Modeling Channel Popularity Dynamics in a Large IPTV System

Tongqing Qiu[†] Zihui Ge[‡] Seungjoon Lee[‡] Jia Wang[‡] Qi Zhao[‡] Jun (Jim) Xu[†] *

[†] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
[‡] AT&T Labs – Research, Florham Park, NJ 07901,USA
Email: tongqqiu@cc.gatech.edu {gezihui, slee, jiawang, qzhao}@research.att.com jx@cc.gatech.edu

ABSTRACT

Understanding the channel popularity or content popularity is an important step in the workload characterization for modern information distribution systems (e.g., World Wide Web, peer-to-peer file-sharing systems, video-on-demand systems). In this paper, we focus on analyzing the channel popularity in the context of Internet Protocol Television (IPTV). In particular, we aim at capturing two important aspects of channel popularity – the distribution and temporal dynamics of the channel popularity. We conduct in-depth analysis on channel popularity on a large collection of user channel access data from a nation-wide commercial IPTV network. Based on the findings in our analysis, we choose a stochastic model that finds good matches in all attributes of interest with respect to the channel popularity. Furthermore, we propose a method to identify subsets of user population with inherently different channel interest. By tracking the change of population mixtures among different user classes, we extend our model to a multi-class population model, which enables us to capture the moderate diurnal popularity patterns exhibited in some channels. We also validate our channel popularity model using real user channel access data from commercial IPTV network.

Categories and Subject Descriptors

C.2.3 [**Computer System Organization**]: Computer-Communication Networks—*Network Operations*; C.4 [**Computer System Organization**]: Performance of Systems—*Modeling techniques*

General Terms

Measurement, Performance

Keywords

IPTV, Channel Popularity, Network Measurement, Modeling

*This work is supported in part by NSF grants CNS-0519745, CNS-0626979, CNS-0716423, and CAREER Award ANI-023831.

SIGMETRICS/Performance'09, June 15–19, 2009, Seattle, WA, USA.

1. INTRODUCTION

Understanding the channel popularity or content popularity is an important step in the workload characterization for modern information distribution systems such as World Wide Web, P2P file-sharing systems, IPTV networks, video-on-demand (VOD) systems, content distribution networks, publish/subscribe systems, and RSS feeds distribution systems. The proper modeling of the distribution of user's interest in various contents and media in the system is a key building block for system design and performance analysis. For example, it has been well known that web site popularity is highly skewed and can be characterized by a Zipf-like distribution [14], a factor that carries important implication in evaluating different DNS caching policies. Similar popularity skewness has also been observed in other systems including P2P file-sharing [7], VOD [20], web servers [1], and IPTV [4].

Another important aspect of the content or channel popularity is its temporal dynamics, which captures the popularity changes over time. Examples of such dynamics are the shift of users' search and download interest among files in a P2P file-sharing system, the change of subscriber numbers among different topics in a publish/subscribe system, and the growth/shrink of community groups in a social network. The popularity dynamics can be either attributed to the stochastic nature of users' interest at the time, or attributed to the change of active users' population at the time, or a combination of both. Understanding the process of popularity dynamics can provide important insight into service design and optimization. For example, properties on TV channel popularity dynamics are an essential piece of information for evaluating the proposal of using peer-assisted TV stream distribution (e.g., [12]) in an IPTV system.

In this paper, we focus on analyzing the channel popularity in the context of Internet Protocol Television (IPTV). Our goal is to construct mathematical models to capture the distribution and the time dynamics of channel popularity. This is motivated by recent booming growth among telecommunication companies around the world in the rapid deployment of the IPTV infrastructure and service expansion, and hence the increasing demand in the workload characterization and performance evaluation of the IPTV system. However, we believe that the basic principle and methodology used herein are applicable to other domains (e.g., RSS feeds, news groups).

Our analysis is based on a large collection of user channel access data from a nation-wide commercial IPTV network[1]. We conduct an in-depth analysis of the user channel switch activities and study

[1]To protect the identity of the IPTV network subscribers, individual set top boxes were assigned a non-identifiable ID number for purposes of this research. The authors did not have access to subscriber's identity or address of individual set top boxes.

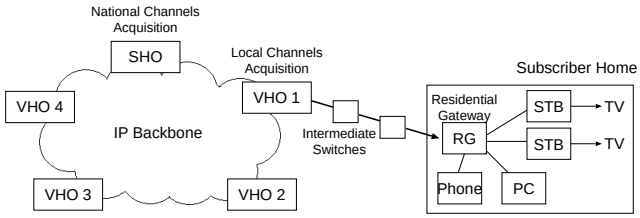


Figure 1: IPTV System Architecture

the channel popularity for different channels, at different time and different aggregation scales (ranging from minutes to days). Then, we identify a stochastic model that matches well in all attributes of interest with respect to the channel popularity. We also explore subsets of user population and investigate whether they intrinsically have different channel preferences from others. We then construct multi-class population models that capture the non-stationary behavior of channel popularity exhibited by its diurnal patterns, which has been reported in previous measurement study [4].

Our contributions can be summarized as follows:

- We observe that channel popularity is highly skewed and can be well captured by a Zipf-like distribution. This holds true at different times of day and at various aggregation scales. We find that the popularity of each individual channel has an exponentially decaying autocorrelation function, a common behavior across different channels. We also examine the change of two channel popularity vectors at adjacent time bins while varying the aggregation step. We find that the cosine similarity between channel popularity vectors exhibits an interesting multi-scale behavior, forming a *V*-shape when the aggregation scale increases from minutes to days.

- We model channel popularity dynamics as an Ornstein-Uhlenbeck process and find that it matches remarkably well with respect to the above properties. The success in capturing the underlying channel popularity dynamics enables our model to produce a satisfying result for channel popularity prediction.

- We develop a method to identify subsets of user population with inherently different channel interests. We apply the K-means clustering algorithm on various features of users, and use a symmetric uncertainty measure and hypothesis test to evaluate the significance of channel popularity difference. By tracking the change of population mixtures among different user classes, we extend our model to a multi-class population model, which enables us to capture the moderate diurnal popularity patterns exhibited in some channels.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the IPTV system and describe the data set used in this study. In Section 3, we present our analysis result on channel popularity distribution and channel popularity time dynamics. In Section 4, we present our stochastic model that captures these behaviors. We extend our analysis to explore the presence of multiple classes of user population in the system and accordingly adapt our model in Section 5. We review related work in Section 6 and conclude in Section 7.

2. OVERVIEW OF IPTV SYSTEM

Figure 1 shows a typical IPTV service system, in which live TV streams are encoded in a series of IP packets and delivered to users through the residential broadband access network. The SHO (Super Head-end Office), which is the primary source of television content, digitally encodes video streams received externally (e.g., via satellite) and transmits them to multiple VHOs (Video Head-end Offices) through a high-speed IP backbone network. The VHOs, each responsible for a metropolitan area, in turn acquire additional local contents (e.g., local news), perform some further processing (e.g., advertisement insertion) and transmit the processed TV streams to end users upon request. Inside a residential home, RG (Residential Gateway) connects to a modem and one or more STBs (Set-Top Boxes) with coaxial cable, receiving and forwarding all data, including live TV streams, STB control traffic, VoIP and Internet data traffic, into and out of the subscriber's home. Finally, behind a STB, there connects a TV.

On the user side, IPTV subscribers use a vendor/provider customized remote controller to control the STB. Similar to conventional TV remote controller, one may use *Up/Down* buttons to sequentially switch channels, use *Return* button to jump back to the channel previously watched, or enter a channel number to jump directly to a specific channel. Many IPTV providers add the capability for a small number of user-defined *favorite channel list*, so that one can easily switch between or scan through the favorite channels. Furthermore, most STBs support the DVR (Digital Video Recording) feature, in which with the help of a local hard drive, a user can pause, rewind, fast forward (up to live play), and record the TV program being played. Some IPTV providers support one channel being recorded to DVR while another channel being played live on TV. Since IPTV utilizes IP multicast to deliver video streams from VHOs to STBs, and due to the overhead of IGMP multicast group management process, there is typically a delay of up to a few seconds when user switch from one channel to another[2]. This limitation is likely to motivate IPTV users to perform more targeted channel switches than randomly or sequentially channel scans compared to users from conventional TV systems.

