Physica A 606 (2022) 128105

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Collective user switching behavior reveals the influence of TV channels and their hidden community structure

Mingyan Wang, An Zeng, Xiaohua Cui *

School of Systems Science, Beijing Normal University, Beijing, 100875, PR China

ARTICLE INFO

Article history:
Received 22 February 2022
Received in revised form 21 July 2022
Available online 27 August 2022

Keywords:
TV ratings
Complex network
PageRank
Community detection

ABSTRACT

Television is the primary medium through which most families access entertainment and information in their daily lives. Thus, understanding users' TV viewing behavior is meaningful for several practical issues, such as evaluating the influence of TV channels and providing personalized TV recommendations. However, most existing works regarding TV viewing data are limited to basic statistics (e.g., TV ratings). In this paper, we analyze a large-scale TV viewing dataset for a city in China via a complex network approach. We construct a directed network that characterizes the collective channel-switching behavior of viewers. By using the PageRank method, we reveal the influential TV channels that are more in line with people's expectations than their rankings based on simple TV ratings. We further construct a network in which channels are linked according to their similarity in users' switching preferences. This network exhibits a clear community structure, which can help TV stations understand which channels are in the bottleneck and which channels have potential. Overall, our work provides a system perspective to evaluate TV channels and their relationships.

1. Introduction

People often post film reviews on websites, purchase goods using electronic payments, and choose programs among hundreds of channels. All these activities produce enormous amounts of data. In the big data era, user behaviors can easily be recorded electronically [1–3]. Using these digital traces, we can implement intelligent recommendation systems [4,5], explore the lifestyles of different groups [6,7], understand the rhythms of human mobility [8,9], model computer virus propagation [10], and study a series of meaningful property features about human behavior. Analyzing human behavior traces can not only capture individual and group characteristics but also help us acquire a better understanding of people's social lives, relationships, and organizations [11,12].

TV watching is one of such human behaviors, which is a major way to acquire information and spend leisure time for most families. Statistics show that in 2020, the total revenue of China's television industry reached 921.46 billion RMB, an increase of 13.66%. TV media still have great achievements in the all-media era. Various types of data are recorded during TV watching, for example, start time, end time, channel name, program name, etc. We can explore viewers' individual preferences and collective tendencies by analyzing data generated by TV-watching behavior. From the perspective of TV stations, such research can also quantify the competitiveness of channels and aid them [13,14].

The audience rating and TV channel attractiveness have long been the focus of TV watching research. The Cambridge Dictionary defines attractiveness as the quality of causing interest or making people want to do something, which is a

* Corresponding author.
E-mail address: xhcui@bnu.edu.cn (X. Cui).

https://doi.org/10.1016/j.physa.2022.128105
0378-4371/

concept that implies comparison. The Audience has hundreds of TV channels to watch but limited attention, so being attracted by one means giving up another (the opportunity cost). An opportunity cost is "the evaluation placed on the most highly valued of the rejected alternatives or opportunities" [15]. For TV channels, the so-called high value is difficult to quantify. The traditional rating model substitute attention received by TV channels for high value, which is not entirely appropriate to some extent. Because television rates such as Rtg and Reach are also determined by additional components such as audience availability, audience demographics, behavioral attributes of the viewers [16], and cast demographics [17]. For example, there are deviations between the ratings of different regions and the overall result, which is affected by different set-top box configurations in different regions [18]. In our data set, the ratings of local TV channels are overestimated. Because when viewers turn on the TV, they will contribute to the viewing time of local channels. Another example is that people may play a boring program as background when there is no program they want to see, which increases the channel's viewing time. But the attractiveness of the TV channel may be less than that of other TV channels. Therefore, "clicking" (called switching in the paper) should be paid enough attention to in the audience rating [19].

In this paper, a network-based TV channel ranking method is proposed from a perspective of switching behavior. We construct a directed network according to the audience's viewing sequences. The edge in the network represents the switches from one channel to another. If the audience actively switches from other channels to channel A, it indicates that channel A is more attractive. If the audience actively switches from channel A to channel B, channel B is more attractive than channel A. This is a process of comparison and diffusion, similar to webpage ranking. Therefore, we creatively apply the PageRank algorithm to channel ranking. The simulation results show that our ranking method can effectively decrease the rankings of local channels aiming to attract people in a specific area, and improve the ranking of high-quality channels for national audiences. We further calculate a channel correlation network in which the links reflect similarities in their probabilities of switching to other TV channels. We detect its community structure after extracting the backbone using the minimum spanning tree algorithm (MST) [20]. The results show large discrepancies in functional features and target audiences for different communities. TV stations can arrange more appropriate programs for channels that are in a bottleneck or have potential.

From the perspective of channel switching behavior, we apply the PageRank to channel ranking, which better captures the attractiveness of channels to viewers and reduces the impact of set-top box configuration or idle viewing on ranking compared to traditional TV rating indicators. Although the time complexity of our method is higher than that of simple statistical indicators, the ranking is more in line with audience and advertiser expectations. Further, we detect communities of channels based on users' collective switching behavior, which is rarely involved in previous research about TV ratings. Each community's ratings and function are consistent with reality, providing possible references for TV stations accordingly. In practice, our method and results can be used to provide personalized TV recommendations for the audience. For example, when the set-top box is turned on, it is suggested to automatically jump to the channel that the user frequently switches to instead of the default channel. Another example of a possible application is to change the adjacent relationship between channels according to the switching behavior of individual users, to reduce the cost of changing channels for users and help TV service providers retain customers.

2. Related work

Television is one of the most influential mass media developed to date. Since the advent of the first commercial TV in 1928, competitiveness between TV stations has been continuous. Ranking channels based on TV ratings is one of the most concerning issues for program producers and sponsors. The most widely accepted measurement of TV ratings is based on Nielsen's method, which was developed by Nielsen Media Research in the 1950s [21]. In the 20th century, limited by lack of technology, TV ratings were generally acquired through the diary, questionnaire, or instrument methods [14]. However, data collected by such methods are neither comprehensive nor objective, which has sequential impacts on the accuracy of indicators. In the past two decades, the development of network techniques has promoted media convergence of telecommunication networks, computer networks, and cable networks, making high-accuracy and large-scale viewing data more generally available [22]. Accordingly, some new research hotspots based on mathematical models for TV rating estimation and prediction have been developed [23–26]. However, there are two limitations that have not been fully discussed. First, although AI algorithms have achieved higher accuracy in predicting TV channels' performance, it has been controversial in giving interpretive audience analysis to TV stations. Second, most of these works focus on viewing time or viewer number and equate these measures with channels' attractiveness. As we mentioned above, traditional measurements cannot fully represent the attractiveness of TV channels [16,17,19]. The TV ratings of default channels, local channels, and channels only broadcasting programs, etc. are often overestimated and get a high ranking. Thus, finding a more scientific ranking method as the auxiliary of traditional methods remains to be addressed.

Competitiveness is a typical relationship between individuals that can naturally be characterized in terms of networks. Therefore, many ranking algorithms based on networks have been proposed, including centrality ranking methods [27], PageRank algorithm [28], Hits algorithm [29], etc. These methods and their variants have been widely applied to rank athletes [30,31], journals [32], authors [33], disease-causing genes [34], etc. PageRank is the most widely used ranking algorithm and is still at the heart of Google and other search engines. The popularity of the algorithm lies in its perceived effectiveness and philosophy easy to understand: instead of ranking objects according to their internal qualities that are

