popularity prediction. IEEE Transactions on Multimedia, 2013, 15(6): 1255–1267
21. Vallet D, Berkovsky S, Ardon S, et al. Characterizing and predicting viral-and-popular video content. Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15), Oct 18–23, 2015, Melbourne, Australia. New York, NY, USA: ACM, 2015: 1591–1600
22. Pinto H, Almeida J M, Gonçalves M A. Using early view patterns to predict the popularity of YouTube videos. Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13), Feb 4–8, 2013, Rome, Italy. New York, NY, USA: ACM, 2013: 365–374
23. Tatar A, Leguay J, Antoniadis P, et al. Predicting the popularity of online articles based on user comments. Proceedings of the 2011 International Conference on Web Intelligence, Mining and Semantics (WIMS'11), May 25–27, 2011, Sogndal, Norway. New York, NY, USA: ACM, 2011: Article 67
24. Chen H Q, Zhong X F, Sun J, et al. Online prediction algorithm of the ews' popularity for wireless cellular pushing. Proceedings of the 2015 IEEE/CIC International Conference on Communications in China (ICCC'15), Nov 2–4, 2015, Shenzhen, China. Piscataway, NJ, USA: IEEE, 2015: 5p
25. Tatar A, Antoniadis P, De Amorim M D, et al. From popularity prediction to ranking online news. Social Network Analysis and Mining, 2014, 4: Article 174
26. Lin Y J, Yeh M Y, Chiu F Y, et al. Predicting popularity of articles on bulletin board system. Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp'16), Jan 18–20, 2016, Hong Kong, China. Piscataway, NJ, USA: IEEE, 2016: 169–176
27. Ma C S, Yan Z S, Chen C W. Forecasting initial popularity of just-uploaded user-generated videos. Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP'16), Sept 25–28, 2016, Phoenix, AZ, USA. Piscataway, NJ, USA: IEEE, 2016: 474–478
28. Kong Q C, Mao W J, Liu C Y. Popularity prediction based on interactions of online contents. Proceedings of the 4th International Conference on Cloud Computing and Intelligence Systems (CCIS'16), Aug 17–19, 2016, Beijing, China. Piscataway, NJ, USA: IEEE, 2016: 5p
29. Lemahieu R, Van Canneyt S, De Boom C, et al. Optimizing the popularity of twitter messages through user categories. Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW'15), Nov 14–17, 2015, Atlantic City, NJ, USA. Piscataway, NJ, USA: IEEE, 2015: 1396–1401

tions of the ACM, 2010, 53(8): 80–88
3. Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences, 2008, 105(41): 15649–15653
4. Chatzopoulou G, Sheng C, Faloutsos M. A first step towards understanding popularity in YouTube. Proceedings of the 2010 IEEE Conference on Computer Communications Workshops, Mar 15–19, 2010, San Diego, CA, USA. Piscataway, NJ, USA: IEEE, 2010: 6p
5. Gill P, Arlitt M, Li Z P, et al. YouTube traffic characterization: a view from the edge. Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC'07), Oct 24–26, 2007, San Diego, CA, USA. New York, NY, USA: ACM, 2007: 15–28
6. Zink M, Suh K, Gu Y, et al. Characteristics of YouTube network traffic at a campus network–measurements, models, and implications. Computer Networks: The International Journal of Computer and Telecommunications Networking, 2009, 53(4): 501–514
7. Cha M, Kwak H, Rodriguez P, et al. Analyzing the video popularity characteristics of large-scale user generated content systems. IEEE/ACM Transactions on Networking, 2009, 17(5): 1357–1370
8. Tan X Y, Guo Y C, Chen Y S, et al. Characterizing user popularity preference in a large-scale online video streaming system. Proceedings of the 6th International Conference on Wireless, Mobile and Multi-Media, Nov 20-23, 2015, Beijing, China. Piscataway, NJ, USA: IEEE, 2015: 246–249
9. Liu W, Zhang G, Chen J, et al. A measurement-based study on application popularity in android and iOS APP stores. Proceedings of the 2015 Workshop on Mobile Big Data, Jun 22–25, 2015, Hangzhou, China. New York, NY, USA: ACM, 2015: 13–18
10. Li C Y, Liu J. Large-scale characterization of comprehensive online video service in mobile network. Proceedings of the 2016 IEEE International Conference on Communications (ICC'16), May 22–7, 2016, Kuala Lumpur, Malaysia. Piscataway, NJ, USA: IEEE, 2016: 7p
11. Cheng X, Dale C, Liu J C. Statistics and social network of YouTube videos. Proceedings of the 16th International Workshop on Quality of Service (IWQoS'08), Jun 2–4, 2008, Enschede, Netherlands. Piscataway, NJ, USA: IEEE, 2008: 229–238.
12. Figueiredo F, Benevenuto F, Almeida J M. The tube over time: characterizing popularity growth of YouTube videos. Proceedings of the 2011 ACM International Conference on Web Search and Data Mining (WSDM'11), Feb 9–12, 2011, Hong Kong, China. New York, NY, USA: ACM, 2008: 745–754
13. Figueiredo F, Almeida J M, Gonçalves M A, et al. On the dynamics of social media popularity: a YouTube case study. ACM Transactions on Internet Technology, 2014, 14(4): Article 24
14. Gonçalves G D, Figueiredo F, Almeida J M, et al. Characterizing scholar popularity: a case study in the computer science research community. Proceedings of the 2014 IEEE/ACM Joint Conference on Digital Libraries, Sept 8–14, 2014, London, UK. Piscataway, NJ, USA: IEEE, 2014: 57–66
15. Yang J, Leskovec J. Patterns of temporal variation in online media. Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11), Feb 9–12, 2011, Hong Kong, China. New York, NY, USA: ACM, 2011: 177–186
16. Borghol Y, Mitra S, Ardon S, et al. Characterizing and modelling popularity of user-generated videos. Performance Evaluation, 2011, 68(11): 1037–1055
17. Wu F, Huberman BA. Novelty and collective attention. Proceedings of the National Academy of Sciences, 2007, 104(45): 17599–17601
18. Ratkiewicz J, Fortunato S, Flammini A, et al. Characterizing and modeling the dynamics of online popularity. Physical Review Letters, 2010, 105(15): Article 158701
19. Cheng X, Liu J C, Dale C. Understanding the characteristics of internet short video sharing: a YouTube-based measurement study. IEEE Transactions on Multimedia, 2013, 15(5): 1184–1194
20. Roy S D, Mei T, Zeng W J, et al. Towards cross-domain learning for social video

