Research (ICWR)*, 22–27 (2019).

5. Cui, H. & Kertész, J. Attention dynamics on the chinese social media sina weibo during the covid-19 pandemic. *EPJ data science* 10, 8 (2021).

6. Asur, S., Huberman, B. A., Szabo, G. & Wang, C. Trends in social media: Persistence and decay. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, 434–437 (2011).

7. Thij, M. *et al.* Modelling of trends in twitter using retweet graph dynamics. In *International Workshop on Algorithms and Models for the Web-Graph*, 132–147 (Springer, 2014).

8. Ratkiewicz, J. *et al.* Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, 249–252 (2011).

9. Ma, H. *et al.* Towards modeling popularity of microblogs. *Front. Comput. Sci.* 7, 171–184 (2013).

10. Romero, D. M., Meeder, B. & Kleinberg, J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 21st international conference on World wide web*, 695–704 (2011).

11. Tsur, O. & Rappoport, A. 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*, 643–652 (2012).

12. Lehmann, J., Gonçalves, B., Ramasco, J. J. & Cattuto, C. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, 251–260 (2012).

13. Pervin, N., Phan, T. Q., Datta, A., Takeda, H. & Toriumi, F. Hashtag popularity on twitter: Analyzing co-occurrence of multiple hashtags. In *International Conference on Social Computing and Social Media*, 169–182 (Springer, 2015).

14. Ma, Z., Sun, A. & Cong, G. On predicting the popularity of newly emerging hashtags in twitter. *J. Am. Soc. for Inf. Sci. Technol.* 64, 1399–1410 (2013).

15. Yu, H., Hu, Y. & Shi, P. A prediction method of peak time popularity based on twitter hashtags. *IEEE Access* 8, 61453–61461 (2020).

16. Khan, H. U., Nasir, S., Nasim, K., Shabbir, D. & Mahmood, A. Twitter trends: A ranking algorithm analysis on real time data. *Expert. Syst. with Appl.* 164, 113990 (2021).

17. 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).

18. 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).

19. Staff, R. China punishes microblog platform weibo for interfering with communication. *Reuters* https://www.reuters.com/article/us-china-censorship-weibo-idUSKBN23H1J2 (2020).

20. Chen, L., Zhang, C. & Wilson, C. Tweeting under pressure: analyzing trending topics and evolving word choice on sina weibo. In *Proceedings of the first ACM conference on Online social networks*, 89–100 (2013).

21. Vuori, J. A. & Paltemaa, L. The lexicon of fear: Chinese internet control practice in sina weibo microblog censorship. *Surveillance & society* 13, 400–421 (2015).

22. Yu, L., Asur, S. & Huberman, B. A. Artificial inflation: The true story of trends in sina weibo. *arXiv preprint arXiv:1202.0327* (2012).

23. Wu, L., Qi, J., Shi, N., Li, J. & Yan, Q. Revealing the relationship of topics popularity and bursty human activity patterns in social temporal networks. *Phys. A: Stat. Mech. its Appl.* 588, 126568 (2022).

24. Time in china. *Wikipedia* https://en.wikipedia.org/wiki/Time_in_China.

25. Center, S. W. D. Weibo 2020 user development report. *Weibo Report*. https://data.weibo.com/report/reportDetail?id=456 (2021).

26. 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).
27. Administrator, W. Weiibo hot search regulation rules. Sina Weiibo https://weibo.com/1934183965/KuKyPkp8Y?type=repost (2021).
28. Center, S. W. D. 2015 weibo user development report. Weibo Report. https://data.weibo.com/report/reportDetail?id=333 (2016).
29. Kernel density estimation. Wikipedia https://en.wikipedia.org/wiki/Kernel_density_estimation.
30. Rogers, E. M. Diffusion of Innovations (The Free Press, New York, 2010).
31. Zhang, Y. Microblogging and its implications to chinese civil society and the urban public sphere: A case study of sina weibo. PhD Thesis at Univ. Qld. (2016).
32. Hewitt, D. Weibo brings changes to china. BBC News https://www.bbc.com/news/magazine-18773111 (2012).
33. Harris, C. R. et al. Array programming with NumPy. Nature 585, 357–362, DOI: 10.1038/s41586-020-2649-2 (2020).
34. Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat. Methods 17, 261–272, DOI: 10.1038/s41592-019-0686-2 (2020).
35. Tavenard, R. et al. Tslearn, a machine learning toolkit for time series data. J. Mach. Learn. Res. 21, 1–6 (2020).
36. Spline interpolation. Wikipedia https://en.wikipedia.org/wiki/Spline_interpolation.
37. Kris wu sex scandal. Wikipedia https://en.wikipedia.org/wiki/Kris_Wu_sex_scandal (2021).

Acknowledgements
We acknowledge support supported by the European Union – Horizon 2020 Program under the scheme “INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities”, Grant Agreement n.871042, “SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics” and SAI enabled by FWF (I 5205-N) within the EU CHIST-ERA program.

Author contributions
HC and JK conceived the idea and designed the study. HC carried out the data collection, HC and JK did the data analysis. Both authors drafted paper, read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information

"Born in Rome" or "Sleeping Beauty": Emergence of hashtag popularity on a microblogging site

Hao Cui\(^1\) and János Kertész\(^1\)

\(^1\)Department of Network and Data Science, Central European University, Quellenstrasse 51, A-1100 Vienna, Austria

*Correspondence: kerteszj@ceu.edu

SI1. Evolution of repost networks before getting to the HSL (movies)

Here we show how the repost networks of different categories of hashtags evolve before they get to the HSL.

https://drive.google.com/drive/folders/1JULm8eNswUSOY4PC-cjTtvP4ocJzvaM-

SI2. Hashtag categorization

We have classified the hashtags into the following categories: 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. The following table contains example hashtags and their translation.

| Social                          | Stars                          | International                      | Others                                      |
|---------------------------------|--------------------------------|------------------------------------|---------------------------------------------|
| #31 省份养老金已全部上涨#     | #梁正贤空间管理大师#            | #白宫内部咖啡厅员工确诊新冠#       | #假如所有生物都变成猫#                      |
| #Pensions in 31 provinces have all risen# | #Liang Zhengxian Master of Space Management# | #Employees in White House cafes diagnosed with covid# | #If all creatures became cats#               |
| #76 岁老人救起 200           | #宋丹丹 60 岁生日宴#            | #阿联酋将与以色列实现关系全面正常化# | #我的青春疼痛#                              |
|                                |                               | #My Youth Pain#                     |                                             |
### 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.

| 多斤溺水者# | #76-year-old man rescues 200-pound drowning man# |
|-----------|-------------------------------------------------|
| #山西警方破获 30 年前故意杀人案# | #周杰伦晒 20 年前旧照# |
| # Shanxi police cracked down on intentional homicide 30 years ago# | #Jay Chou’s old photos from 20 years ago# |
| #UAE to fully normalize relations with Israel# | #澳大利亚墨尔本将实施宵禁# |
| #Melbourne, Australia will implement curfew# | #如何优雅地背着被子去学校# |
| #How to gracefully carry a quilt to school# |

*Figure 1. Examples of hashtag “rebirth”.*

- As shown in Fig 1A, the hashtag #杨幂跳无价之姐# (#Yang Mi dances Priceless Sister#) 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 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 1B, the hashtag #9 号台风美莎克# (#Typhoon Maysak No. 9 #) 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 1C, the hashtag #大连新增3例确诊# (#3 new infected cases confirmed in Dalian#) first appeared on the HSL on 2020 August 2 at 08:17. It was born on 2020 July 23 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 $t_{\text{HSL}} < 5$ days.