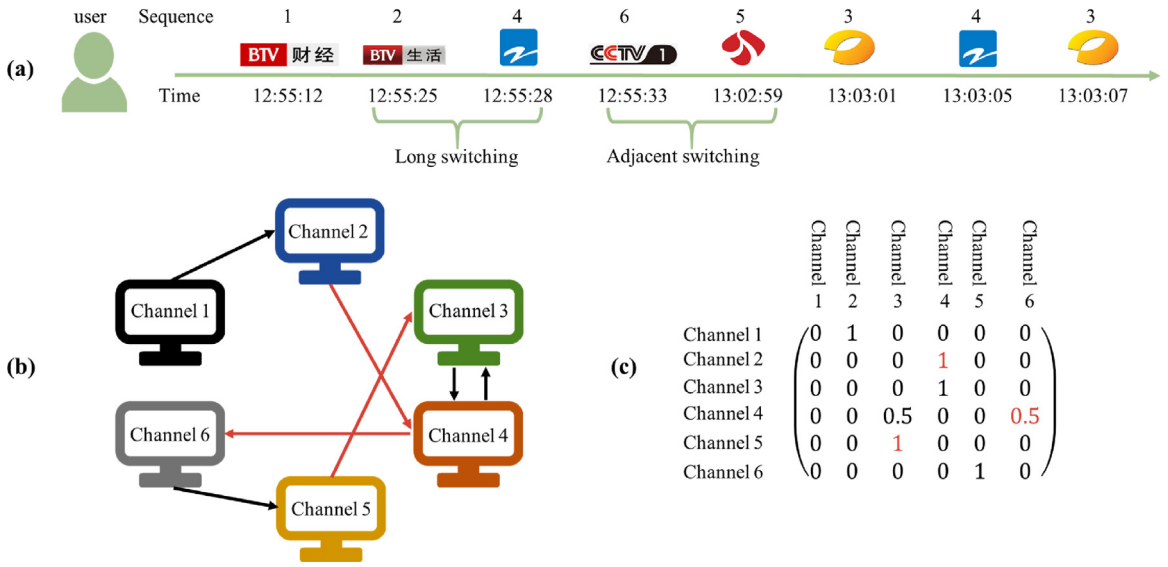

Fig. 1. Switching behavior and transition probability matrix. (a) A user's viewing records sorted by time. The sequence on the top reflects the channels' serial numbers. (b) Switching between TV channels represented by a network. When a viewer switches from channel 1 to channel 2, a directed edge is formed between the two channels. The black edges represent adjacent switching (between adjacent channels) and the red edges indicate long switching (between nonadjacent channels). (c) The transition probability matrix S between channels corresponding to (b). Taking $S(4,:)$ as an example, the viewer switches from channel 4 to channel 3 and channel 6, so $S(4,3) = S(4,6) = 0.5$. The black indicates adjacent switching and the red indicates long switching.

difficult to measure (such as the utility of web pages or the attractiveness of TV channels), it is better to use the collective wisdom of the network to interpret each link as an internal vote [35]. Competitiveness between channels is similar to that between web pages, authors, and athletes. Thus, how to convert the competitiveness to a network to achieve an effective ranking is one of the problems we aim to address.

Representing links between components with networks is not only useful for ranking but also contributes to exploring other features of complex systems, such as vulnerability [36], segmentation [37], and clustering [38]. Clusters (or communities) are relatively independent parts of a network, functioning as organs of bodies. Community detection is of great significance for mining information because it can help in inferring the inner properties among components. Divisive algorithms, modularity-based methods, spectral algorithms, dynamic algorithms, and others are commonly used to detect communities [39,40]. Community detection can reveal people's group characteristics in real datasets on human behaviors. For example, clustering based on mobile phone data can help in understanding the patterns of human mobility [41,42]. Similarly, clustering data from online platforms can help to detect social structures [43,44]. These results inspired us to investigate the inherent characteristics of channel networks through community detection.

3. Data description

3.1. TV viewing data

Since the 1990s, television rating research in China has developed rapidly. Currently, China has established the world's largest TV rating survey network. The method used for such research has also changed from the simple diary card method to the instrument method. The current spread of digital TV has made it easy to obtain accurate, objective, and massive viewing data from set-top boxes after 2010 in China [45]. When the TV is turned on, the equipped set-top box will automatically record the operation of the audience until the TV is turned off without missing any viewing record. In this paper, we analyze TV viewing records from a certain city in China for the period from July 1, 2015, to September 30, 2015. The records include user ID, set-top box ID, date, start time, end time, channel name, and channel sequence, totaling 39,625 users and 268 TV channels.

When an audience is watching TV, it tends to switch channels until finding the program it likes, as illustrated in Fig. 1(a). There are usually two channel-switching modes. One involves jumping directly from the current channel to the target channel. However, if it does not have a specific target, it will switch to a nearby channel, browsing in a channel-by-channel fashion until it is attracted by some program. Jumping from the current channel to a nonadjacent channel is termed long switching, while mechanical forward or backward is termed adjacent switching in this paper. Long switching can better reflect the attractiveness of channels because people give up the current channel for the nonadjacent channel that attracts them more.

The switching between TV channels can be naturally represented by a directed network, whose mathematical form is a transition probability matrix originally proposed by Markov [46] as shown in Fig. 1(b)(c). The transition probability matrix is calculated as follows:

$$s_{i,j} = \frac{C_{i,j}}{\sum_{k=1}^{N} C_{i,k}}$$
$$S = \begin{bmatrix} s_{1,1} & \cdots & s_{1,N} \\ \vdots & \ddots & \vdots \\ s_{N,1} & \cdots & s_{N,N} \end{bmatrix} \tag{1}$$

where N is the number of channels; $C_{i,j}$ is the switching count from i to j in a certain period; $s_{i,j}$ is the switching probability from i to j; S indicates the transition probability matrix in the period. The i_{th} row represents the probabilities of switching from the i_{th} channel to other channels. The sum of elements in each row is 1. Further, we can also define two kinds of transition probability matrices according to the two channel-switching modes. If we only retain the long switches in the viewing sequences, we can calculate the long transition probability matrix. The adjacent transition probability matrix is defined in the same way.

3.2. Basic statistics

In this section, we calculate four common indexes to understand the basic picture of TV viewing in the city. We plot the daily average viewer number per hour (Fig. 2(a)), which shows that there are two peaks (at 12 PM and 8 PM) and that the lowest occurs at approximately 5 AM each day. Fig. 2(b) indicates that channel viewing time is a straight line under the semilogarithmic axis, revealing that the ratings between channels are extremely uneven. The probability function of the daily average viewing time of all viewers is characterized by a log-normal distribution (Fig. 2(c)). In addition, the 10 channels with the longest viewing time account for more than 85% of the audience's total viewing time (Fig. 2(d)). In general, TV channels are heterogeneous in their ratings, and viewers have their fixed favorites among hundreds of channels. In other words, audiences have clear channel preferences. We aim to quantify these preferences to rank and evaluate TV channels.

3.3. Channel-switching modes

We defined two switching modes in the previous section: long switching and adjacent switching. Based on these initial definitions, we further define the long switching rate and the adjacent switching rate as follows:

$$l = \frac{L}{C}$$
$$a = \frac{A}{C} \tag{2}$$

where C is the total switching count, L is the long switching count and A is the adjacent switching count within a certain period. l and a respectively represent the proportion of long switches and adjacent switches. Audiences have different long switching rates at different times. Fig. 3(a) shows that l is about 0.76 ± 0.04, fluctuating slightly in a day. It is clear that long switching occurs frequently at 6 AM, 12 PM and 6 PM corresponding to periods such as after getting up and after work when people just turn on the TV and tune to their preferred channels. Fig. 3(b) is the distribution of all users' long switching rates with $\bar{l} = 0.71 \pm 0.14$. This means that most users prefer long switching (with few exceptions), which is more reflective of users' subjective choices. Fig. 3(c) and (d) reveal that after a long switch, people are more likely to perform additional long switching. However, after an adjacent switch, a long switch or an adjacent switch are equally probable. Fig. 3(c) and (d) confirm that the audience probably will continuously do long switching, which is the main switching mode. These also ensure that we will not delete many continuous adjacent switching records when only long switches are retained to construct the transition probability matrix. Fig. 3(e)(f) shows the dependence of two switching rates on the watched time. With the longer watched time, the long switching rate increases slightly, which is in line with the situation that the audience will not easily switch channels aimlessly when they are immersed in a channel for a long time. The above information reveals that people often have clear targets in mind when switching channels and that long switching is dominant and continuous among the two modes. It encourages us to consider ranking TV channels in terms of channel switching, especially long switching.

4. Application of PageRank in channel ranking

There are many existing TV ratings indicators. The two most critical indicators are *Rtg* and *Reach*. These are the basis of other derivative indicators, and they are calculated as follows:

$$Rtg = \frac{WT_{total}}{T \times N} \times 100\% \tag{3}$$

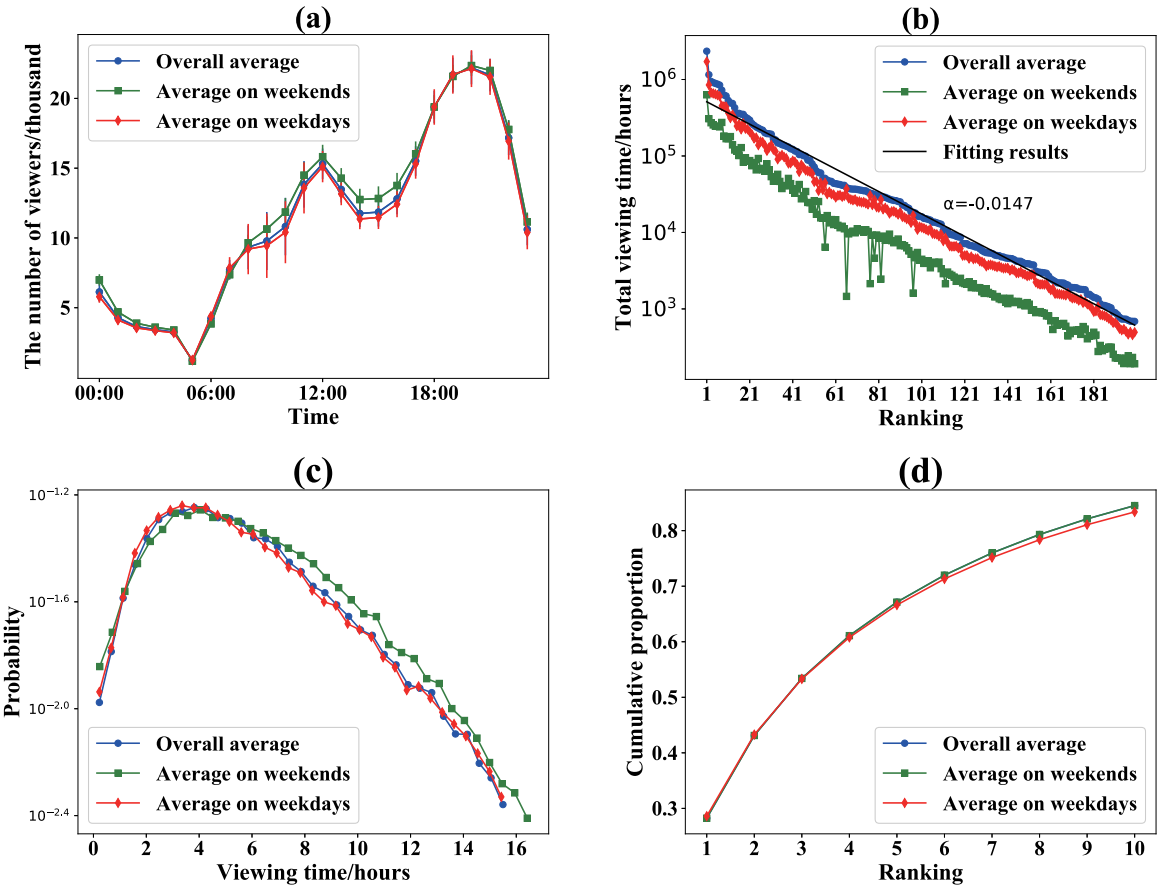

Fig. 2. Basic statistics of the TV viewing data. The results are obtained from the viewing records of a Chinese city from July 2015 to September 2015. Other figures in the paper are also based on this dataset. We count four indicators on weekdays (red diamonds), weekends (green squares), and all 92 days (blue circles) to learn the ratings of the region. (a) The number of viewers in each hour averaged over 92 days. The lowest is at 5 AM, and the two peaks are at 12 PM and 8 PM. (b) The viewing time of the top 200 channels. The abscissa is the ranking of channels according to viewing time and the ordinate is the logarithm of their viewing time. It is a straight line with a slope of −0.0147 under the semilogarithmic axis. (c) The probability function of daily average viewing time for all users. It is a lognormal distribution. (d) Cumulative proportion of viewing time of the 10 longest-watched channels, which is averaged over all users. The top 10 channels account for 85% of users' TV viewing time. These four figures illustrate that channels' TV ratings vary greatly and that viewers have channel preferences.

$$Reach = \frac{WV_{total}}{N} \times 100\% \tag{4}$$

where WT_{total} is the total viewing time of all viewers on a certain TV channel, WV_{total} is the total number of viewers in the period, T is the total duration of the period, and N is the number of viewers. Eqs. (3) and (4) are used to calculate *Rtg* and *Reach*, which respectively refer to the proportion of watched time and audience of a TV channel, standardized by Nielsen [21]. However, these two indicators have some limitations. For example, the default boot TV channel is always its local channel in a certain area. When people are in Beijing, the default boot channel is usually BTV1-HD. Thus, BTV1-HD's *Rtg* and *Reach* are overestimated when the TV is turned on. Another example occurs when a viewer does not have a target. In that case, it may stay on the default boot channel or other channels after a round of searches [18,19]. In our dataset, these local channels' *Rtg* and *Reach* are overestimated. Therefore, *Rtg*, *Reach* and their sub-index do not accurately reflect the channels' attractiveness and competitiveness. Considering these limitations, we apply the concept of PageRank to channel ranking to acquire better results.

The PageRank algorithm was proposed to rank web pages for the search engine [28,47]. The core idea behind PageRank is that if many pages link to a page, that page is important; consequently, the pages to which it links are also important. PageRank algorithm is very practical and reflects the nature of ranking to a certain extent. Therefore, it is innovatively applied in many fields, such as athlete ranking [30,31], paper ranking [32], and scientist ranking [33] and so on. The principle can also be naturally applied to TV channels. If a channel is sufficiently attractive, the audience will switch to it from other channels. Therefore, we decide to construct a transition probability matrix to rank channels according to viewers' switching behavior, defining a complicated mechanism of attractiveness diffusion from channel to channel. But