further propose the specialized models, which are built based on the popularity evolution patterns. We run regressions with our dataset, and show our models can better describe the early-future popularity relationship than the state-of-the-art model [2]. In fact, based on the log-linear method [2], many general models (that is using the same set of model parameters for all the instances) for online content popularity prediction have been proposed, such as: the constant scaling model [25], the multivariate linear with radial bias function (RBF) model [22], the ordinary least squares (OLS) multivariate linear regression model [14], and etc. We note that it is out of the scope of this paper to build a prediction system and compete all those methods. Instead, our work aims to show that when using the early observations of online content popularity to predict the future popularity, it can obtain great performance improvement to consider the popularity evolution pattern for each individual video and build specialized models for each kind of pattern. However, such method will require the estimation of the popularity evolution pattern for each video. Possible solutions for the pattern prediction should be based on classification techniques and effective features extracted from video/author properties, content/textual information [26], daily popularity measures [22], related video recommendations [27], online contents interactions [28] and etc.

Moreover, we find there are some previous studies that proposed specialized models based on the video category [22,29]. For comparison, we also build such category-specific models (top 10 categories and the rest, total 11 models) using our dataset, and find it can only slightly improve the regression results (overall MRSE is 9.89%). This indicates that the category-specific models are less effective than our popularity-evolution-patternspecific models. This is because the evolution patterns are key impact factors for the view count accumulation, while different kinds of popularity evolution patterns lie widely in the same category of videos. Hence, the video category is a less distinctive aspect for building specialized prediction models.

5. Conclusions

In this paper, we have presented a detailed characterization of the dynamics of online video popularity.

Based on a long-term recent dataset crawled from Youku website, four key aspects have been analyzed: the overall popularity distribution, the individual popularity distribution, the popularity evolution pattern and the early-future popularity relationship. Our measurements shed light on how the popularity of newly uploaded Youku videos distributes and evolves over time, as well as how the evolution patterns impact the linear correlations between early and future view count. The findings uncovered in this paper are crucial and reliable for all interested parties of online video service such as content publishers, service providers, online advertisers and network operators, to improve the service design and video delivery for better user experience.

Acknowledgements This work was supported by the Video Super-Resolution Reconstruction Project (20130005110017).

References

1. Cisco visual networking index: forecast and methodology, 2016–2021. San Jose, CA, USA: Cisco
2. Szabo G, Huberman B A. Predicting the popularity of online content. Communica-

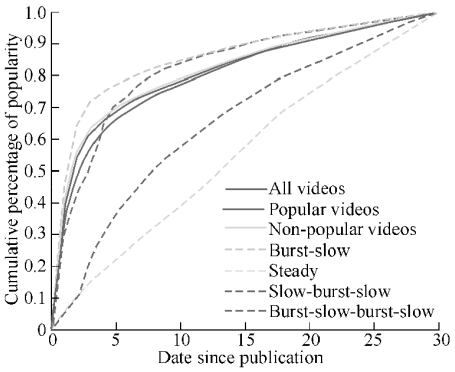


Fig. 10 Cumulative percentage of popularity over the observation period

In our analysis, we consider the differences of early-future popularity relationship across different kinds of videos, and build the specialized log-linear models for the videos with different popularity evolution patterns as follows:

$$\widehat{N_{\mathrm{v}}}(r)=\mathrm{e}^{\ln N_{\mathrm{v}}(i)+\ln \tau(i,r,p)} \tag{5}$$

where $\ln\tau(i, r, p)$ is determined by the videos with same popularity evolution pattern p. We run regressions with our dataset (i = 7 and r = 30, same as in Ref. [2]), and list the estimated model parameters and the mean relative squared error (MRSE) in Table 2. The MRSE of a set of videos V is defined as:

$$E_{\mathrm{MRSE}}=\frac{1}{|V|}\sum_{v\in V}\left(\frac{N_{\mathrm{v}}(r)-\widehat{N_{\mathrm{v}}}(r)}{N_{\mathrm{v}}(r)}\right)^2 \tag{6}$$

For comparison, we also list the results for the general model. It is apparent from the table that the specialized models are more proper for the early-future popularity correlation, with the overall MRSE reducing 24.35% over the general model. And in particular, the specialized models gain large MRSE reductions (55.04%) for the videos with steady popularity evolution pattern.

Table 2 Summary of goodness of fit for different models

Evolution pattern	General model		Specialized models	
	τ	MRSE(%)	τ	MRSE(%)
Overall	1.219 3	10.35	-	7.83
Burst-slow	1.219 3	3.89	1.184 5	3.78
Steady	1.219 3	44.82	2.744 9	20.15
Slow-burst-slow	1.219 3	35.84	1.400 0	30.75
Burst-slow-burst-slow	1.219 3	6.43	1.196 9	6.39
Others	1.219 3	41.25	1.649 2	33.96

In conclusion, the early popularity of a video has a linear correlation on a log-log scale with the future popularity, but we find such relationship is largely impacted by the popularity evolution pattern of the individual video. We

Videos with steady increasing popularity are usually about current hot topics which can naturally continuously attract users. And a dramatic increase in popularity in the middle of lifetime is usually related to online social networking effects [20–21], such as the video is retweeted by a famous person.