2.1 Data Set Description

The data we use in this study are collected from a large-scale IPTV service provider in the United States, which has over one million subscribers and over two million STBs spreading across three different time zones, carrying over 500 different live TV channels. With strict adherence to legal and privacy policy requirements, we have obtained anonymous subscribers' STB logs, control plane messages, network configuration data, and TV channel lists from this service provider. We construct user activities with respect to turning on/off STBs, switching channels, and playing live or recorded TV program by combining these data altogether. We associate each of the user activities in the anonymous STB logs with its origin STB and a timestamp (with the resolution of one second) in this study. There are a few caveats with the quality of the data in this study. The channel switch events capture user requests logged at the STB, with the timestamp indicating the time that the request is received at the STB. Note this is different from the time when the request arrives at the VHO, and different from the time when the streaming content is received at the STB. Requests that are very rapidly followed by a subsequent request in time may not be recorded by the STB, hence are missing from our study. Furthermore, we do not have detailed TV program information when DVR is used – from the STB logs, we know that a recorded video is being played, but we do not know what is played. Therefore, in this paper we exclude the user activities when the STB is tuned to DVR mode. Note that no personally identifiable information is used in the analysis and all data processing is conducted in accordance with the privacy policy in place.

We have collected the aforementioned data for one month (June) in 2008. To account for the time zone differences, we divide the

[2]How to reduce this delay is an active research area [16].

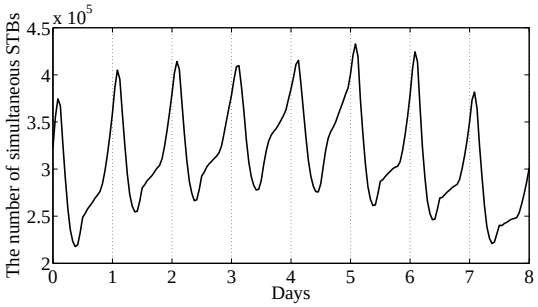


Figure 2: The number of online STBs for each hour during a week.

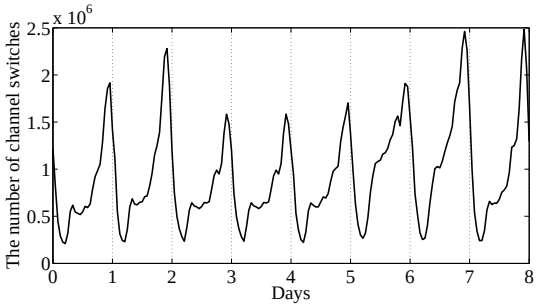


Figure 3: The number of channel switches for each hour during a week.

STBs into three groups, one for each time zone and study their channel popularity properties separately. In the rest of the paper, we will use the data from the Eastern time zone while the results including the other time zones are quantitatively similar.

3. ANALYZING CHANNEL POPULARITY

Channel popularity needs to be precisely defined before we can present our analysis result. There are two common used metrics to measure the channel popularity. The first is based on *channel access frequency* which is defined as the number of channel switching requests to the channel. The other is based on *channel dwell time* defined as the amount of time STBs stay tuned in the channel. They measure two different aspects of channel popularity: weighted by visit frequency vs. weighted by watching time.

Figure 2 and Figure 3 show the time series of the number of online STBs and the total number of channel switches respectively. As expected, we find both of them highly variable, exhibiting strong diurnal patterns. To account for the variation due to the change in the number of active users, we focus on the probability distribution (i.e., normalized among all channels) instead of the absolute value of the channel popularity measure.

3.1 Distribution of Channel Popularity

We first examine the long term distribution of channel popularity (over the entire month) of all channels using both metrics. Figure 4 shows the cumulative distribution function (CDF) of channel popularity ranked by access frequency and dwell time. We observe a close match between the CDF curves of the two different popularity metrics. Both distribution functions exhibit high skewness – the top 10% of channels account for more than 90% of channel access.

We next focus on the short term distribution of channel popularity with respect to the two metrics. We examine this property at different time scales and at different points in time. Interestingly, we

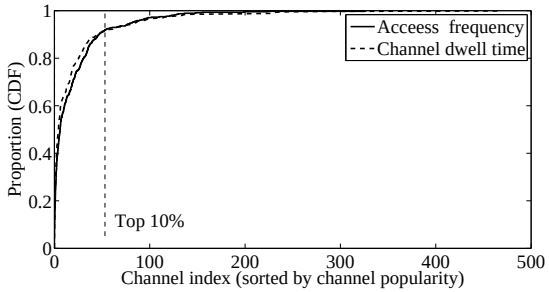


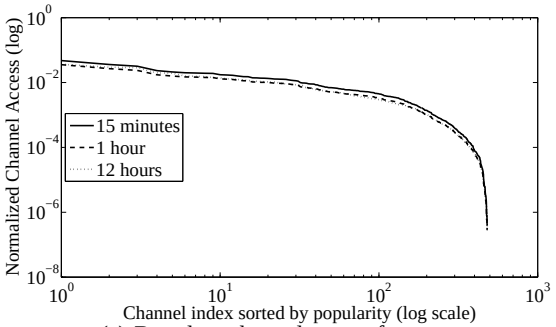
Figure 4: CDF of channel popularity.

find the channel popularity distribution is nearly *invariant* across a large range of measurement time scales or over different time periods. For example, in Figure 5, we show the average popularity probabilities (i.e., normalized channel access ratio and normalized channel dwell time ratio) measured during 15-minutes, 1-hour, and 12-hour periods starting from 0 AM. We sort the channels in a decreasing order of popularity and plot their rank in x-axis. We find that the three curves almost overlap on top of each other. In Figure 6, we also show the channel popularity probabilities of the same day using 4-hour aggregation granularity (0 AM to 4 AM, 8 AM to 12 PM, and 8 PM to 12 AM). Again, we find the curves very close to each other. We emphasize that the nearly invariant distribution function does not necessarily imply the channel popularity itself being invariant — the rank order of the channels is actually different from time to time and from one scale to another. We will turn to the temporal dynamics of channel popularity in Section 3.3.

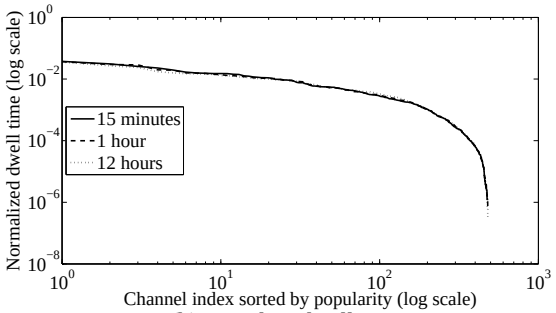
The log-log scale curves in Figure 5 and Figure 6 also suggest that channel popularity is highly skewed in all cases. To simplify computation, in the rest of the paper, we focus only on the top 150 channels which account for over 98% of the channel accesses. We acknowledge that modeling the tail part may be important for some applications. However, this simplification should have little impact on the analysis of overall time dynamics of channel popularity, which is the main focus of our study.