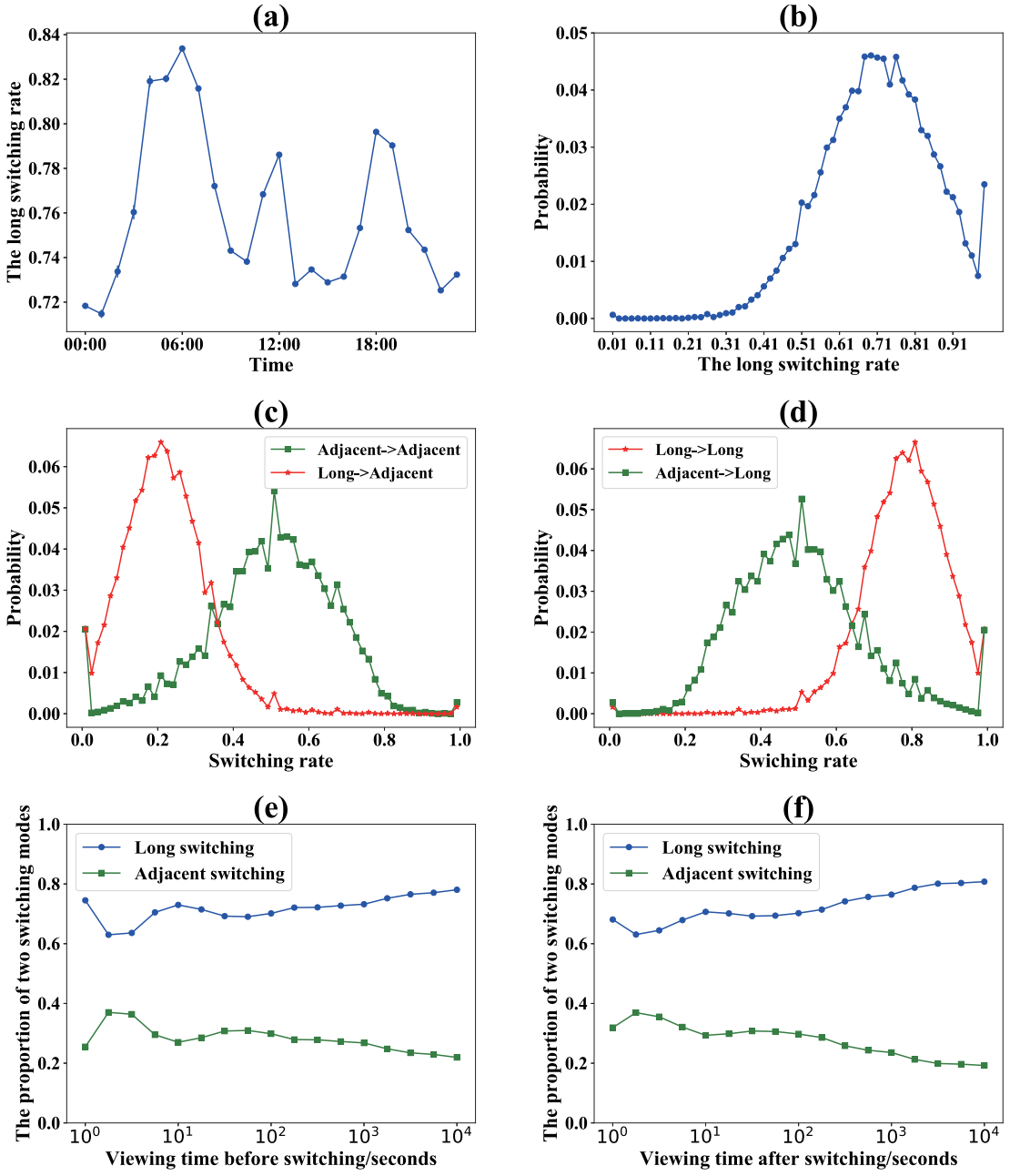

Fig. 3. Properties of user switching behavior. (a) The long switching rate l per hour averaged over 92 days. $l = \frac{L}{C}$, where L is the long switching count in an hour and C is the total switching count. l is higher at 6 AM, 12 PM and 6 PM, corresponding to when people just turn on the TV to find what they want to watch. (b) The probability function of the long switching rate l for all users with $\bar{l} = 0.71 \pm 0.14$. This means that most users prefer long switching (with few exceptions), which better reflects the audience's subjective choices, and encourages us to apply the PageRank to channel ranking. (c) The probability function of two switching rates under the adjacent switching condition. For a user, when a jump is adjacent switching, we calculate the probability that the previous switching is long or adjacent, and the conditional probability of all users forms the distribution in Fig. 3(c). The green squares denote adjacent switching, and the red diamonds denote long switching. (d) The same as (c) but for the long switching condition. (c) and (d) show that it is more likely to be a long switch after long switching with the probability of 0.8, but the probability of choosing long switching or adjacent switching after adjacent switching is equal. (c) and (d) indicate that the audience probably will continuously do long switching and also ensure that we will not delete many continuous adjacent switching records when only long switches are retained to construct the transition probability matrix. (e) Proportions of the two switching modes under different viewing time before switching. The green squares denote adjacent switching, and the blue circles denote long switching. (f) The same as (e) but under different viewing time after switching. (e)(f) illustrate that the longer viewers watch, the more likely they are to do a long switch, which is in line with the situation that the audience will not easily switch channels aimlessly when they are immersed in a channel for a long time. Fig. 3 reveals that long switching is the main channel-switching mode. It encourages us to consider ranking TV channels in terms of channel switching, especially long switching.

Table 1
Kendall's rank correlation coefficients of the ranking results.

	Rtg	*Reach*	*PR* (*all*)	*PR* (*long*)	*PR* (*adjacent*)	*PR* ($time-weighted$)	*Friday* (*long*)
Rtg	1.000	0.954	0.742	0.740	0.182	0.842	0.749
Reach		1.000	0.733	0.728	0.185	0.833	0.741
PR (*all*)			1.000	0.840	0.202	0.753	0.810
PR (*long*)				1.000	0.124	0.739	0.902
PR (*adjacent*)					1.000	0.194	0.109
PR ($time-weighted$)						1.000	0.726
Friday (*long*)							1.000

PR means PageRank.

one problem is that people sometimes browse channels aimlessly when watching TV, that is, the adjacent switching defined in the paper. These behaviors cannot fully reflect the attractiveness of the channels. On the contrary, long switching is the subjective behavior of the audience with a clear purpose. And from Section 3.3, we have concluded that long switching is the main switching mode, accounting for about 80% of all switches. Therefore, we separately extract the long switching behaviors and construct a long transition probability matrix to sort the TV channels to obtain more reasonable sorting results. We built a transition probability matrix between 8 PM and 10 PM (prime time) according to the methods mentioned above and ranked channels using PageRank, which can be iterated as shown in Eq. (5):

$$\begin{aligned} M &= \alpha S^T + \frac{1-\alpha}{N} EE^T \\ P_{t+1} &= MP_t \end{aligned} \tag{5}$$

where α is the probability of a random switch; M is the matrix for iteration; P_t is the column vector of channels' scores at time t. The iteration rule refers to the method of Page Lawrence [47].

It is universally known that viewing time is also a critical indicator to measure the quality of TV channels [16,17,21,27]. In addition to the above-mentioned transition probability matrix based on the number of switches (Eq. (1)), we also consider another method of constructing the matrix by time weighting, shown in Eq. (6):

$$\begin{aligned} d_{i,j} &= \frac{\tau_{i,j}}{\sum_{k=1}^{N} \tau_{i,k}} \\ D &= \begin{bmatrix} d_{1,1} & \cdots & d_{1,N} \\ \vdots & \ddots & \vdots \\ d_{N,1} & \cdots & d_{N,N} \end{bmatrix} \end{aligned} \tag{6}$$

where $\tau_{i,j}$ is the viewing time at j after switching from i to j; $d_{i,j}$ is the proportion of $\tau_{i,j}$ in the total viewing time of channel j. Eq. (6) is also essentially a transition probability matrix [46]. The probability is not calculated by the number of switching but by the viewing time after switching. All ranking results are shown in Fig. 4.

It is very difficult to determine which ranking is better, whether for TV channels, web pages, or athletes. Generally, there are two approaches: one is to compare the ranking with the authoritative ranking and explain why the differences are reasonable; the other is to look for evidence to support from the side [30–33]. Firstly, the traditional Rtg/Reach ranking can be used as a benchmark. We hope that the new ranking is consistent with the traditional ranking overall but with explainable differences. This is because if they are completely contradictory, it means that the new ranking is completely contrary to the audience's general cognition and it is unreliable. Meanwhile, there should also be differences between the two rankings, which can be explained from the perspective of switching behavior, and also lead to improvements in our method. Thus, we examine ranking results' Kendall's rank correlation coefficients [48] shown in Table 1. The Kendal correlation between the two rankings is about 0.7–0.8, and they do have the same trend. Fig. 4(a)(b)(c) also clearly show that the PageRank ranking and the traditional ranking are consistent at the macro level. And there also exist some outliers. It is indicated that the PageRank algorithm can provide additional information while being as reliable as the statistics of time and viewer number for evaluating channel influence. In Fig. 4(d), the ranking on the basis of adjacent transition is significantly different from the Rtg ranking (Kendall's rank correlation coefficient is only 0.182). Therefore, aimless adjacent switching is not an appropriate index for measuring a channel's attractiveness. Fig. 4(e) shows that once viewing time is taken into account in the transition probability matrix, the ranking results of PageRank and Rtg are very similar (the top 10 channels are nearly the same). So the time weighting in the paper is not as appropriate as the count weighting.

Specifically, BTV1-HD is a remarkable outlier in Fig. 4(a)(b)(c). It is the top-1 channel in Rtg and Reach rankings but drops sharply in PageRank rankings. While other HD TV channels rank very low in all rankings. This is because BTV1-HD is the default boot channel. Thus, it gains extra viewers and viewing time for appealing to local viewers. PageRank results suggest that BTV1-HD acts like a springboard. Most viewers switch out of BTV1-HD, while a few switch back during prime time. Other channels belonging to the Beijing television station have similar problems. On the contrary, some outliers attract more audiences; they have wider audiences and feature higher-quality programs. The ranking of CCTV channels targeting national audiences has increased (red dots in Fig. 4), especially CCTV-1 and CCTV-4. Some popular provincial