We further investigate the popularity evolution patterns of popular videos. As Table 1 shows, the most common pattern of popular videos is burst-slow (66.13%), followed by steady (11.62%), burst-slow-burst-slow (8.54%) and slow-burst-slow (7.19%). It can be noticed that, comparing to the results of all videos, the percentages of popular videos are less biased towards the top 1 evolution pattern.

This indicates the evolutions of popularity for popular videos are usually more complicated.

4.4 Early-future popularity relationship

Generally, a video which receives a large number of views soon after uploading is very likely to become popular in the future. Otherwise, a small early popularity usually corresponds to an unpopular video. Szabo et al. [2] first observed there is an approximately linear relationship between the early and future popularity on a logarithmic scale. Based on this characteristic, they proposed a log-linear model using the log-transformed early view count to predict the log-transformed future popularity as followed:

$$\widehat{N_{\mathrm{v}}}(r) = \mathrm{e}^{\ln N_{\mathrm{v}}(i)+\ln \tau(i,r)} \qquad (4)$$

where N_v (r) is the predicted view count of video v on the rth day, and N_v (i) is the early view count on the ith day. τ (i,r) is a general model parameter, learned from the training set treating all the videos as a whole at the same time. This model and its modification versions have been widely verified on various kinds of online content datasets: YouTube and Digg [2,22], 20Minutes [23], iFeng News [24] and etc.

Fig. 10 shows the cumulative percentage of the average normalized popularity over the 30 days for all videos, popular/non-popular videos and videos with different kinds of popularity growth patterns. It can be noticed that the curves of popular and non-popular videos almost overlap the curve of all videos. This indicates the popularity of videos hardly impacts the early-future popularity relationship. However, the curves of videos with different popularity evolution patterns deviate the curve of all videos a lot. Burst-slow and burst-slowburst-slow videos tend to obtain influential amount of popularity soon after the uploading, whereas slow-burstslow and steady videos usually receive relatively small fraction of popularity at the early stage. For instance, on the 7th day, the former two kinds of videos averagely receive 81.21% and 78.20% of the total view counts.

While for the latter two kinds of videos, the percentages are only 46.61% and 29.09%, respectively. Hence, using a general model without considering the popularity evolution pattern of the videos to describe the relationship between early video popularity and future video popularity [2] is not accurate and proper.

Table 1 Summary of the top 4 popularity evolution patterns

Pattern	Number of videos	Proportion of videos/(%)	Number of popular videos	Proportion of popular videos/(%)
Burst-slow	8 220	76.84	2 060	66.13
Steady	816	7.63	362	11.62
Slow-burst-slow	691	6.46	224	7.19
Burst-slow-burst-slow	649	6.07	266	8.54
Others	321	3.00	203	6.52

Fig. 9 illustrates the typical curve of view count over time for each of the top 4 evolution patterns. From Fig. 9, it can be learned that burst-slow is the most common pattern, covering as many as 76.84% of the videos. This is as expected, since the recommendation scheme of service providers and the social influence of video uploaders may usually bring a large number of initial views for the newly uploaded videos. Meanwhile, we find there are indeed noticeable proportions of videos that experience no (steady) or multiple (burst-slow-burst-slow) bursts, or experience bursts in the middle of the lifetime (slow-burst-slow), as shown in Table 1.

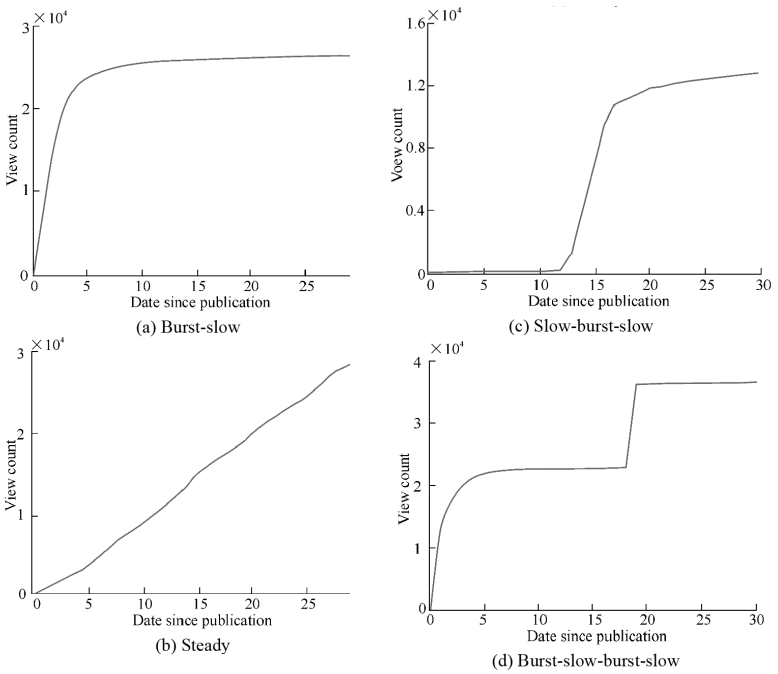


Fig. 9 Typical view count curves over time for the top 4 popularity evolution patterns

Next, based on the number and temporal locations of popularity bursts, we further define the popularity evolution pattern for an individual video, in order to describe the popularity growth trend. More specifically, for each video we can get a sequence of the popularity growth trend based on Eq. (3). We denote the sequence as $S = [s_1, s_2, ..., s_{30}]$, where s_i=1 if the growth trend is burst on the ith day, and s_i=0 if the growth trend is slow.

Then, by merging the consecutive same trends, we try to generate the final popularity evolution pattern of the video. For instance, for the sequence [1,1,1,0,0,0,...0] the evolution pattern will be '10' (burst-slow), which indicates the video experiences a sudden increase in popularity at the beginning of the lifetime, and receives limited daily views afterwards. If no burst is found during the observation period (i.e. no 1 in S), the evolution pattern will be regarded as steady.