3.2 Correlation between Channel Accesses and Channel Dwell Time

We have observed in the previous subsection that channel popularity based on access frequency and based on dwell time produces a very similar result. This may be an indication for a strong correlation between these two popularity measures, which turns out to be true as illustrated in Figure 7, based on the entire period of trace. Figure 7(a) shows the scatter plot of the ranks of the channels in which the popularity rank according to channel access frequency is shown on the x-axis and the rank according to channel dwell time on the y-axis. Figure 7(b) shows the similar scatter plot of the actual probability value by the two metrics instead of the corresponding ranks. In both figures, we find that the points are spread well along the diagonal line — their Spearman rank correlation value equals to 0.98 and their Peterson correlation coefficient equals to 0.97 – demonstrating the strong correlation between the two popularity metrics. We believe that the relatively long delay during channel switches (described in Section 2) and the convenient TV guide and favorite-channel-list features are both contributing factors to this high correlation, as people are more likely to switch directly to the channel that they intend to watch. Another factor that may influence the observed high correlation comes from a limitation of our data source. As described in Section 2.1, channel switch requests that are rapidly followed by a subsequent request

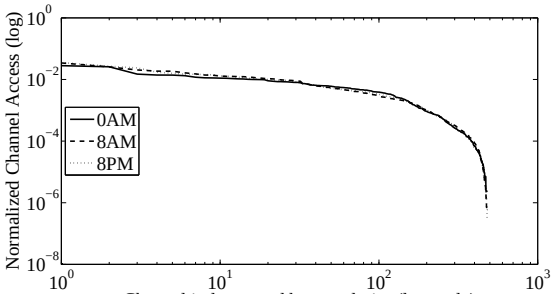


(a) Based on channel access frequency

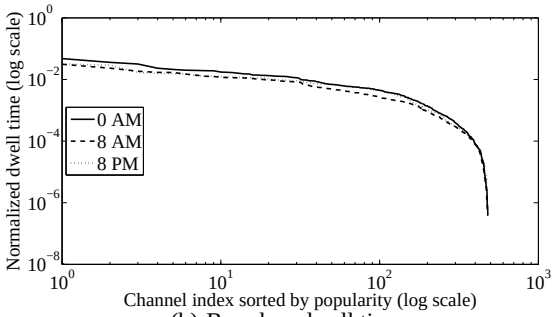


(b) Based on dwell time

Figure 5: Channel popularity distribution (varying time period).



(a) Based on channel access frequency



(b) Based on dwell time

Figure 6: Channel popularity distribution (varying start time).

are not recorded in our logs.

In the rest of this paper, we use the channel access frequency as the metric for channel popularity when illustrating our findings.

3.3 Temporal Dynamics of Channel Popularity

We now turn our attention to the time dynamics of popularity for individual channels. We start by looking at the time series of the channel popularity. (We refer to channel popularity measured by channel access frequency simply as channel popularity). Figures 8 and 9 show the time series of 9 days for a kids channel K and a news channel N respectively. In contrast to the time-invariant behavior reported in Section 3.1, both time series exhibit strong fluctuations over time. We next follow classic time series analysis processes to analyze these channel popularity series. In particular, we examine their stationarity, their first-order and second-order statistics, and their autocorrelation structure.

To test the stationarity of the channel popularity series, we apply the nonparametric *runs test* [2]. Given a time series $X(t)$, the runs test works as follows: (i) divide the series into equal-length time intervals and compute a mean value $\bar{X}_i$ for each bin, (ii) compute the median value of $\bar{X}_i$ over all bins and mark the ones below the median as "$-$" and the rest as "$+$", (iii) consider a consecutive sequence of "$+$" or a consecutive sequence of "$-$" as a *run* and count the total number of runs, and (iv) compare the number of runs against known run-count-distribution for stationary random data.

At the 95-th percentile confidence interval, we find that 92% of the channels pass the stationarity test when aggregated at 15-minute intervals. A small number of channels that fail the runs test exhibit non-trivial daily pattern, to which we will offer explanation in Section 5.

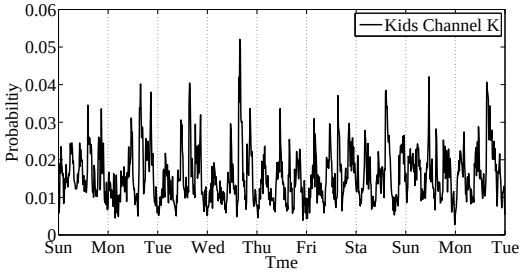


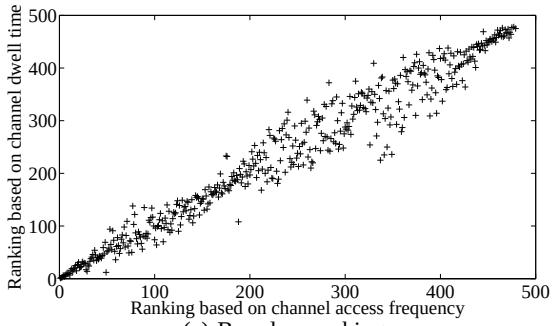
Figure 8: The dynamics of channel popularity of kids channel K, 1 point every 15 minutes

We also calculate the coefficient of variation (CoV) for the channel popularity series. Figure 10 shows the distribution of CoV's of the channels. Despite the wide difference in their mean value (shown in Figure 4), we find that the CoV's of channel popularity series are narrowly centered around 0.6. For example, the CoV for the series of the kids channel K (in Figure 8) is 0.57 and that for the news channel N (in Figure 9) is 0.68. We will see how the empirical CoV help in our model in Section 4.

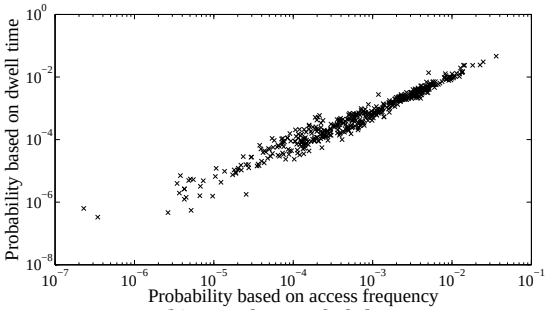
We further study the autocorrelation structure of the channel popularity series, defined by their autocorrelation function (ACF):

$$R(\tau) = \frac{\mathrm{E}[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2}$$

Figure 11 shows the ACF for the channel popularity series of the kids channel K and news channels N (other channels are similar). The lag ranges from 15 minutes to 8 days. The roughly straight



(a) Based on ranking



(b) Based on probability

Figure 7: The correlation between channel access frequency and dwell time.

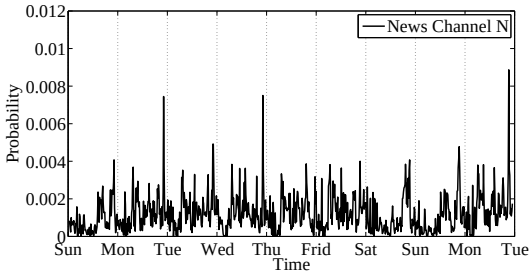


Figure 9: The dynamics of channel popularity of news channel N, 1 point every 15 minutes

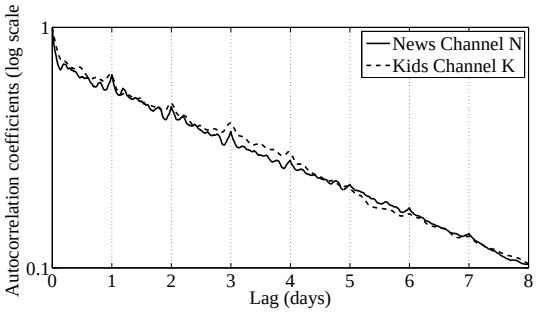


Figure 11: The autocorrelation function of both news channel N and kids channel K

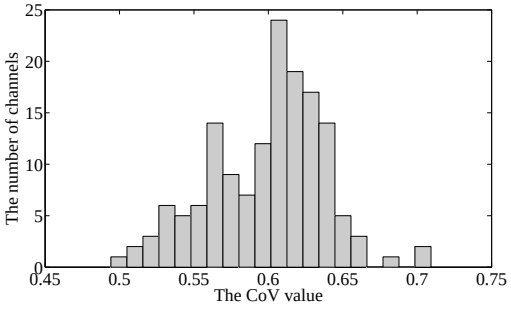


Figure 10: The distribution of CoV

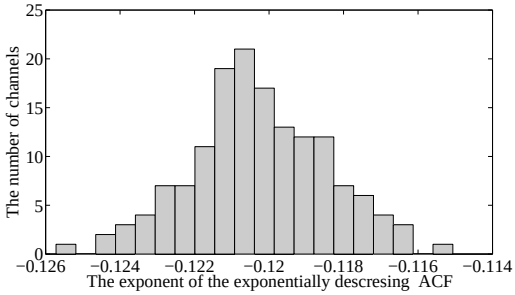


Figure 12: The distribution of the slope in ACF

lines indicate that the autocorrelation decays exponentially (note the log scale in y-axis) as the time lag increases, a typical behavior often observed in auto-regression processes. We further observe that the slope of the decreasing curves, which is the exponent of the exponentially decreasing ACF, are close among all channels. Using least square fitting, we obtain the best estimate of the exponent for each channel and plot their distribution in Figure 12. We find that their values concentrate at around -0.12.