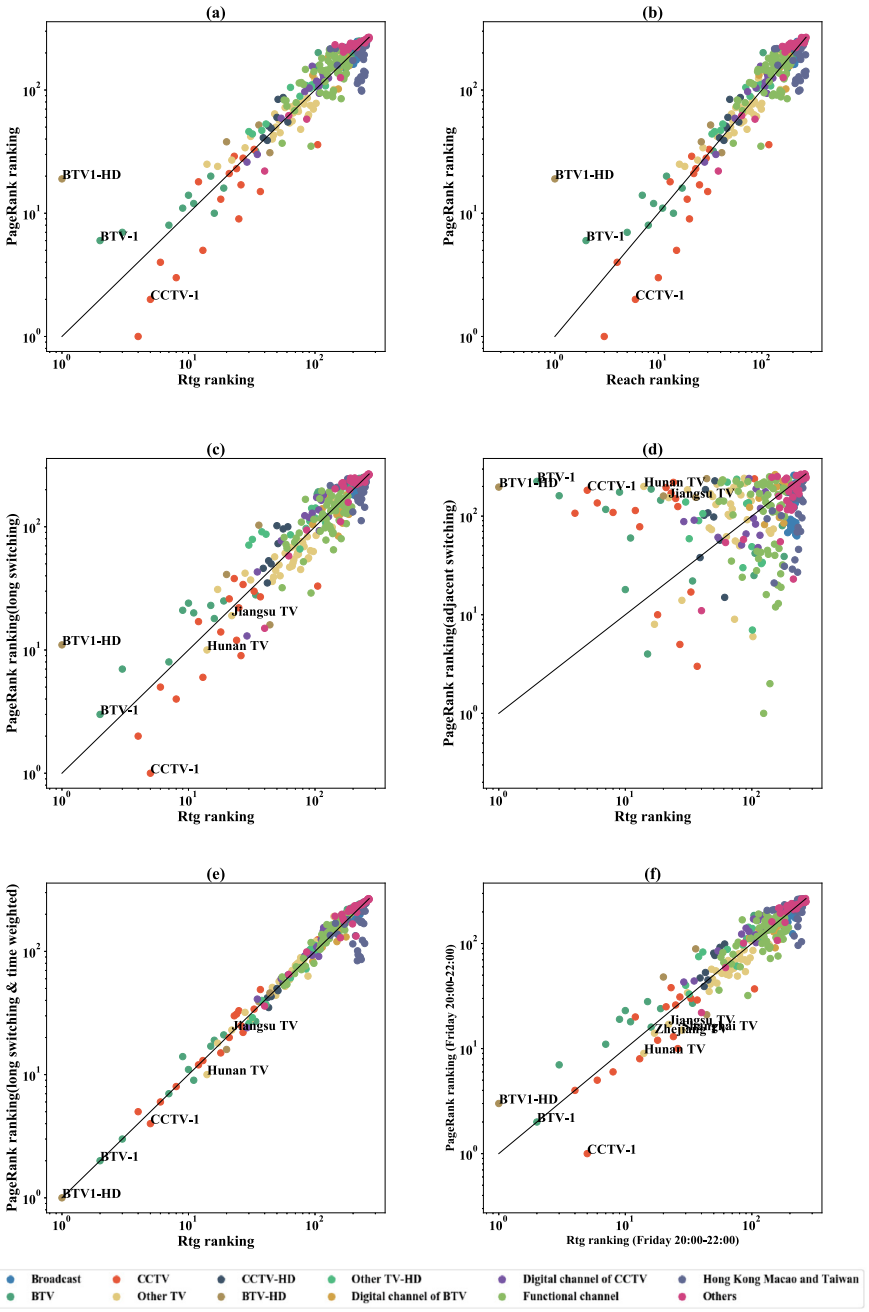

Fig. 4. Comparisons of ranking results. In the figure, the abscissa and ordinate represent the channel rankings of conventional methods and the PageRank methods, respectively. The conventional methods include Rtg ranking and Reach ranking. The PageRank methods include ranking based on the whole transition probability matrix, ranking based on the long transition probability matrix (counting only long switches), ranking based on the adjacent transition probability matrix (counting only adjacent switches), and ranking based on the time-weighted transition probability matrix. The channels are divided into 12 groups according to their features, and the rankings are logarithmic to highlight the differences at the top. (a) PageRank versus Rtg. (b) PageRank versus Reach. (c) PageRank based on long switching versus Rtg. (d) PageRank based on adjacent switching versus Rtg. (e) PageRank based on time-weighted matrix versus Rtg. (f) PageRank(long) versus Rtg on Friday evening. Fig. 4(a)(b)(c) shows that the PageRank method (whole and long) and the conventional methods are consistent at the macro level. The outliers are channels that garner much switching. Fig. 4(e) shows that once viewing time is taken into account in the transition probability matrix, the ranking results of PageRank and Rtg are very similar (the top 10 channels are nearly the same). So the time weighting in the paper is not as appropriate as the count weighting. Fig. 4(f) shows that on the most competitive Friday evening, there is a large variation between the two ranking results, and the rankings of high-quality TV channels (Hunan TV, Jiangsu TV, Zhejiang TV, Shanghai TV, etc.) in PageRank are higher, which is consistent with the fact that they won TV awards in 2015.

Table 2
Kendall's rank correlation coefficients of two TV rankings and ad price ranking.

	Adprice	*Rtg*	*PageRank* (*long*)
Adprice	1.000	0.486	0.593
Rtg		1.000	0.703
PageRank (*long*)			1.000

channels such as Hunan TV and Zhejiang TV also rank better in our ranking. In order to further illustrate the superiority of our method, we focus on the comparison of two ranking results on Friday evening (shown in Fig. 4(f)), on which the competition for ratings is the fiercest and TV channels tend to broadcast their trump programs. Hunan TV, Jiangsu TV, Zhejiang TV, and Shanghai TV rose by 36%, 18%, 46%, and 23% respectively in our ranking. Corresponding to this, Hunan TV, Jiangsu TV, Zhejiang TV, and Shanghai TV won the most influential provincial TV in 2015 (the selection was sponsored by the magazine China Radio Film and TV under the National Press and Publication Administration).

Second, the advertising price also reflects the sponsor's evaluation of the TV quality to some extent. Therefore, we collected the advertising prices of all provincial TVs' trump programs on Friday evening, and calculated the Kendall's rank correlation coefficients of the two rankings and advertising price ranking (shown in Table 2). PageRank is more in line with the channel quality reflected by advertising ranking than Rtg (0.593 versus 0.486). Notably, advertising quotation is not only related to the channel influence but also involves regional economy and policies, so the correlation between advertising ranking and rating ranking will not be particularly high. PageRank has raised the Kendall's rank correlation coefficient from 0.486 to 0.593, which is a 20% increase. We consider it a relatively large improvement.

Note that the time complexity of PageRank is $O(\epsilon(t)n^2)$, where n is the node number, $\epsilon(t)$ is the number of iteration steps. The time complexity of classical methods *Rtg* or *Reach* is $O(n)$. Every coin has two sides. Although PageRank's time complexity is higher, it uses the collective wisdom of the network to interpret each link as an internal vote, reducing the impact of set-box configuration, idle viewing, etc., on channel ranking. The ranking results are more in line with reality, which is evidenced by influential TV awards sponsored by China Radio Film and TV and advertising expenses.

5. Community detection of channels

If the audience shows similar behavior characteristics when watching several TV channels, it indicates that these channels are similar in program content or positioning in the TV channel system. A typical example is that people who pay close attention to the news may jump between several news channels at about 7 PM every day. After the news broadcast, they will jump to other channels or turn off the TV. Therefore, we aim to classify TV channels based on their similarity in users' switching behavior, which can be calculated from the transition probability matrix. If the audience switch from A to C_1, C_2, and C_3 with the probability of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and they also switch from channel B to C_1, C_2, C_3 with probability $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Starting from A and B and arriving at the same channel with the same probability, it can be concluded that A and B are similar in audience preferences and they play the same role in the audience's channel-switching. In Eq. (1), the ith row of the transition probability matrix S represents the probabilities of the i_{th} channel switching to other channels, denoted by v_i. In this paper, we call it a switch-out vector. We represent channels' similarities in preferences with similarities of their switch-out vectors. We construct a similarity network in which the nodes are the TV channels, and two nodes are linked when they have similar switching. This similarity network is denoted by the matrix R, which is calculated as shown in Eq. (7):

$$\begin{aligned} v_i &= (s_{i,1}, s_{i,2}, \ldots, s_{i,N-1}, s_{i,N}) \\ r_{i,j} = corr(v_i, v_j) &= \frac{\sum_{k=1}^{N}(s_{i,k} - \bar{s_i})(s_{j,k} - \bar{s_j})}{\sqrt{\sum_{k=1}^{N}(s_{i,k} - \bar{s_i})^2}\sqrt{\sum_{k=1}^{N}(s_{j,k} - \bar{s_j})^2}} \\ R &= \begin{bmatrix} r_{1,1} & \cdots & r_{1,N} \\ \vdots & \ddots & \vdots \\ r_{N,1} & \cdots & r_{n,N} \end{bmatrix}. \end{aligned} \tag{7}$$

R is a symmetrical fully connected matrix with noise caused by some occasional long switching, and performing community detection directly on R does not provide much meaningful information. Therefore, we first use the MST algorithm to extract the backbone of the network to obtain better detection results. Then we select the fast-unfolding method for community detection because it is widespread and proved to be very efficient when dealing with large networks [41]. We partition the communities on two networks (based on the long switching) consisting of the top 100 channels and all channels. The former shows the classification of the top channels in more detail, while the latter can be used to determine all the functional communities at the overall level.

Among 99 channels (one isolated channel is deleted), 8 communities are detected. We first measure the intra-community switching rate *intra*. It is about 0.94±0.06, with a large fluctuation in the late night, as shown in Fig. 6(a). Fig. 6(b) is the distribution of all users' intra-community switching rates with $\overline{intra} = 0.95 \pm 0.03$. The majority of switches

Table 3
Average generalized measures obtained for the 4 categories.

Category	D^{out}	D^{in}	I^{out}_{ext}	I^{in}_{ext}	H^{out}	H^{in}	I^{out}_{int}	I^{in}_{int}
1	0.29	0.55	0.32	0.64	0.78	−0.61	1.30	−1.13
2	−0.48	0.95	−0.46	0.99	−0.60	0.85	−0.39	−0.48
3	1.98	0.77	1.96	0.50	1.54	0.53	0.67	−0.42
4	−0.22	−0.64	−0.29	−0.60	−0.41	−0.26	−0.25	0.62

Table 4
Categories detected with the generalized measures.