However, we find there may be some stochastic disturbances lying in the growth trend sequence. For instance, a burst-slow pattern may result in a sequence like [1,1,1,0,1,1,0,...,0], with a one-day 'noise' trend 0. If we directly merge the sequence for final pattern, the result will be '1010' (burst-slow-burst-slow) instead of '10'. To tackle this problem, we apply a smoothing algorithm on the growth trend sequence before the merge step, to filter out those disturbances. The algorithm iterates the sequence with a window of specified length ω , and leverages the output of the last state in the window according to other state values in the window and the scale parameter κ. The pseudo code of our algorithm is shown in Algorithm 1.

Algorithm 1 State sequence smoothing algorithm

Input: s, ω, κ
Output: P
Procedure: SmoothFiltering ($s, \omega = 5, \kappa = 0.675$)
1: P={}
2: for $i = 0$ to $\omega - 1$ *do*
3: if sum(s [0 :$\omega - 1$]) < ($\omega - 1$) / 2 then
4: $P_i = 0$
5: else
6: $P_i = 1$
7: end if
8: end for
9: for i =ω to len(s) do
10: if κsum(s [$i - \omega + 1$: $i - 1$]) /$\omega + s_i (1 - \kappa) < 0.5$ then
11: $P_i = 0$
12: else
13: $P_i = 1$
14: end if
15: end for

Table 1 shows the top 4 popularity evolution patterns in our dataset, which cover 97% of the total videos. For each pattern, we list the number of videos, the proportion of total videos, the number of popular videos and the proportion of total popular videos.

tively steady. In our analysis, we set β =4.0, where the steady decrease begins.

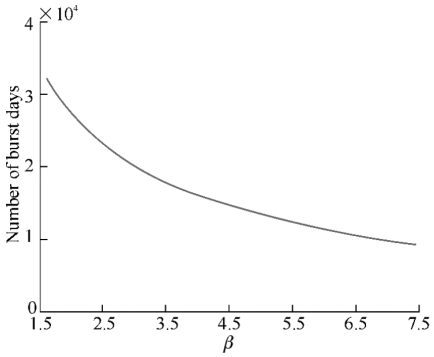


Fig. 7 Burst day count with different β

Fig. 8(a) shows the CDFs of burst day counts for all the videos, popular videos and non-popular videos. It can be noticed that most of the videos (92.37%) will experience 1 or more bursts during their lifetime. 42.57% of the videos just have 1 burst, and 39.09% of the videos have 2 bursts.

Videos with more than 2 bursts account for only 10.71%. For non-popular videos, their burst day counts concentrate in 1 (48.51%) and 2 (38.13%) days. While for popular videos, the burst day counts distribute more evenly, with 11.62% having 0 burst day and 18.84% having more than 2 bursts.

We then look through the temporal locations of the burst days. Fig. 8(b) shows the per-day histogram of the burst day count for all the videos over the observation period. It is apparent from Fig. 8(b) that most of the bursts happen on the first two days, while there are indeed some bursts appearing in the middle of the observation period.

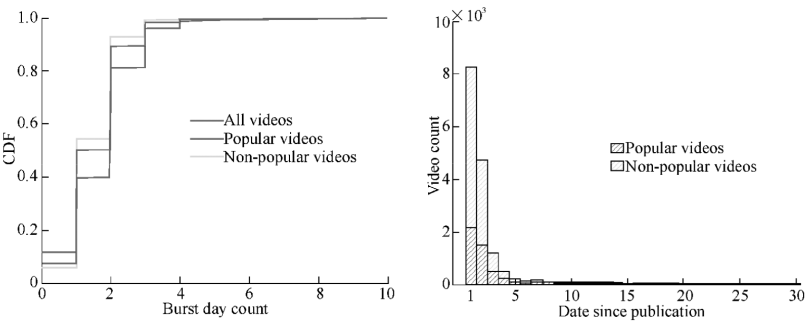


Fig. 8 CDF and histogram of burst days over the observation period

patterns in Fig. 6 are quite different. On active days, the rush hours of video playback are at noon, and the average view counts are relatively low at night. While on inactive days, with a peak appearing at noon, the view count curve starts increasing in the evening and reaches the top at late night. This indicates the variances of user behaviors on different kinds of days.

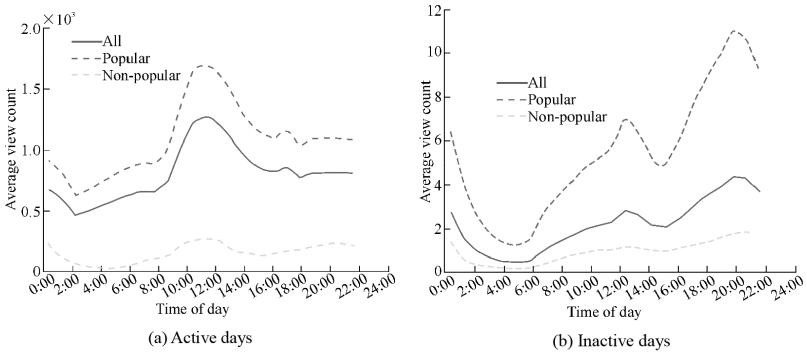


Fig. 6 Average view count in each hour of day during the active days and inactive days

Fig. 6 also show the per-hour average view count for popular videos and non-popular videos. We find the trends of the curves with different popularity are similar, both proportional to the curve of all videos. Hence, video popularity generally does not impact the per-hour view count distribution.

4.3 Popularity evolution pattern

In this section, we analyze how the popularity of an individual video evolves over time since its publication. Instead of a simple logarithmic growth trend, as previous research concluded [19], we find the popularity evolutions of Youku videos are rather complicated. Some videos may receive most of their views on a small number of days, while other videos may remain active with similar daily views all over their lifetime.