In Figure 11, besides the sharp decreasing trend, we also observe small increases at the lags around day boundary (1 day, 2 days, etc.). This implies that there indeed exist some diurnal patterns, although minor, in the channel popularity. We address this issue in Section 5.

3.4 Multi-scale Property of Channel Popularity Similarity

We have examined the autocorrelation structure of the popularity measure of each individual channel. To quantify the similarity (or dissimilarity) of the channel popularity collectively among all channels, we adopt the metric named *cosine similarity*. Cosine similarity measures the similarity between two vectors by finding the cosine value of the angle between them. For a pair of vectors $\mathbf{A}$ and $\mathbf{B}$, the cosine similarity is given by:

$$similarity(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \, ||\mathbf{B}||}$$

Its value ranges from -1 to 1, with value closer to 1 indicating higher similarity between $\mathbf{A}$ and $\mathbf{B}$. Cosine similarity has been widely used in high-dimensional data analysis such as applications in text mining [19].

In the context of IPTV, we expect the channel popularity to be relatively stable over time. This is indeed true—the average cosine similarity between adjacent 15-minute time bins is around 0.97, indicating the distribution of the channel popularity is quite stable

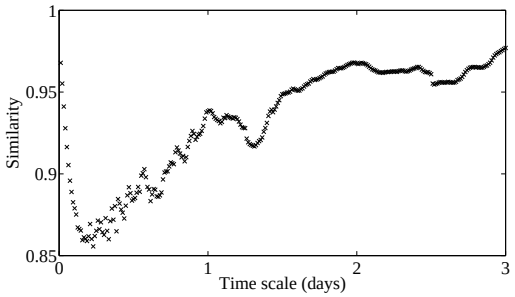


Figure 13: The average cosine similarity for different aggregation time scales

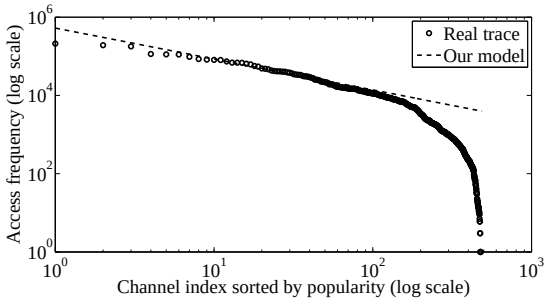


Figure 14: Channel popularity distribution.

in a short time frame. We further investigate whether the similarity becomes more pronounced with different time scales and perform the following multi-scale analysis.

We discretize our data traces by fixed-interval time bins, with interval lengths ranging from 15 minutes to 3 days. At each interval, we calculate the channel access probability of different channels for each time bin. Then, for each pair of adjacent time bins, we compute the cosine similarity of channel popularity vectors. Based on these values, we calculate the average for each aggregation interval. Figure 13 shows the result where x-axis is the aggregation time scale (interval length) and y-axis is the average similarity.

In Figure 13, we observe that the curve forms a *V*-shape as we increase the aggregation level. Specifically, the similarity value first decreases as the aggregation times increases, reaching its minimum at around 3-4 hour aggregation scale. After that, we observe an increasing trend as we increase the aggregation time scale. This is because when the time scale is short, the similarity/dissimilarity of the channel popularity is determined by the TV program (shows) of the time. On the other hand, when the time scale is long, the similarity/dissimilarity is determined by the overall type of TV program on the channels. Both the viewer base of individual TV shows and the long term user affinity to the type of program are relatively more stable, which makes the time scale in between the weakest in term of channel popularity stability. We can also gain some intuition from the perspective of process analysis. Specifically, Figure 11 shows the exponentially decaying autocorrelation function of channel popularity (note the log-scale on y-axis), causing the fast decreasing stability in short time scales. As the aggregation level becomes sufficiently large, the short term disturbances are smoothed out, converging to long term average, and hence improving the stability. We next present our model and demonstrate that our proposed model can closely match this behavior.

4. MODELING CHANNEL POPULARITY

We now present our model for channel popularity. We will use Zipf-like model to capture the long term channel popularity distribution among different channels and mean reversion model to capture the stochastic process of popularity fluctuation for individual channels.

4.1 Zipf-like model

The Zipf-like distribution has been proved successful in capturing the skewness in content popularity such as Web [1] and VOD [17]. In the Zipf-like distribution, an object of the rank i has the access probability of $C/i^{1-\alpha}$, where C is a normalization constant and α is the distribution skew parameter. In Section 3, we have observed that the channel popularity is also highly skewed. We naturally model it using Zipf-like distribution. Figure 14 shows the access frequencies of all channels in the order of decreasing popularity from the real trace and the fitted Zipf-like distribution (the dashed line with $\alpha = 0.55$). We observe a very good match up to around 150 channels, which account for over 98% of the channel-switches (see Figure 4).

4.2 Mean reversion model

Modeling the temporal dynamics of channel popularity is more challenging. Based on our analysis in Section 3, we choose a class of stocastic models, namely *mean reversion model* for this purpose. Mean reversion model has been widely used in financial data analysis. The basic idea is that the price of a stock or a commodity may fluctuate but will revert to its long-term equilibrium level. Ornstein-Uhlenbeck (OU) process $\{X_t : t > 0\}$ is the most widely used mean reverting stochastic process in financial modeling [18]. It is stationary, Gaussian, Markovian, and continuous in probability. The Ornstein-Uhlenbeck process is characterized by the following linear stochastic differential equation (SDE) [9]:

$$\mathrm{d}X_t = \lambda(\mu - X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \tag{1}$$

where $\lambda > 0$ is the mean reversion rate, μ the long-term mean, and σ the volatility. W_t denotes a Wiener process (also known as Brownian motion), which is characterized by: (i) $W_0 = 0$, (ii) W_t is almost surely (i.e., with probability one) continuous, and (iii) W_t has independent increments with distribution $W_t - W_s \sim \mathcal{N}(0, t-s)$ for $0 \leq s < t$.

To understand the OU process, we can view the RHS of Eq (1) as summation of a deterministic term (the first term in RHS) and a stochastic term (the second term in RHS). When $X_t > \mu$, the deterministic term $\lambda(\mu - X_t)$ is negative, resulting in pulling back down toward the equilibrium level (i.e., μ); if $X_t < \mu$, the deterministic term is positive, pushing X_t back up to the equilibrium level. As a result, every time the stochastic term makes X_t deviate from the equilibrium, the deterministic term will act in such a way that X_t will head back to the equilibrium μ.

For an OU process, we have the moments:

$$\mathrm{E}(X) = \mu \tag{2}$$

$$Cov(X_s, X_t) = \frac{\sigma^2}{2\lambda} e^{-\lambda|s-t|} \tag{3}$$

This implies that the autocorrelation function of an OU process decays exponentially as the lag $|s-t|$ increases, which would match well with the empirical ACF of channel popularity series in Figure 11.

We now determine the model parameters from the analysis result in the previous section. It is straightforward to see that the long term equilibrium μ can be derived from Eq (2), which we further model by the Zipf-like distribution. From Eq (3), we find that the

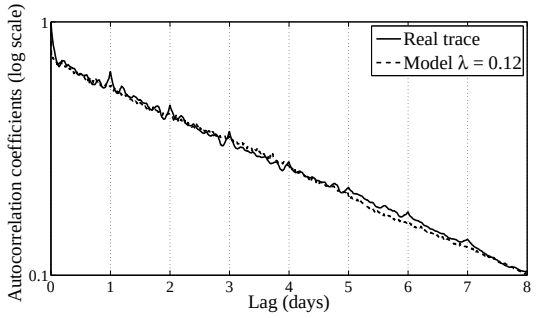


Figure 15: Fitting autocorrelation function with $\lambda = 0.12$

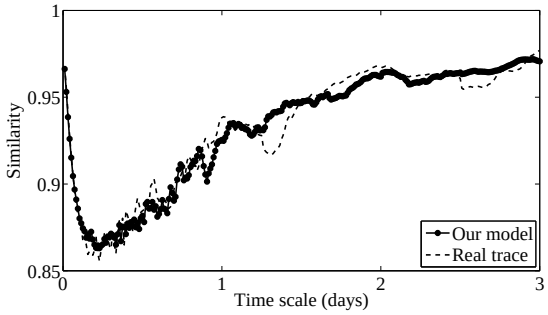


Figure 16: Cosine similarity using a simulated trace based on the mean reversion process.

autocorrelation decreases with lag at the rate $e^{-\lambda}$. Use the value extracted from in Figure 12, we set $\lambda = 0.12$. Finally, to determine σ, we use the coefficient of variation (CoV). Using Eq (3), we can derive σ as follows:

$$\sigma = \mu \times \sqrt{2\lambda} \times \text{CoV}$$

Using fixed time steps of 1, we can obtain a discrete version of the OU process and derive a first-order autoregressive sequence of X_t:

$$X_{i+1} = X_i e^{-\lambda} + \mu(1 - e^{-\lambda}) + \sigma\sqrt{\frac{1 - e^{-2\lambda}}{2\lambda}}\mathcal{N}_{0,1} \qquad (4)$$

where $\mathcal{N}_{0,1}$ is a standard Gaussian random variable. This can be used to drive simulation of IPTV channel popularity.