Category	Role	Typical channels	Proportion
1	Source within the community	BTV-1, CCTV-Kids	18%
2	Sink among communities	Zhejiang TV, Hunan TV	22%
3	Connector	BTV-9, BTV-KAKU	10%
4	Sink within the community	CCTV-8, CCTV-NEWS	49%

occur within the community, proving the effectiveness of community division. Most BTV and CCTV channels form a community (CCTV is the Chinese national television station and BTV is the capital television station). Popular regional TV channels such as Zhejiang TV also form a community, while less attractive provincial channels such as Qinghai TV are grouped into another community. In addition, the CCTV-HD community and the regional TV-HD community exist. Remarkably, YouManKaTong, Eagle Animation, CCTV-Kids, and similar channels form a community for children. Fortune, Drama, Global Go-SD, Chess, etc. form a community for grownups; these channels integrate finance, information, and entertainment. In addition, CCTV-stock, Family Financing, Oriental Biz, etc. form a special financial community. These channels are clustered due to their functions, so we term them functional communities in the paper. The community detection results indicate that channels in the same community have similar main contents and audience groups. For instance, children who like to watch cartoons will switch between various children's channels. Hence, the switch-out vectors of channels for children have a high degree of similarity, which results in these channels being clustered on the network. The formation of the grownup community and financial community follow the same pattern. In addition, the partition results are also consistent with the TV ratings to a certain extent. For example, the channels in the CCTV and BTV community rank close to the top in PageRank and Rtg. Similarly, channels with lower ratings tend to gather in a community.

However, the community ownership of some channels is not consistent with common sense. CCTV-music belongs to the CCTV, but it is assigned to the grownup community. Because the main content of the channel is to broadcast concerts, providing leisure for adults. Hunan TV is detected as a CCTV channel or BTV channel, but it does not belong to CCTV or BTV. It indicates that Hunan TV has become from a provincial TV to a heavyweight TV channel facing the national audience with its high-quality TV programs. Although BTV 9, BTV 6, and BTV KAKU belong to BTV, the audience often switches from these channels to other provincial channels. So they are assigned to other regional TV channels community. For further quantifiable explanation, we use the generalized measures [49,50] to characterize the position of vertices relative to the community structure. We calculate diversity (D^{out}, D^{in}), external intensity (I^{out}_{ext}, I^{in}_{ext}), heterogeneity (H^{out}, H^{in}), internal intensity (I^{out}_{int}, I^{in}_{int}) of the 100 channels, and divide them into 4 categories according to the 8 indicators (see Appendix). The results are shown in Tables 3 and 4. Category 1 has a positive outcoming *internal intensity*, which means the audience always jumps to other channels in the same community through these channels, such as BTV-1, and BTV-1-HD. Category 1 is the source within the community, and the channel belonging to Category 1 can be regarded as a springboard channel. Category 2 has a positive incoming *external intensity*, named sink among channels, which means the audience is often attracted to these channels from other communities. Some high-quality provincial channels underestimated in traditional rating measurement belong to the category, such as Hunan TV. Category 3 has higher *diversity*, *external intensity*, *heterogeneity*, and outcoming *internal intensity*. The audience switches into or out of the community through these channels, which is the role of the connector, like BTV-9, BTV-6, and BTV KAKU mentioned at the beginning of this paragraph. TV stations can consider putting some high-quality programs on these channels to prevent the loss of audience. Category 4 has positive incoming *internal intensity*. Contrary to Category 1, it is the sink within the community. The audience always switches to these channels after browsing in the community. Many of these channels broadcast TV dramas or news, such as CCTV-8, and CCTV-NEWS. In this way, each TV channel can refer to the characteristics of its community to arrange programs, to better retain the audience and improve their ratings.

As shown in Fig. 7, performing community detection on the network of all channels reveals several comprehensive, macroscopic features. In Fig. 7, there are also the BTV community, the popular TV community, the HD-TV community, and so on, which is consistent with Fig. 5. For the functional community, in addition to the children's community, the grownup community and the financial community mentioned above, there is a shopping community, including Enjoy shopping, Home shopping, BAMC shop, etc. In addition, there is a daily leisure community, including Channel Pet, BAMC Music, and others as well as a community consisting of radio channels. In general, the community detection of all channels

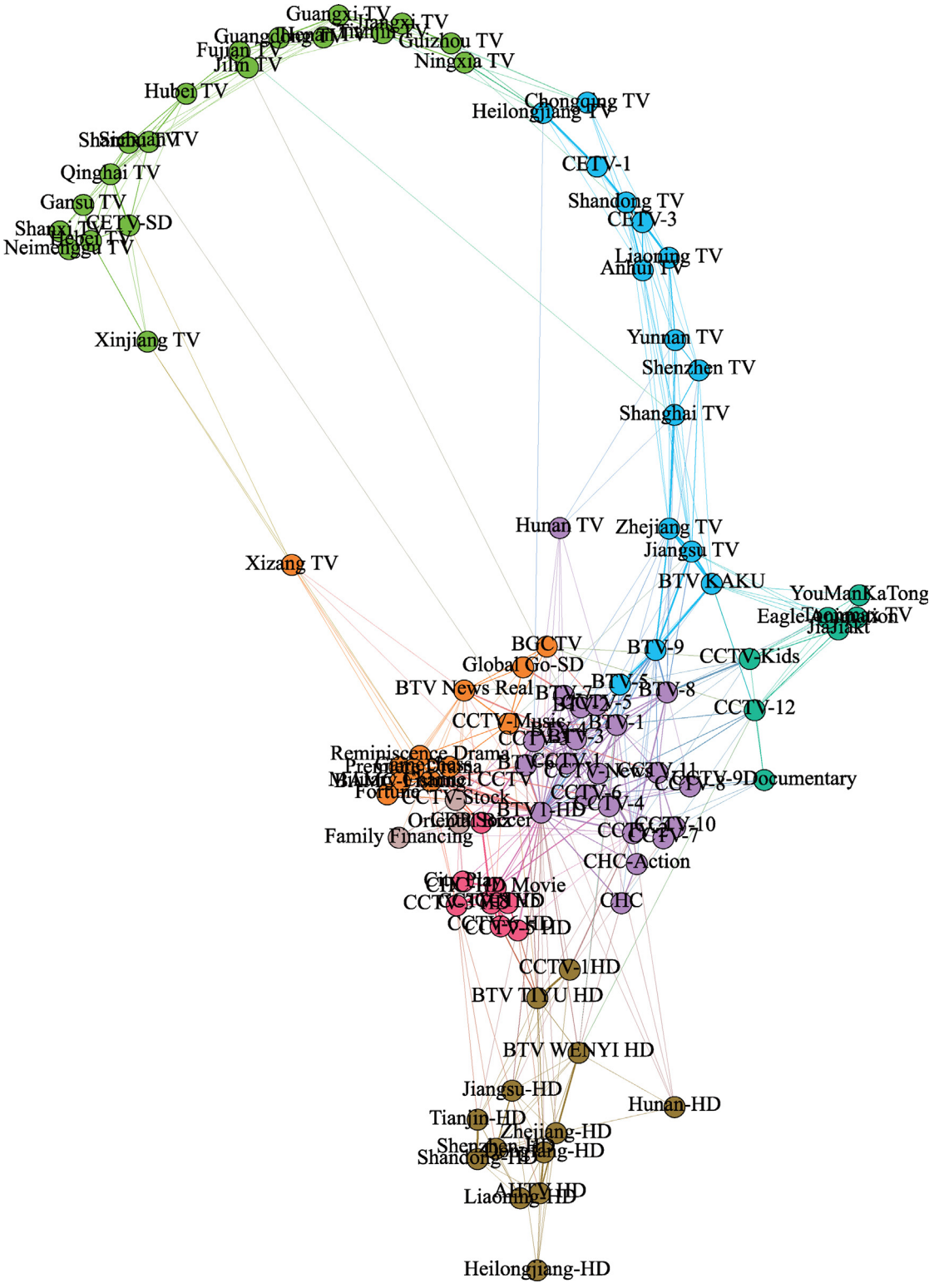

Fig. 5. Visualization of community detection in the PageRank top 100 channels. The clustering results reveal 8 communities, which include a CCTV and BTV community, a CCTV-HD community, a regional TV-HD community, a popular regional TV community, a less popular regional TV community, a financial community, a children's community, and a grownup community. Community detection is achieved using the fast-unfolding method which has lower time complexity.