To measure the growth trend of video popularity, we propose the notion of popularity burst. More specifically, let N_v (i) be the increase in view count for video v on the ith day. If the ith day is an active day and N_v(i) is larger than a threshold V_a, we regard the popularity growth trend on that day as burst. Otherwise the popularity growth trend will be regarded as slow. V_a is defined as:

$$V_a = \beta \frac{N_v(n)}{n} \qquad (3)$$

β is a scale parameter over the average increase in daily views (i.e.N_v(n) / η) for a burst day. To get a proper value, we vary β from 1.6 to 8.0, and calculate the number of burst days for all the videos, as shown in Fig. 7. We find the number of burst days decreases with the value of β increasing. When β is small, the number drops sharply, and when b is large, the decrease gets rela-

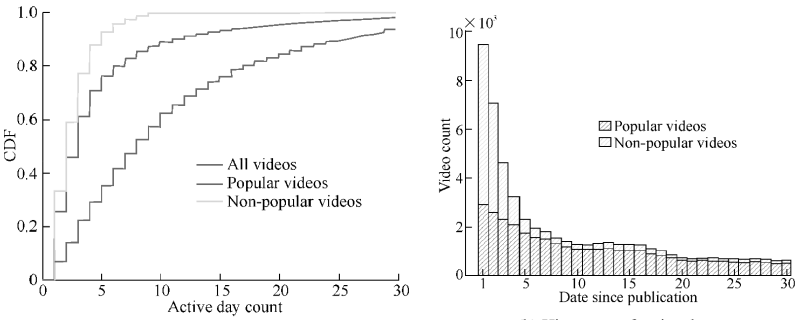
eos. Clearly, inactive videos have poor abilities to attract users, and are of far fewer values for service providers and network operators in practical. Hence, we focus on the active videos in our following analysis.

Fig. 5(a) shows the CDF of active day counts for all the active videos in our dataset. It can be noticed that most of the videos are with quite small active day count. For instance, around 25% of the videos only have 1 active day, and around 80% of the videos have 6 or fewer active days.

It is apparent from Fig. 5(α) that non-popular videos tend to have less active days, whereas popular videos are usually with large active day counts. For instance, 92.85% of the non-popular videos have 5 or fewer active days. In contrast, the percentage of popular videos is only 35.41%.

Around 40% of the popular videos are with active days more than 10 days. Such difference of the active days between popular videos and non-popular videos is intuitive, since more active days will usually bring more views to a video and finally make the video popular.

Next, we look through the temporal location distribution of the active days during the observation period. Fig. 5(δ) shows the per-day histogram of the active day count for all the videos over the observation period. It is clear from Fig. 5(δ) that for non-popular videos, most of the active days concentrate at the early stage of video lifetime; while for popular videos, the active days distribute more evenly across the observation period, with the video count at the early period slightly higher than on other days.



(a) CDF of active day count

(b) Histogram of active days

Fig. 5 CDF and histogram of active days over the observation period

This observation is of great importance for service providers and network operator to design and adjust the video caching schemes. More specifically, the golden days of non-popular videos are usually very short, hence caching those video for only the first several days will be just enough. As for popular videos, more caching days should be employed, in order to cover their whole active periods.

We also investigate how the popularity of an individual video distributes within a day at the hour granularity. Averagely, the view count on active days is much larger than that on inactive days (1 082.43 vs. 5.50). Hence, in Fig. 6, we plot the average view count in each hour for all the videos during the active days and inactive days, respectively. It can be noticed that the circadian

It can be noticed that, when h is smaller than 300, the active day count decreases dramatically. When η is larger than 300 and smaller than 500, the number of active days still drop obviously. While h is larger than 500, the number of active days varies steadily. Therefore, in our analysis, we set η= 500.
However, as analyzed in Sect. 4.1, for around 80% of the videos, the 30th day view counts are less than 1 000, thus the average daily views are less than 34, which are far less than η = 500 . This indicates only using the threshold η to evaluate the video activeness of the day may be too strict for the less popular videos. To tackle this problem, we introduce two thresholds δ and α in both absolute and relative terms of daily views to measure the active day for the less popular videos. That is, if a video receives the daily views more than δ , and α times of the average daily increase in view count, we can also regard that day as the active day for the video. Fig. 4 (b) shows the number of active days over d with different α values.

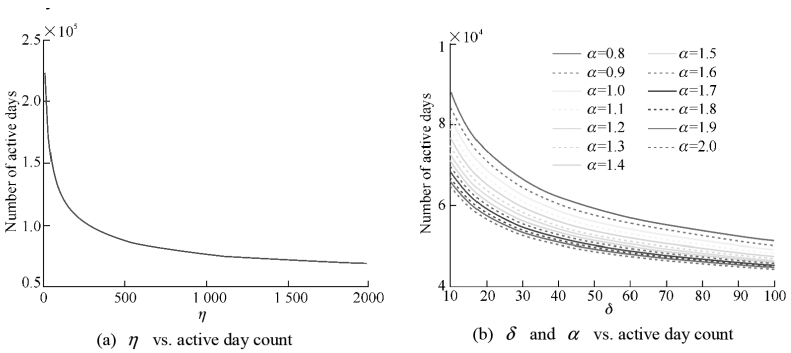


Fig. 4 Active day count with different η , δ and α

It is apparent from Fig. 4(δ) that the shapes of the curves with different a are quite similar. When η is very
small, with the value of d increasing the number of active days decreases significantly. While, since around 40, when η is relatively large, the active day number decreases steadily. Hence, in our analysis we set η = 40.
And for the choice of α , we find when α is small, the distance between two adjacent curves is large. While, since around α = 1.6, the curves are nearly overlapped. Hence, we set α = 1.6 to obtain steady results of the active day count. Overall, we define the threshold of views a V that a video V should receive on an active day as:

$$V_{\mathrm{a}} = \min\left(\eta, \max\left(\delta, \alpha \frac{N_{\mathrm{v}}(n)}{n}\right)\right) \qquad (2)$$

where n is the total number of the observation days. v N (n) is the total view count that v receives during the n days. And h = 500, d = 40, a = 1.6.
Based on Eq. (2), we calculate the number of active days for each video in our dataset. We find 17 076 videos have no active day during the observation period. We regard those videos as inactive videos, and the rest as active vid-

from the beginning of lifetime to the long-term. We note our observation is different from the previous studies [19,16], which fit the popularity of YouTube videos with Weibull or Log-Normal distribution.