We validate our model against measurement data. Figure 15 shows the ACF of the news channel N compared against the simulation from the model with $\lambda = 0.12$. We observe a remarkable match, with the exception of the surges at lags that are multiple of one day. We further examine the multi-scale property as shown in Figure 13. Figure 16 shows that our model faithfully reproduce the V-shape behavior in the cosine similarity of channel popularity vectors.

4.3 Forecasting channel popularity

We have shown that with properly chosen parameters, the OU process can nicely capture various properties on the channel popularity. We now explore whether we can use it as the underlying process to perform forecasting. More specifically, given the historical states from X_0 to X_i for a channel, how accurately can we predict X_{i+1}?

This can be viewed as a linear regression problem due to the AR(1) model of the sequence of X_i in Eq (4). To facilitate the regression analysis, we rewrite Eq (4) as:

$$X_{i+1} = aX_i + b + \epsilon \qquad (5)$$

The first objective of linear regression analysis is to best-fit the data by estimating the parameters of the model. Of the different criteria that can be used to define what constitutes a best fit, the least squares criterion is a very powerful one. Using the least squares criterion, we obtain the model parameters as follows.

$$a = \frac{nX_{xy} - X_x X_y}{nX_{xx} - X_x^2}$$

$$b = \frac{X_y - aX_x}{n}$$

$$sd(\epsilon) = \sqrt{\frac{nX_{yy} - X_y^2 - a\,(nX_{xy} - X_x X_y)}{n(n-2)}}$$

where

$$X_x = \sum_{i=1}^{n} X_{i-1}, \qquad X_y = \sum_{i=1}^{n} X_i,$$

$$X_{xx} = \sum_{i=1}^{n} X_{i-1}^2, \qquad X_{xy} = \sum_{i=1}^{n} X_{i-1}X_i, \qquad X_{yy} = \sum_{i=1}^{n} X_i^2$$

We take the trace of a news channel to evaluate the performance of our forecasting model. We find that a small resulting mean squared error (MSE) ($= 8 \times 10^{-8}$) is obtained compared to its mean value 0.0014 and variance 9.3×10^{-7}. This means our forecasting model predicts the dynamics of channel popularity reasonably well. We have performed the prediction on various channels and observed the similar results.

5. MULTI-CLASS POPULARITY MODELING

So far we have presented our analysis on the dynamics of channel popularity and developed a mean reversion process to model them. Although the model in the previous section works reasonably well, it is incapable of capturing some diurnal patterns in the dynamics such as the small increases around daily boundary in ACF (see Figure 15). In this section, to enhance our model to capture these diurnal patterns, we investigate whether we can identify subgroups of STBs that have distinct channel preference, compared to the overall pattern and dynamics of channel popularity. We first explore various features that we can use to group STBs (Section 5.1), and then analyze properties of different groupings to identify a desirable grouping that we can use for our multi-class modeling (Section 5.2 and 5.3). This grouping method actually provides an interesting insight behind the dynamics of channel popularity (Section 5.4), and we employ this finding to develop a multi-class popularity model that better captures the channel popularity dynamics (Section 5.5).

5.1 Grouping STBs

Given the data set we have, we have a number of different ways to group the appearing STBs. Here we choose the following attributes which can best characterize a STB for grouping:

- **TV watching time:** For each STB, we consider various aspects of TV watching time, such as daily average, hourly average, and average nightly watching time.

- **Channel change frequency:** We consider the daily average and hourly average of channel changes to group STBs.

- **Dwell time per channel change:** For each channel change, we determine how long a STB stays on the channel. This dwell time can be reported long when a user does not watch the channel, but leaves the STB on. To minimize such effect in our analysis, we investigate both the median value and the average value of dwell time per channel change.
- **Location:** We use the network location of a STB to group STBs.

We use the first 15 days of the logs to calculate the attributes for each STB. In other words, we use the data in these days as the training set. As described later in this section, we use the remaining data to evaluate the properties of grouping. We next describe different grouping strategies that we use to identify various groupings of STBs, which can be classified into two categories. One is threshold-based grouping. The other is clustering algorithm-based group.

5.1.1 Threshold-based

In this grouping, we select a grouping attribute and a set of corresponding thresholds to group STBs. These thresholds are chosen by the common sense of viewers instead of some specific computer algorithm.

- **Daily watching time (**WT-D**):** We consider the daily average TV watching time for each STB. Specifically, we call a STB a *heavy-watcher* if the STB spends more than 12 hours on average, and a *light-watcher* if it spends less than 1 hour. We call the remaining STBs *medium-watchers*. In our data, 28% of STBs are heavy-watchers, and 36% of them are light-watchers. In the rest of this section, we call this grouping outcome WT-D.
- **Daytime vs. Nighttime (**DN-D**):** We define a STB as a *daytime-watcher* if the average TV watching time during the day (from 6am to 6pm) is more than twice the time during the night (from 6pm to 6am). We define a *nighttime-watcher* similarly. We call the remaining STBs *all-time-watchers*. We observe 31% of STBs are daytime-watchers, and 39% of them are nighttime-watchers.
- **Daily channel change count (**CHG-D**):** We use the average channel change count per day. We define the STBs that switch channels more than 200 times on average as *frequent-switchers* (24% of STBs), and the STBs that switch the channel less than 10 times as *infrequent-switchers* (12% of all). We call the remaining 64% of STBs *moderate-switchers*.
- **Median dwell time (**DWL**):** For each STB, we use the log of 15 days and find the median value for the dwell time per channel change. Then we use 10, 20, and 30 minutes as thresholds to divide them into four groups.
- **Location (**LC**):** We use the metropolitan area as the granularity of grouping STBs based on the location.

5.1.2 Clustering Algorithm-based

In this category, we employ unsupervised clustering algorithms to group STBs. While we have explored multiple clustering algorithms, we focus on the results based on the well-known *K-mean* algorithm [13], which is effective for large data sets. In this algorithm, we need to provide the number groups K as input parameter. While there are several ways to find the optimal K, we use the intra-cluster dissimilarity W_K as the measure:

$$W_K = \sum_{k=1}^{K} \sum_{C(i)=k} ||x_i - \hat{x}_k||^2,$$

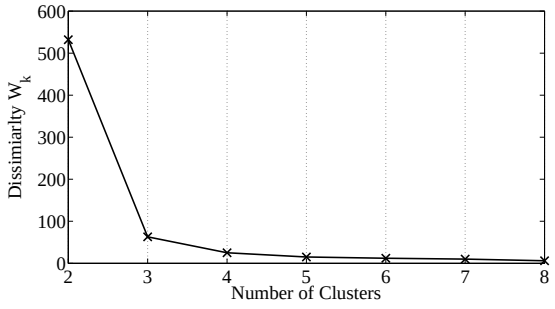


Figure 17: Dissimilarity vs. K when we use TV watching time as grouping feature.

where x_i is the data item, and $\hat{x}_k$ is the center of items in k-th cluster. We vary $K \in \{1, 2, \ldots, K_{max}\}$ and obtain a separate grouping result and the corresponding W_K for each K. Then, we consider the trade-off between dissimilarity and the number of clusters when we choose the optimal K. We present a concrete example below when we use the following set of attributes as input feature to the clustering algorithm. Different from the threshold-based category, they are all feature vectors.