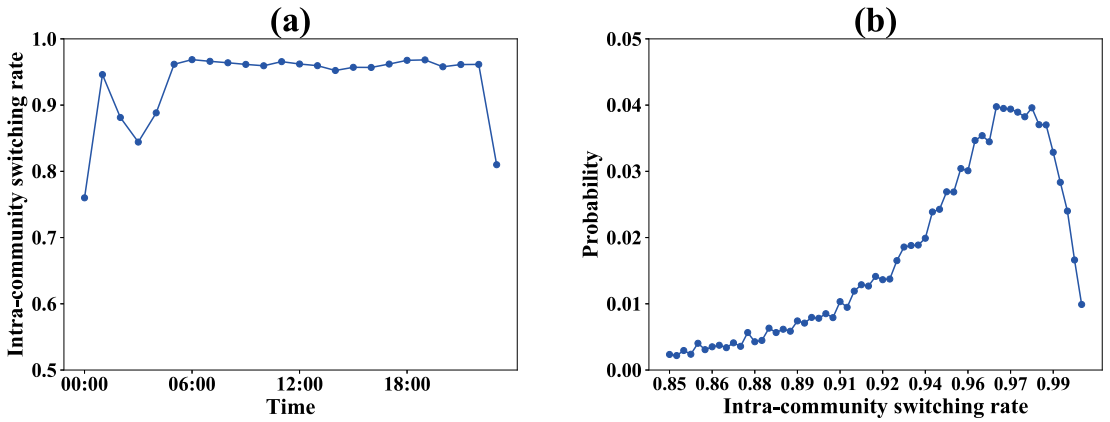

Fig. 6. Switching rate at the community level. (a) The intra-community switching rate *intra* per hour averaged over 92 days, which is lower at late night when the audience looks for TV programs. (b) The probability function of the intra-community switching rate *intra* for all users with $\overline{intra} = 0.95 \pm 0.03$. This means that the vast majority of channel switching occurs within the community after community detection.

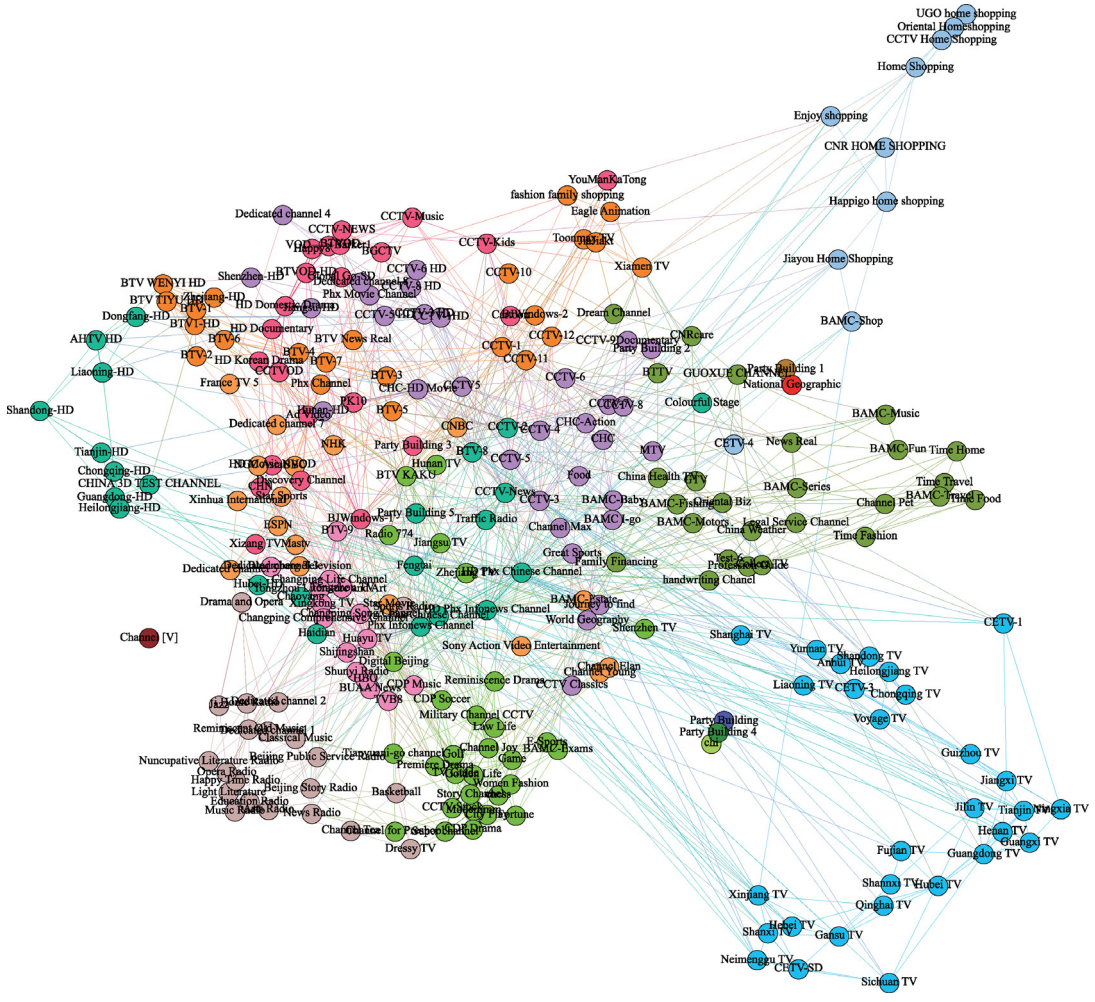

Fig. 7. Visualization of community detection in all channels. These results show that there are 11 major communities, including a BTV community, a CCTV community, a TV-HD community, a popular regional TV community, a less popular regional TV community, a children's community, a grownup community, a financial community, a daily leisure community, a shopping community, and a radio community. Community detection here is also achieved using the fast-unfolding method.

is consistent with that of the top 100 on the whole, but there are also some differences. For example, CCTVs and BTVs are split into two communities. The functional communities are further subdivided, and there is a female-oriented functional community including Dreesy TV and Modern Women, and a male-oriented functional community including BAMC-Fishing and BTTV (military-related).

6. Conclusion

In this paper, we analyze three months of TV-watching records collected by set-top boxes in a city in China. The core of our work is to quantify the TV channels' influence and reveal their hidden community structures by analyzing switching behavior. First, basic statistics are collected from the perspectives of channels and audiences. These statistics reveal that the ratings among channels are extremely unbalanced and that audiences have strong viewing preferences: the top channels attract the largest audiences. Then, we define two switching modes: long switching and adjacent switching. We explore their properties, including hourly variations in the long switching rate, the probability distribution of the long switching rate, the conditional probability distributions of the two switching modes, and the relationship between the two switching modes and watch time. The results show that long switching is the main switching mode and that people usually have clear targets when switching channels.

Long switching represents people's subjective thoughts, indicating the attractiveness diffusion among TV channels. Therefore, we construct a channel transition probability matrix based on the switching behaviors. We apply the PageRank algorithm to rank the channels and compare the results with those of traditional ranking methods. Interestingly, we find that all the ranking method results are consistent at the macro level, but that some outliers exist, which can reveal the advantages of using the transition matrix for ranking. Finally, we partition TV-channel networks based on the similarity of channel switch-out vectors. The partitioned results not only reflect the audience rating of channels but also reveal the function of each community.

Generally, our results are meaningful theoretically. It can help TV stations make more appropriate program arrangements according to their rankings in PageRank and positions in functional communities. In practice, our method and results can be used to provide personalized TV recommendations for the audience. For example, when the set-top box is turned on, it is suggested to automatically jump to the channel that the user frequently switches to instead of the default channel. Another example of a possible application is to change the adjacent relationship between channels according to the switching behavior of individual users, to reduce the cost of changing channels for users and help TV service providers retain customers. Some extensions could be made based on this work. For TV rating analysis, the transition probability matrix, the basis of PageRank and community detection, may be better designed. We simply experimented time-weighted transition probability matrix and the results showed little improvements. Considering the transition probability multiplied by an exponential time factor may make better use of viewing time to improve ranking performance. As for communities formed by TV channels, we can further explore whether the community a channel belongs to will change and whether the position of a channel in the community will change if we can acquire a longer viewing record. Furthermore, the principles used in the paper can be extended to other systems such as online shopping, and online course learning. For example, we can properly rank online courses according to students' online learning trajectories, or detect communities of courses to judge which are basic courses, and which are advanced courses. Our methodology can be widely applied to systems with switching behavior.

CRediT authorship contribution statement

Mingyan Wang: Software, Validation, Writing – original draft. **An Zeng:** Methodology, Formal analysis. **Xiaohua Cui:** Conceptualization, Methodology, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 11775020, and 11675001).

Appendix. Generalized measures used to characterize the position of vertices

Dugué proposed 4 measures aiming at representing separately the aspects of connectivity: *diversity*, *external intensity*, *heterogeneity*, and *internal intensity*. Because we deal with directed links that represent switching, each one of these measures exists in two versions: incoming and outgoing, effectively resulting in 8 measures. All the measures are expressed as z-scores within the community.