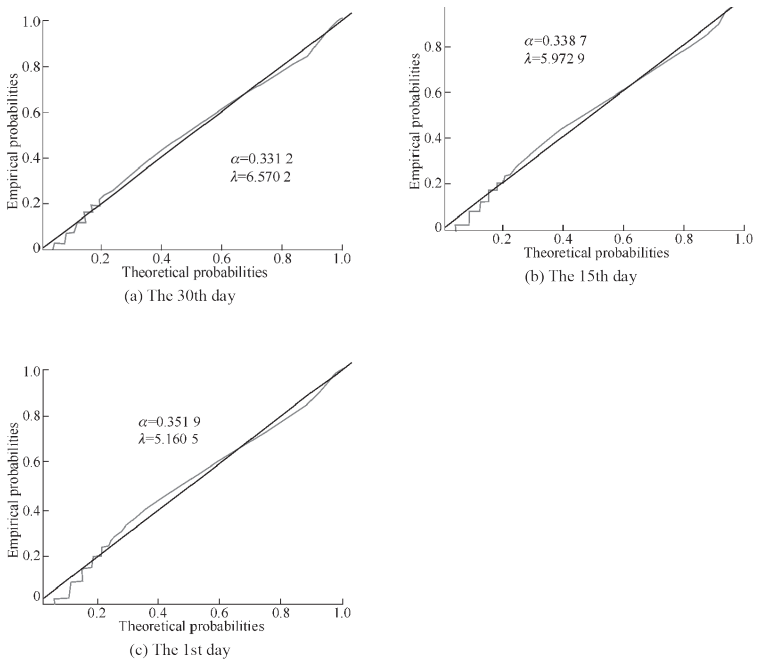


Fig. 3 P-P plots of the empirical distribution vs. the Pareto Type 2 distribution for video view count

Although the specific forms of those popularity distributions vary across service providers and datasets, they are all heavy-tail distributions.

4.2 Individual popularity distribution

In this section, we first propose the notion of active day to study the view distribution of an individual video during the observation period. At the day granularity, we find on some days a video may be widely watched by users, while on other days the video may only receive limited views. To measure the ability of an individual video to obtain the views on different days, we propose the notion of active days. More specifically, if a video is able to receive adequate views more than a threshold a V on one day, we consider that day is an active day for the video. Otherwise, if the daily increase in view count of the video is less than V_a, that day will be regarded as an inactive day for the video.

An important question is how to get a proper value of V_a for each individual video. First, we define a threshold η, indicating the minimal requirement for the views that a video should receive on an active day. We vary the value of η from 0 to 2 000, and calculate the number of active days for all the videos in our dataset, as shown in Fig. 4(a).

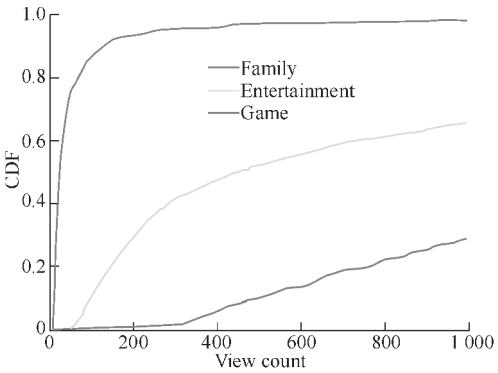


Fig. 2 CDF of long-term view count for videos in different categories

In addition, we analyze the influence of video tags as impact factors of video popularity. We first extract the hot tags which appear in more than 10 videos in our dataset. In total, there are 563 hot tags, accounting for 2.5% of all the tags. This shows that most of the tags do not appear frequently. Then, we check the hot tags of the popular videos and non-popular videos, respectively. We find while most tags are common among videos on different popularity levels, there are indeed some tags that only appear in the videos on certain popularity level. More specifically, 431 hot tags appear in both popular and non-popular videos, while 97 hot tags appear only in popular videos and 35 hot tags appear only in non-popular videos. Taking the video tag 'Survivor Games' for instance, it appears in 28 videos in our dataset. All of the 28 videos are popular videos, with the average view count equal to 50 833 and the largest view count equal to 123 943. Hence, the presence of certain tags in the video tag list can affect the video popularity.

We further study if there is a mathematical model that can describe the overall distribution of Youku video popularity. After comparing with various hypothetical distributions, we find that the Pareto Type 2 (Lomax) distribution can well fit the empirical distribution (i.e. the distribution of video view count over different date in our case). The probability density function of the Pareto distribution (i.e. for a random variable X, the value at a given point xî X) is given by:

$$f(x) = \frac{\alpha \lambda^{\alpha}}{(x+\lambda)^{\alpha+1}}$$

where a > 0 is the shape parameter, and l > 0 is the scale parameter. We run regressions with the popularity of all the videos in our dataset on the 30th day, 15th day and 1st day, respectively. Probability-probability (P-P) plots, together with the estimated parameters, are shown in Fig. 3 to visualize the goodness of fit.

It can be noticed that from Fig. 3, most of the dots in the scatter plots follow a straight line with the slope is 1 and the intercept is 0 (i.e. the black curve). This indicates that the Pareto distribution is a good approximation of the overall popularity distribution, for the newly uploaded videos thoroughly

It can be noticed that the popularity of different videos varies substantially, with the view counts in the horizontal axis spanning over 8 decades. Most of the videos are barely noticed by users, whereas a small number of videos may achieve very large view counts. For instance, 60.71% of the videos are with view count less than 100, and 79.01% of the videos are with view count less than 1 000.

Meanwhile, 11.22% of the videos are watched for more than 10 000 times, and the most popular video (VID = XMTU0MjYzMjE0OA==) is watched for as many as 57 040 334 times. In particular, in our analysis we define the videos with view count more than 10 000 as popular videos, and the rest as non-popular videos. The skewness of video popularity is apparent from Fig. 1(b): the top 1% videos account for as many as 82.88% of the total views, and the top 10% videos account for almost 99% of the total views.