- **Hourly TV watching time (**WT-H**):** We assign a 24-element tuple to each STB, where each value corresponds to the average TV watching time per hour in a day. Then we perform the K-mean algorithm to cluster the STBs. In a sense, WT-H simultaneously considers the two features used in WT-D and DN-D. In Figure 17, we plot W_K as a function of K. We can observe that as small a value as $K = 3$ provides a good grouping result. These three clusters cover 60%, 27%, and 13% of STBs, respectively.
- **Hourly channel change (**CHG-H**):** Similar to WT-H, we collect the number of channel changes for each hour in a day and assign a 24-element vector to each STB. In this grouping, $K = 4$ leads to the optimal grouping result, where the clusters have 47%, 25%, 21%, and 7% of STBs, respectively.
- **Hourly dwell time (**DWL-H**):** For each hour, we calculate the average dwell time per channel change. Then, we assign 24-element vector to each STB, to obtain four groups with 37%, 31%, 24%, and 8% of STBs from the K-mean algorithm, respectively.
- **Hourly median dwell time (**MDWL-H**):** Unlike DWL-H, we use 1-hour intervals and calculate the median dwell time value for each 1-hour bin and input them into the K-mean algorithm. From this grouping, we obtain four groups with 41%, 20%, 10%, and 29% of STBs, respectively.
- **Channel preference (**PREF**):** For each STB, we calculate the access probability to each of top 150 channels (which covers 98% of channel popularity as shown in Figure 4). Then, using Table 1[3], we classify these channels based on their program contents and obtain aggregate access probabilities for 8 types, which we use as the grouping attribute. We use $K = 8$ after the number of types.

[3]In this classification, both educational channels and documentary channels belong to "science." The category "others" includes channels that offer diverse programs (e.g., news, TV series, shows, etc.) as well as less known channels that are not easy to classify.

Table 1: Classification of top 150 IPTV channels

Type	Examples	# channels
News	CNN, NBC News	13
Kids	Disney, Cartoon Network	15
Sports	ESPN, Star games, NBA TV	20
Movies	HBO, Cinemax	15
Science	Discovery channel, Animal planet	20
Music	MCM, MTV	21
Foreign	TF1, BFM, Al jazeera, CCTV	18
Others	TBN, EWTN	28

In the rest of this section, we investigate whether we observe any correlation among the features and corresponding groupings and we can explain underlying channel popularity dynamics based on the identified sub-groups.

5.2 Measuring Difference in Channel Preferences of STB Groups

In this part, we examine whether STBs in different groups exhibit different channel preferences. We use *mutual information* in measuring the *difference* of channel preferences of STBs belonging to different groups.

In probability theory and information theory, the mutual information of two random variables quantitatively measures their mutual dependence. Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right) \qquad (6)$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p_1(x)$ and $p_2(y)$ are the marginal probability distribution functions of X and Y respectively. The smaller the mutual information value is, the larger the difference between X and Y is.

We conduct *significance testing* to determine whether the channel preference of a given STB group G is significantly different from that of all STBs S. For this, we first compute the mutual information I_G between channel preference vector of G and that of S using Equation (6). Here, X and Y are two variables describing channel preferences. In particular, $p_1(x = X)$ is the probability to choose a type X channel for group G. Similarly, $p_2(y = Y)$ is a probability to choose a type Y channel for S. $p(x = X, y = Y)$ is the probability of choosing type X channel in G and choosing type Y channel in S.

Then we randomly select a subset S_i of S, which has the same size as group G. Similarly, we compute the mutual information I_{S_i}. After taking a large number of random selections of S_i, we can get the empirical distribution of I_{S_i}. According to the Central Limit Theorem, $\bar{I}_{S_i}$ is approximately normally distributed with mean $\hat{\mu}$ and deviation $\hat{\delta}$. Here, our null hypothesis H_0 is: group G is not significantly different from S in terms of channel preferences. For the sampled distribution, we compute the *p-value* $\Pr[\bar{X} \leq I_G | (\hat{\mu}, \hat{\delta})]$. If the p-value is very small, e.g., < 0.005, we shall reject H_0. Using this method, we can verify if a group G has a significant difference in the channel preference compared with all STBs S. We can also apply the same method on a given type of channels to determine if G has a significant difference in preference for that type of channels.

Table 2 shows channel preferences of all STBs as well as STB groups based on PREF. There are eight STB groups, each of which corresponds to one type of channels. The size of STB groups varies from 45% of all STBs to 2% of all STBs. The STB group preferring news channels is the largest, and STB groups preferring music and foreign channels are the smallest. We highlight the identified significant difference in channel preference in bold. Compared to all STBs, we observe that each group clearly exhibits distinct preference for the corresponding type of channels. For example, group1 shows significant preference for news channels. These results indicate the potential benefit of modeling different groups seperately, which we focus on later in Section 5.5.

5.3 Identifying Best Grouping Methods

Now we propose a generic method for selecting the best grouping methods. In our case, a "good" grouping should achieve the following two goals. First, the method should yield STB groups that well represent the channel preferences. Second, the resulting STB groups should be stable over time.

To identify grouping methods that yield good representation of channel preferences of STBs, we compute mutual information between STB groups based on PREF and those based on each of other grouping methods (denoted as M) using Equation (6). Here, we consider each STB group as a random variable. $p_1(x = X)$ is the probability that a STB belongs to group X according to PREF. $p_2(y = Y)$ is the probability that a STB belongs to group Y according to a given grouping method M. The joint distribution $p(x = X, y = Y)$ is the probability that a STB belongs to group X based on PREF and belongs to group Y based on M.

It is likely that different grouping methods yield different number of groupings. For example, the location based grouping will yield over 150 clusters while other grouping methods usually yield a handful of groups. In such a case, the mutual information $I(X;Y)$ defined in Equation (6) can be misleading. In order to perform a fair comparison on different grouping methods, we adopt a normalized metric called *symmetric uncertainty*, which is defined as:

$$U(X,Y) = 2 \frac{I(X;Y)}{H(X) + H(Y)} \qquad (7)$$

where $I(X;Y)$ is the mutual information defined in Equation (6) and H is the entropy:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i), \qquad (8)$$

When X and Y are independent, $U(X,Y) = 0$. When X is a function of Y, $U(X,Y) = 1$.

Table 3 shows the symmetric uncertainty between the channel preferences (i.e., PREF) and different grouping methods described in Section 5.1. We find that clustering algorithm-based on hourly median dwell time (MDWL-H) and on hourly TV watching time (WT-H) yield the highest and lowest symmetric uncertainty values (0.513 and 0.123) among all the grouping methods. Intuitively, this can be explained as follows. Users who watch TV at the same time during a day does not necessarily watch the same set of channels (i.e., they do not necessarily have a clear mutual interest in channels). However, users who switch channels at the same time during a day may have a strong preference for the type of channels they watch. This is because most of the channel change behaviors are impacted by the start/end times and commercial breaks of the TV program. The symmetric uncertainty values for threshold-based grouping methods range from 0.179 to 0.314, with the grouping based on the daily watching time WT-D having the highest value and grouping based on the location LC having the lowest value.

We also prefer a grouping method that yields STB groups that are stable over time. We perform a stability test on our grouping results in the following way. We use the percentage of STBs that stay in the same group over a certain time period (e.g., 15 days) as the metrics to measure the stability of STB groups. In our analysis, we divided our data traces into two parts, where each part lasts

Table 2: Channel preferences of STB groups based on PREF.