Diversity The diversity D evaluates the number of communities to which a node is connected (other than its own).

External intensity The external intensity I_{ext} of a node measures the amount of links it has with communities other than its own.

Heterogeneity The heterogeneity H of a node measures the variation of the number of links a node has, from one community to another.

Internal intensity The internal intensity I_{int} of a node measures the amount of links it has with its own communities.

Then the k-means algorithm is used to classify the nodes. In this paper, k=4 is the optimal choice determined by the number of nodes and the category result.

References

[1] J. Mervis, Agencies rally to tackle big data, 2012.
[2] A.S. Pentland, The data-driven society, Sci. Am. 309 (4) (2013) 78–83.
[3] J.J. Hox, Computational social science methodology, anyone? Methodology (2017).
[4] Z. Wang, X. Yu, N. Feng, Z. Wang, An improved collaborative movie recommendation system using computational intelligence, J. Vis. Lang. Comput. 25 (6) (2014) 667–675.
[5] P. Vilakone, K. Xinchang, D.-S. Park, Personalized movie recommendation system combining data mining with the k-clique method, J. Inf. Process. Syst. 15 (5) (2019) 1141–1155.
[6] Y. Yoshimura, S. Sobolevsky, J.N. Bautista Hobin, C. Ratti, J. Blat, Urban association rules: uncovering linked trips for shopping behavior, Environ. Plan. B: Urban Anal. City Sci. 45 (2) (2018) 367–385.
[7] R. Di Clemente, M. Luengo-Oroz, M. Travizano, S. Xu, B. Vaitla, M.C. González, Sequences of purchases in credit card data reveal lifestyles in urban populations, Nature Commun. 9 (1) (2018) 1–8.
[8] J.L. Toole, C. Herrera-Yaqüe, C.M. Schneider, M.C. González, Coupling human mobility and social ties, J. R. Soc. Interface 12 (105) (2015) 20141128.
[9] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, M.C. González, The timegeo modeling framework for urban mobility without travel surveys, Proc. Natl. Acad. Sci. 113 (37) (2016) E5370–E5378.
[10] H.-J. Li, W. Xu, S. Song, W.-X. Wang, M. Perc, The dynamics of epidemic spreading on signed networks, Chaos Solitons Fractals 151 (2021) 111294.
[11] J. Gao, Y.-C. Zhang, T. Zhou, Computational socioeconomics, Phys. Rep. 817 (2019) 1–104.
[12] D.V. Shah, J.N. Cappella, W.R. Neuman, Big data, digital media, and computational social science: Possibilities and perils, Ann. Am. Acad. Political Soc. Sci. 659 (1) (2015) 6–13.
[13] J. Webster, P.F. Phalen, L. Lichty, Ratings Analysis: Theory and Practice, Routledge, 2005.
[14] K. Buzzard, Tracking the Audience: The Ratings Industry from Analog to Digital, Routledge, 2012.
[15] S.A. Spiller, Opportunity cost consideration, J. Consum. Res. 38 (4) (2011) 595–610.
[16] R. Weber, Methods to forecast television viewing patterns for target audiences, in: Communication Research in Europe and Abroad Challenges of the First Decade, DeGruyter, Berlin, 2002.
[17] D. Meyer, R.J. Hyndman, The accuracy of television network rating forecasts: The effects of data aggregation and alternative models, Model Assist. Stat. Appl. 1 (3) (2006) 147–155.
[18] J. Chai, X. Pan, F. Yin, J. Chai, Study on the impact of big data on radio and television ratings, in: 2015 International Conference on Automation, Mechanical Control and Computational Engineering, Atlantis Press, 2015.
[19] J. Gillan, Television and New Media: Must-Click TV, Routledge, 2010.
[20] R.L. Graham, P. Hell, On the history of the minimum spanning tree problem, Ann. Hist. Comput. 7 (1) (1985) 43–57.
[21] E.E. Massetti, Audience rating system for digital television and radio, uS Patent 5, 974, 299 (Oct. 26 1999).
[22] L. Yin, X. Liu, A gesture of compliance: Media convergence in china, Media Cult. Soc. 36 (5) (2014) 561–577.
[23] S. Wakamiya, R. Lee, K. Sumiya, Twitter-based tv audience behavior estimation for better tv ratings, in: DEIM Forum, 2011, http://db-event.jpn.org/deim2011/.
[24] P.J. Danaher, T.S. Dagger, M.S. Smith, Forecasting television ratings, Int. J. Forecast. 27 (4) (2011) 1215–1240.
[25] M. Takahashi, S. Clippingdale, M. Naemura, M. Shibata, Estimation of viewers' ratings of tv programs based on behaviors in home environments, Multimedia Tools Appl. 74 (19) (2015) 8669–8684.
[26] N. Ma, S. Zhao, Z. Sun, X. Wu, Y. Zhai, An improved ridge regression algorithm and its application in predicting tv ratings, Multimedia Tools Appl. 78 (1) (2019) 525–536.
[27] Y. Yang, D. Xu, F. Nie, J. Luo, Y. Zhuang, Ranking with local regression and global alignment for cross media retrieval, in: Proceedings of the 17th ACM international conference on Multimedia, 2009, pp. 175–184.
[28] A.N. Langville, C.D. Meyer, Deeper inside pagerank, Internet Math. 1 (3) (2004) 335–380.
[29] C. Ding, X. He, P. Husbands, H. Zha, H. Simon, Pagerank, hits and a unified framework for link analysis, in: Proceedings of the 2003 SIAM International Conference on Data Mining, SIAM, 2003, pp. 249–253.
[30] F. Radicchi, Who is the best player ever? a complex network analysis of the history of professional tennis, PLoS One 6 (2) (2011) e17249.
[31] S. Motegi, N. Masuda, A network-based dynamical ranking system for competitive sports, Sci. Rep. 2 (2012) 904.
[32] J.D. West, T.C. Bergstrom, C.T. Bergstrom, The eigenfactor metricstm: A network approach to assessing scholarly journals, Coll. Res. Libr. 71 (3) (2010) 236–244.
[33] F. Bibi, H.U. Khan, T. Iqbal, M. Farooq, I. Mehmood, Y. Nam, Ranking authors in an academic network using social network measures, Appl. Sci. 8 (10) (2018) 1824.
[34] C. Jing, B.J. Aronow, A.G. Jegga, Disease candidate gene identification and prioritization using protein interaction networks, BMC Bioinformatics 10 (2009) 1–14.
[35] G. Ghoshal, A.-L. Barabási, Ranking stability and super-stable nodes in complex networks, Nature Commun. 2 (1) (2011) 1–7.

[36] H.-J. Li, L. Wang, Z. Bu, J. Cao, Y. Shi, Measuring the network vulnerability based on markov criticality, ACM Trans. Knowl. Discov. Data (TKDD) 16 (2) (2021) 1–24.
[37] H.-J. Li, Z. Wang, J. Pei, J. Cao, Y. Shi, Optimal estimation of low-rank factors via feature level data fusion of multiplex signal systems, IEEE Trans. Knowl. Data Eng. (2020).
[38] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3–5) (2010) 75–174.
[39] H. Li, W. Xu, C. Qiu, J. Pei, Fast markov clustering algorithm based on belief dynamics, IEEE Trans. Cybern. (2022).
[40] H.-J. Li, L. Wang, Y. Zhang, M. Perc, Optimization of identifiability for efficient community detection, New J. Phys. 22 (6) (2020) 063035.
[41] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (10) (2008) P10008.
[42] E. Ferrara, P. De Meo, S. Catanese, G. Fiumara, Detecting criminal organizations in mobile phone networks, Expert Syst. Appl. 41 (13) (2014) 5733–5750.
[43] A.L. Traud, P.J. Mucha, M.A. Porter, Social structure of facebook networks, Physica A 391 (16) (2012) 4165–4180.
[44] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, A. Provetti, Extraction and analysis of facebook friendship relations, in: Computational Social Networks, Springer, 2012, pp. 291–324.
[45] Q. Chang, The development history of chinese tv audience ratings and its critical thinking, Adv. Journal. Commun. 2014 (2014).
[46] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, SIAM, 1999.
[47] L. Page, S. Brin, R. Motwani, T. Winograd, The Pagerank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford InfoLab, 1999.
[48] M.G. Kendall, Rank Correlation Methods, 1948.
[49] R. Guimera, L.A. Nunes Amaral, Functional cartography of complex metabolic networks, Nature 433 (7028) (2005) 895–900.
[50] N. Dugué, V. Labatut, A. Perez, A community role approach to assess social capitalists visibility in the twitter network, Soc. Netw. Anal. Min. 5 (1) (2015) 1–13.