Various reasons underlie such highly asymmetric video popularity. Besides the content subjects of different videos naturally demonstrate different appeal, the recommendation strategy of the service provider also plays an important role in the video popularity. For instance, Youku usually lists the newly uploaded television (TV) episodes and variety shows on the first page of its portal to attract users.

Those videos are much more visible to users, resulting in a rich-get-richer effect on their popularity.

Next, we aim to shed light on the impact factors (video properties in our case) of the popularities of different videos. Some pearson's correlation coefficients with the video view count on the 30th day are as followed: video duration (0.034), tag count (0.003), video resolution (0.137), uploader's video count (0.003) and uploader's follower count (0.002). No significant correlations can be observed between these properties and the long-term video view count. Nevertheless, we find that the category of a video has a great impact on its view count. Fig. 2 shows the CDFs of video view counts on the 30th day for three representative categories: 'family', 'entertainment' and 'game'. Significant popularity differences can be observed between different video categories. On average, the videos in the three categories are watched for 865, 9 786 and 21 780 times, respectively. Videos in the 'family' category tend to have very small view count. 93% of them are watched for less than 200 times. While the percentages of those kinds of videos in the 'entertainment' and 'game' categories are around 30% and lower than 5%. Videos in the 'entertainment' category have a relatively uniform popularity distribution. And for the 'game' videos, the view counts are usually very large. Over 71% of them are watched for more than 1 000 times. The differences of view counts reflect the differences of user interests in these videos. Videos in the 'family' category are usually private videos uploaded by users to share with their close friends. Other users have limited interests in those user-generatedcontent (UGC) videos. However, most of the videos in the 'entertainment' category are video-on-demand (VoD) contents released by the video portal to attract various users. And recordings and commentaries of computer games account for a large proportion of the 'game' category. These videos are very popular among the young users, who contribute to the majority of the view counts.

lion daily video views. To study the popularity dynamics of Youku videos, we crawled and tracked the view counts of a set of newly uploaded videos for a period of time via Youku open application programming interface (API) 'cloud.youku.com'. The data were gathered with a two-step procedure.

Step 1 On a certain date we initially crawled the video list every hour from the portal's 'most recently uploaded videos of the day' section.

Step 2 We tracked the popularity of these videos for 30 consecutive days since the uploaded date, by retrieving their updated view counts every hour via the application program interface (API). Thus, for each video in the video set collected in Step 1, we can obtain its per-hour view count time series for 30 d.

In Step 1, we collected the video lists in April 15~21, 2016, and got 7 sets of the daily newly uploaded video lists.

In Step 2, we carried out the 30-day popularity tracking for each video in the 7 video sets. The whole collection period ended on May 23rd, 2016. During the collection, 1 217 videos were either deleted by the uploaders or blocked by Youku. We excluded those videos from our dataset.

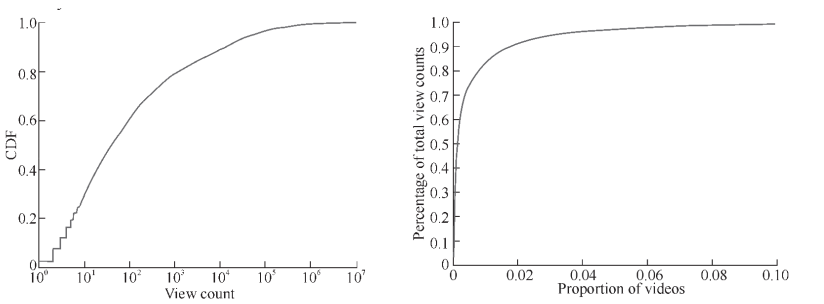
Eventually, our dataset consists of 27 773 videos, which are watched 1 628 600 925 times by users.

Note the Youku API also provides several other kinds of video lists, such as 'most viewed videos of the day' or 'most favorited of the day'. Nevertheless, in our analysis, we choose to collect the most recently uploaded videos rather than others, in order to avoid the sampling bias [2,16], and get a complete history of video popularity since the uploaded date.

4 Analysis of popularity characteristics

4.1 Overall popularity distribution

We first look through the overall popularity distribution for a set of newly uploaded videos. Many previous studies have observed that the popularity of online content usually distributes unevenly, and exhibits a heavy tail [17–18]. This characteristic also holds for the Youku videos. Fig. 1(a) shows the cumulative distribution function (CDF) of the view counts for all the videos in our dataset on the 30th day since the publication date. For a better visualization, the horizontal axis of the figure is logarithmically rescaled.



(a) CDF of video view count on the 30th day

(b) Percentage of total view count over the proportion of videos

Fig. 1 CDF of video view count on the 30th day and percentage of total view count over the proportion of videos

2 Related work

2.1 Popularity distribution of online videos

Gill et al. [5] collected traffic in a campus network and characterized the usage patterns, video properties, video popularity and transfer behaviors of the YouTube service. Zink et al. [6] also collected traffic in a university campus network, and analyzed the session duration, session data rate, video popularity and user access patterns of YouTube videos. Cha et al. [7] analyzed the video popularity characteristics of two video sharing systems, YouTube and Daum. They examined the distribution and evolution of video popularity, and investigated mechanisms to improve the video delivery. Tan et al. [8] collected a large dataset of user watching behavior records from PPTV, and studied central tendency, dispersion tendency and asymmetry of users' popularity preferences distributions. Liu et al. [9] collected five-month daily download traces from four mobile application stores, and analyzed the common and different application popularity features between Android and iOS stores. Li et al. [10] collected large-scale and long-term dataset in the mobile network, and characterized the user activity, user network, video properties and video popularity of Youku. Those works provide valuable insights into the popularity distribution of online videos.