	News	Kids	Sports	Movies	Science	Music	Foreign	Others	Group size (%)
All STBs (%)	52.3	14.4	5.2	3.1	1.8	0.3	0.4	22.4	100
Group1 (%)	**67.8**	9.7	3.5	2.1	1.2	0.2	0.3	15.2	45
Group2 (%)	49.5	**19.0**	4.9	2.9	1.7	0.3	0.4	21.2	12
Group3 (%)	50.2	13.8	**9.0**	3.0	1.7	0.3	0.4	21.5	5
Group4 (%)	50.7	14.0	5.1	**6.0**	1.8	0.3	0.4	21.8	6
Group5 (%)	50.6	13.9	5.0	3.9	**5.1**	0.3	0.4	21.7	3
Group6 (%)	51.0	14.8	5.1	3.0	1.8	**3.0**	0.4	21.8	2
Group7 (%)	51.6	14.2	5.1	3.0	1.9	0.3	**1.8**	22.1	2
Group8 (%)	48.9	13.5	4.9	2.9	1.7	0.3	0.4	**27.5**	27

Table 3: Symmetric uncertainty between PREF and different grouping methods

	WT-D	DN-D	CHG-D	DWL	LC	WT-H	CHG-H	DWL-H	MDWL-H
PREF	0.314	0.305	0.254	0.309	0.179	0.123	0.206	0.430	0.513

Table 4: Stability of different grouping methods

PREF	WT-D	DN-D	CHG-D	DWL	LC	WT-H	CHG-H	DWL-H	MDWL-H
67.1%	83.5 %	79.4%	77.6%	74.3%	100%	70.4%	72.3%	66.5%	69.4 %

15 days. We compute STB groups on each 15-day trace separately and examine the stability of STB groups. Note that for clustering algorithm based grouping methods, because the group centers are determined non-deterministically, we group the second 15-day trace by using the same group centers as those that are identified in the first 15-day trace. Then, for each STB, we compute the distance between attribute vector obtained from the second 15-day trace and each group center identified in the first 15-day trace. The STB is assigned to the group of which the center is closest.

Table 4 shows the stability of different grouping methods. We have four key observations. First, the grouping based on channel preference PREF is not stable over time. This indicates that PREF may not be a good grouping method to be used in our model even though Table 2 shows PREF clearly represents distinct channel preferences in each STB group. Second, we find that all the grouping methods based on hourly features (i.e., WT-H, CHG-H, DWL-H, and MDWL-H) have low stability over time. Hence, they are not considered good grouping methods to be used in the model. Third, we observe that grouping based on location LC yields perfect stability of STB groups. This is expected because STBs location is less likely to change over time. However, since LC has a low symmetric certainty value as shown in Table 3, it is not considered a good choice either. Finally, we observe that the grouping based on daily TV watching time WT-D has the highest stability among all grouping methods other than LC. In addition, WT-D also has a relative high value in symmetric uncertainty as shown in Table 3 (it is the highest among the threshold based grouping methods). Thus, we identify WT-D to be the best grouping method based on our data trace.

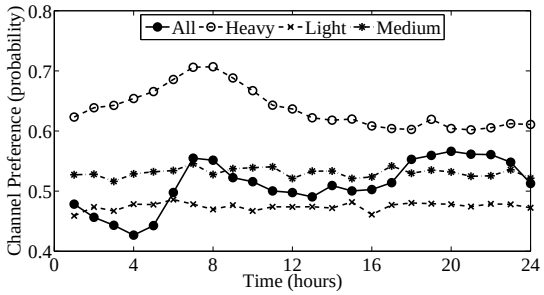


Figure 18: Time-of-day dynamics for news channels, comparing WT-D with all watchers

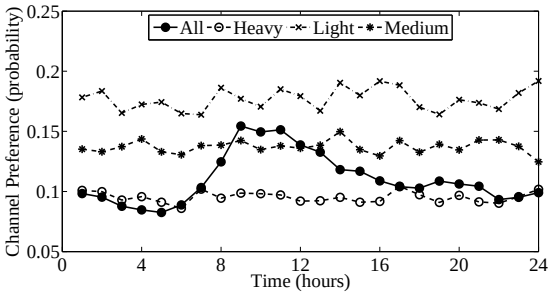


Figure 19: Time-of-day dynamics for kids channels, comparing WT-D with all watchers

5.4 Explaining Channel Popularity Dynamics

In Section 3, we have observed diurnal patterns in channel access popularity (Figures 8 and 9). In this subsection, we examine whether some of these groups exhibit different channel access preference. Based on the result in the previous subsection, we focus on the grouping result by WT-D, because it has the highest stability (except for LC) as well as a reasonably high symmetric uncertainty measure against PREF. Table 5 compares the channel preference of each group based on WT-D with that of all STBs. Based on our significance testing, we find that heavy-watchers group and light-watchers group have distinct preferences to news and kids channels, which are marked in bold. In this subsection, we focus on the preference for these two channel types. Although we do not present here, we also investigated other groupings and observed a similar result in many cases.

In Figures 18 and 19, we show the access probability of news channels and kids channels (as defined in Table 1), respectively. We display one line for each group in WT-D as well as an additional line for the all-STBs case (denoted by "all-watchers"). We observe

Table 5: Channel preferences of STB groups based on WT-D.

	News	Kids	Sports	Movies	Science	Music	Foreign	Others	Group size (%)
all watchers (%)	52.3	14.4	5.2	3.1	1.8	0.3	0.4	22.4	100
heavy-watchers (%)	**62.6**	9.7	4.9	2.3	2.0	0.4	0.3	19.6	28
light-watchers (%)	47.4	**17.5**	5.4	2.3	1.7	0.4	0.4	25.3	36
medium watchers (%)	53.3	13.9	4.7	3.0	1.9	0.3	0.3	22.5	36

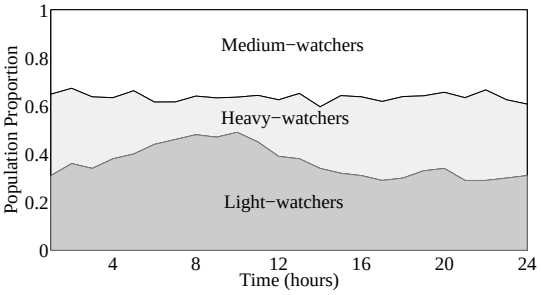


Figure 20: Population mix for each group based on WT-D

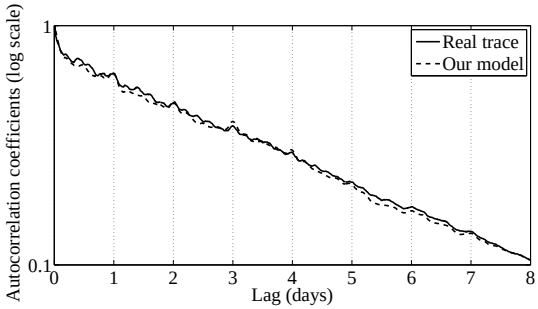


Figure 21: The autocorrelation function based on multi-class model.

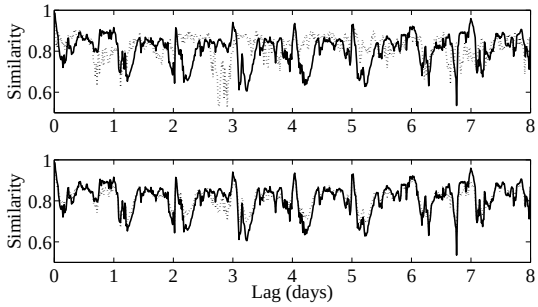


Figure 22: The cosine similarity function when varying lag. The solid line represents the real trace, the dash line represents the model. Top: single-class, bottom: multi-class.

that while the line for all-watchers shows a clear diurnal access pattern (e.g., with peaks around 7am and 8pm for news channels), the channel preference within a group is more stable throughout the day, except for heavy-watchers' increased preference during the morning time. Specifically, in Figure 18, at least 60% of channel changes by heavy-watchers are for news channels throughout the day, which is significantly higher than the overall average 52.3%. While the average preference of medium-watchers for news channels (53.3%) is similar to the daily average, this group also exhibits a more stable access pattern, compared to all-watchers. In Figure 19, while the group-level access probabilities for kids channels fluctuate more than for news channels, the values are still more stable than that of all-watchers. These results illustrate that we can identify sub-groups that have distinct channel preference, and although the channel popularity within the groups may vary over time, some groups often have fairly constant channel preference for some channels.

In Figures 18 and 19, we observed that group-level channel preference stays reasonably stable all day, but the overall channel popularity shows a diurnal pattern. How can we explain two seemingly conflicting results? In Figure 20, we show the proportion of active STBs for each group by WT-D. We determine that a STB is active if there is a channel change during the 1-hour window. This figure shows that medium-watchers constitutes around 35% of active STBs throughout the day, heavy-watchers between 15% and 35%, and light-watchers between 30% and 50%. We observe that there is strong correlation between channel popularity change and population mix change. For instance, in Figure 19, the overall channel access probability for kids channels peaks between 8am and noon. This coincides with the population gain by light-watchers (Figure 20), which has significantly higher preference for kids channels. In Figure 18, while the overall channel popularity peaks at the morning time (7am) due to the change in preference by heavy-watchers, the increase of heavy-watchers in the population mix obviously explains the other peak at later time (around 8pm).