However, they focused mostly on the overall popularity distribution of a group of videos, and collected the data at either a single or several days (i.e. snapshots). Our study complements these works by tracking and analyzing the video popularity for a whole observation period (30 consecutive days) since the uploading. Moreover, our work is more fine-grained, analyzing the per-hour popularity and individual video popularity.

2.2 Popularity evolution of online videos

Crane et al. [3] studied the video popularity with an epidemic model. They found the relaxation process followed a power law, and classified the popularity evolutions into four patterns according to the peak fraction. Cheng et al. [11] crawled data from YouTube and studied the video properties, access pattern, popularity distribution, popularity growth trend of YouTube videos. They used a simple logarithmic model to describe the growth trend of video popularity. Szabo et al. [2] studied the relationship between early and future video popularity. They found early access of the content could reflect long-term user interest, and presented a log-linear model to predict the popularity of online content. Figueiredo et al. [12–13] analyzed the characteristics of popularity evolutions and referrer types of YouTube videos. Gonçalves et al. [14] characterized the scholar popularity in the computer science research community. They used K-spectral clustering [15], a time series clustering algorithm, to identify the dynamics of the scholar popularity. Our study complements those previous works by proposing more proper popularity evolution patterns that can describe the whole observation period rather than only the peak days.

Moreover, we take the number and temporal locations of the popularity bursts into consideration. Based on the analysis of popularity evolution, we further correlate early view counts with future view counts according to different popularity growth patterns.

3 Dataset

The data used in this paper were collected from the online video service provider Youku. Youku is one of the most popular online video service providers in China, with more than 500 million monthly active users and 800 mil-

traffic in 2016, and will be up to 82% in 2021. Every second, nearly a million minutes of videos cross the Internet, and it will take an individual over 5 million years to watch all the newly uploaded videos each month. Considering the limited time and diversity of user attention, it is not surprising that the popularity of online videos is allocated in a rather asymmetric way [2]: a small fraction of the videos receives most of the user views, whereas the vast majority are barely noticed by users.

Given the large volume and uneven popularity of online videos, a better understanding of the popularity dynamics is critical to the interest parties in various contexts. For instance, with the popularity information of the newly published content, effective information services (e.g. video recommendation and searching) can be better designed and supported by the service providers. And advertisers in the online marketing can place advertisements, plan campaigns and estimate costs based on the video popularity information, in order to maximize their revenues. Network operators can deploy and adjust their network infrastructures such as cache servers accordingly in advance for better video delivery. And in the social aspect, a thorough understanding of popularity evolution will uncover the rules governing collective human behavior. That is, to reveal where the popularity growth of an individual video comes from: caused by random user choices, endogenous effects (e.g. listed on the home page) or exogenous effects (e.g. shared on other websites) [3].

In this paper, we analyze the dynamics of video popularity of a leading online video service provider in China, namely Youku (www.youku.com). Our study is based on an up-to-date dataset containing 27 773 newly published videos collected from Youku website for 30 consecutive days. We use view count, i.e. how many times a video is watched by users, as the metric of online video popularity. View count is widely used in the video popularity study, and other metrics such as comment count or favorite count are largely correlated with it [4]. Based on those, we study the characteristics of online video popularity from four key aspects: overall popularity distribution, individual popularity distribution, popularity evolution pattern and early-future popularity relationship. The main contributions of our work are summarized as follows:

1) We collect a long-term, up-to-date and fine-grained dataset of online video popularity. We further release our whole dataset to the public for research purpose (https://github.com/lichenyu/datasets/tree/master/youku_popularity_perhour_160415_160421/).
2) We analyze both of the overall and individual distribution of online video popularity. In particular, we investigate the per-day popularity distribution throughout the observation period and the fine-grained per-hour popularity distribution within a day, which are seldom analyzed in previous works.
3) We investigate the popularity growth trend of an individual video, and reveal a small number of evolution patterns. In particular, the number and temporal locations of sudden popularity bursts are taken into account in our analysis to describe the popularity growth.
4) We shed light on the influence of popularity evolution patterns on the linear correlation between early popularity and future popularity. We utilize specialized models on a logarithmic scale to model the correlation, which performs better than the current state-of-the-art model.

The rest of this paper is organized as follows. Sect. 2 briefly discusses the related work. Sect. 3 describes the dataset and collection method used in our study. The main analysis results are presented in Sect. 4. In Sect. 5 we conclude the paper.

Analyzing the dynamics of online video popularity

Ouyang Shuxin ▪
School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
Li Chenyu ▪
School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
Li Xueming ▪
Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing 100876, China

abstract
Given the large volume of video content and the diversity of user attention, it is of great importance to understand the characteristics of online video popularity for technological, economic and social reasons. In this paper, based on the data collected from a leading online video service provider in China, namely Youku, the dynamics of online video popularity are analyzed in-depth from four key aspects: overall popularity distribution, individual popularity distribution, popularity evolution pattern and early-future popularity relationship. How the popularity of a set of newly upload videos distributes throughout the observation period is first studied. Then the notion of active days is proposed, and the per-day and per-hour popularity distributions of individual videos are carefully studied. Next, how the popularity of an individual video evolves over time is investigated. The evolution patterns are further defined according to the number and temporal locations of popularity bursts, in order to describe the popularity growth trend. At last, the linear relationship between early video popularity and future video popularity are examined on a log-log scale. The relationship is found to be largely impacted by the popularity evolution patterns. Therefore, the specialized models are proposed to describe the correlation according to the popularity evolution patterns. Experiment results show that specialized models can better fit the correlation than a general model. Above all, the analysis results in our work can provide direct help in practical for the interested parties of online video service such as service providers, online advisers, and network operators.

Keywords
online video service, online content popularity, popularity evolution pattern, early-future popularity relationship

1. Introduction

With the emergence of Web 2.0 services, a huge amount of video content has been brought into the Internet. Online video service is currently the dominating service on the Internet. As the white paper of Cisco System notes [1], global video traffic accounted for 73% of all the Internet