In sum, our results show that some sub-groups have different channel preference, and their population mix change has a strong correlation with overall popularity change. In the rest of this section, we further investigate these findings and demonstrate that we can better model channel popularity dynamics by employing this grouping methodology.

5.5 Modeling and Simulation

Now we can use our classification results to further improve the modeling results shown in Section 4. We use the grouping result from the feature WT-D since it has the highest stability over time and a reasonably large symmetric uncertainty value. Assume that all channels still follow the mean reversion model in each group. But the parameters need to be revisited. Let X_{ij}^t, μ_{ij}, λ_{ij} and σ_{ij} denote the popularity measure, the long-term mean, the mean reversion rate and the volatility of the group j on the channel i, respectively. The estimation procedure described in Section 4 can be easily adapted to derived the parameters for every $(channel, group)$ combination.

Then to simulate the temporal popularity dynamics for a channel i, we mix all $(i, j), j = 1, 2, ...,$ using the empirical population proportion for each group (see Figure 20) as the mixture weight. In other words,

$$X_i^t = \sum_j W_j^t \times X_{ij}^t$$

where X_i^t denotes the popularity of channel i at time t and W_j^t denotes the proportion of STBs in group j at time t.

To evaluate the above multi-class model, we use the model to simulate the process of the popularity dynamics and compare it to

the real trace in two aspects. First, in Section 4 we have shown that the mean reversion model on a single class cannot model the daily bumps on autocorrelation curve well (see Figure 15). Here, we can do much better as shown in Figure 21. It is not hard to observe that our multi-class model captures the most of bumps at the daily boundary. As the result, the MSE of the model is equal to 1.6×10^{-5} which is more than one order of magnitude smaller than that ($= 2.4 \times 10^{-4}$) of the single-class model in Section 4.

Second, we compute cosine similarity (defined in Section 3) on the trace generated by our models. Again, we compare the cosine similarity of both single- and multi-class models with that from the real trace. In Figure 22, given a fixed lag, we compute the cosine similarity between the channel popularity vectors of two adjacent 15 minute-time-bins and take average on the length of 9 days. We repeat this by gradually varying the length of lag. This resulting curve from the real trace reflects the degree of similarity of channel popularity across the time domain. The top and bottom subfigures in Figure 22 compare the single-class model and multi-class model to the real trace in terms of cosine similarity, respectively. It is clear that the multi-class model can capture the high daily similarity, but single class model fails to do so. As the result, the MSE of multi-class model is 10^{-3} which is around one order of magnitude smaller than that ($= 9 \times 10^{-3}$) from single-class model. In summary, taking advantage of a good grouping feature with high stability and symmetric uncertainty scores, our multi-class model can generate a more accurate temporal dynamics process to simulate the real scenario than the previous single-class model.

6. RELATED WORK

The channel popularity or content/media popularity, in general, has been widely studied in different applications. Costa et al. [8] analyzed user activities and media distribution in media streaming applications. Cherkasova et al. [5], Chesire et al. [6], and Tang et al. [17] modeled workload of media streaming service. Yu et al. [21] studied the user activities to access a Video-on-Demand (VoD) system. Cha et al. [3] explored how users access videos in the YouTube system. Guo et al. [10] compared access patterns of different types of media content on the Internet including Web, P2P, VoD, and live streaming. However, the findings in these studies may not be applicable to IPTV systems as the user behavior can be inherently different from those in other applications.

More recently, there are a number of studies on IPTV system. Cha et al. [4] report various findings about user watching behavior by analyzing control messages in an IPTV system. While some of our findings are consistent with those reported in their study, we focus developing a multi-class population model of channel popularity based on key observations in our analysis. Smith [16] models bandwidth demand to support both multicast and unicast for fast channel change, where channel switching is modeled as a renewal process. However, the work does not consider the temporal dynamics within a day. Hei et al. [11] and Silverston et al. [15] report their measurement studies on P2P-based IPTV systems, while our work focuses on analyzing and modeling a large commercial IPTV system.

7. CONCLUSIONS

In this paper, we analyze and model channel popularity based on user channel access data in a nation-wide commercial IPTV system. We find that the channel popularity is highly skewed and can be well captured by a Zipf-like distribution. We also observe a fair amount of channel access popularity change during a short time window, although we find that channel popularity during moderately long time windows stays relatively stable. We demonstrate that we can model such popularity dynamics using a mean reversion process. Further, we develop a method for identifying groups of users which show intrinsic difference in their channel preference. We demonstrate that we can combine this grouping and the change of population mix to obtain a multi-class population model, which enables us to capture diurnal patterns in channel popularity dynamics. Although the focus in this paper is on analyzing and modeling channel popularity in an IPTV system, our methodology can be applicable to other systems, which we plan to investigate in our future work.

8. REFERENCES

[1] P. Barford and M. Crovella. Generating representative web workloads for network and server performance evaluation. In *SIGMETRICS*, pages 151–160, 1998.

[2] J. Bradley. *Distribution-free statistical tests*. Prentice-Hall., 1968.

[3] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proceedings of ACM IMC*, 2007.

[4] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatrianin. Watching Television Over an IP Network. In *Proceedings of ACM IMC*, 2008.

[5] L. Cherkasova and M. Gupta. Characterizing locality, evolution, and life span of accesses in enterprise media server workloads. In *NOSSDAV*, 2002.

[6] M. Chesire, A. Wolman, G. M. Voelker, and H. M. Levy. Measurement and analysis of a streaming media workload. In *USITS*, pages 1–12, 2001.

[7] J. Chu, K. Labonte, and B. Levine. Availability and locality measurements of peer-to-peer file systems. In *Proceedings of ITCom: Scalability and Traffic Control in IP Networks*, 2002.

[8] C. P. Costa, I. S. Cunha, A. B. Vieira, C. V. Ramos, M. M. Rocha, J. M. Almeida, and B. A. Ribeiro-Neto. Analyzing client interactivity in streaming media. In *WWW*, 2004.

[9] J. L. Doob. The Brownian movement and stochastic equations. *Annals of Math*, 40(1):351–369, 1942.

[10] L. Guo, E. Tan, S. Chen, Z. Xiao, and X. Zhang. The stretched exponential distribution of internet media access patterns. In *PODC*, pages 283–294, 2008.

[11] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross. A measurement study of a large-scale p2p iptv system. *IEEE Transactions on Multimedia*, 9(8):1672–1687, 2007.

[12] Y. Huang, T. Z. J. Fu, D.-M. Chiu, J. C. S. Lui, and C. Huang. Challenges, Design and Analysis of a Large-scale P2P-VoD System. In *Proc. ACM SIGCOMM*, 2008.

[13] J. B. MacQueen. Some methods for classification and analysis of multivariate observations,. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[14] J. Nielsen. Zipf curves and website popularity, www.useit.com/alertbox/zipf.html, 1997.

[15] T. Silverston, O. Fourmaux, K. Salamatian, and K. Cho. Measuring p2p iptv traffic on both sides of the world. In *CoNEXT*, page 39, 2007.

[16] D. E. Smith. IPTV Bandwidth Demand: Multicast and Channel Surfing. In *INFOCOM*, pages 2546–2550, 2007.

[17] W. Tang, Y. Fu, L. Cherkasova, and A. Vahdat. Medisyn: a synthetic streaming media service workload generator. In *NOSSDAV '03*, pages 12–21, 2003.

[18] G. Uhlenbeck and L. Ornstein. On the Theory of Brownian Motion. *Physical Review*, September 1930.

[19] Y. Yang. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In *SIGIR 94*, pages 13–22, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[20] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding user behavior in large-scale video-on-demand systems. *SIGOPS Oper. Syst. Rev.*, 40(4):333–344, 2006.

[21] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding user behavior in large-scale video-on-demand systems. In *EuroSys*, pages 333–344, 2006.