Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices

M. Zubair Shafiq[†] Lusheng Ji[‡] Alex X. Liu[†] Jia Wang[‡]
[†]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA
[‡]AT&T Labs – Research, Florham Park, NJ, USA
{shafiqmu,alexliu}@cse.msu.edu, {lji,jiawang}@research.att.com

ABSTRACT

Understanding Internet traffic dynamics in large cellular networks is important for network design, troubleshooting, performance evaluation, and optimization. In this paper, we present the results from our study, which is based upon a week-long aggregated flow level mobile device traffic data collected from a major cellular operator's core network. In this study, we measure and characterize the spatial and temporal dynamics of mobile Internet traffic. We distinguish our study from other related work by conducting the measurement at a larger scale and exploring mobile data traffic patterns along two new dimensions – device types and applications that generate such traffic patterns. Based on the findings of our measurement analysis, we propose a Zipf-like model to capture the volume distribution of application traffic and a Markov model to capture the volume dynamics of aggregate Internet traffic. We further customize our models for different device types using an unsupervised clustering algorithm to improve prediction accuracy.

Categories and Subject Descriptors

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

General Terms

Experimentation, Measurement, Performance, Theory

1. INTRODUCTION

1.1 Motivation

Since the emergence of cellular data networks, the volume of data traffic carried by cellular networks has been growing continuously due to the rapid increase in subscriber base size, cellular communication bandwidth, and cellular device capability. The recent unprecedented cellular data volume surge as the result of dramatic growth in the popularity of smart phones strongly suggests that the trend of cellular data growth will continue to accelerate as technology and application availabilities further improve [1]. To cope with the explosive cellular data volume growth and best serve their customers, cellular network operators need to design and manage cellular core network architectures accordingly. To achieve this, the first step is to understand the spatial and temporal patterns of Internet traffic carried by cellular networks. Understanding the spatial and temporal patterns of traffic can help to estimate both short- and long-term changes in network resource requirements.

1.2 Limitations of Prior Art

Cellular data traffic has not been well explored in prior work, although some attempts have been made [16, 14, 6]. The studies by Williamson *et al.* [16] and Trestian *et al.* [14] focused on jointly characterizing temporal dynamics of network traffic and user mobility. Their traffic traces contained data from about $10,000$ and $280,000$ users, respectively. Falaki *et al.* characterized diversity in smart phone activities (both in terms of user interaction with smart phones and the generated traffic) and linked it to battery consumption patterns [6]. Their traffic trace was collected from 255 users.

Prior work on cellular data traffic has four major limitations. First, the scales of these studies are not sufficient to be representative for the purpose of strategic level cellular operation planning. Second, no prior work has studied the behavior of different device types used to access cellular networks. However, understanding the behavior of different device types is important for billing and network resource planning. For example, knowing the different specifics of traffic that different device types tend to generate may help operators to construct appropriate promotions and rate plans. Third, no prior work has studied the behavior of network applications in cellular network traffic. However, understanding the behavior of different network applications is important because different applications have different demands on the quality of service. For example, if the volume of VoIP traffic (*e.g.* Skype) dominates P2P traffic (*e.g.* torrents), the service provider faces more demands on the quality of service, as compared to the opposite case. Finally, no prior work has developed predictive models for the spatial and temporal dynamics of cellular network traffic. However, the development of predictive models for cellular network is important for forecasting traffic trends and adjusting network resources accordingly.

SIGMETRICS'11, June 7–11, 2011, San Jose, California, USA.

1.3 Key Contributions

In this work, we study the traffic dynamics of a large operational cellular network. Our data set was collected from the core network of a major cellular service provider. In this paper, we first present the findings from our measurement studies. Second, based on the findings of our measurement analysis, we propose a Zipf-like model to capture the distribution and a Markov model to capture the volume dynamics of aggregate Internet traffic. We make key contributions from the following four perspectives:

1. **Scale of Study:** Our data set contains the logs of aggregated IP traffic generated by devices located in a major state of the USA. The usage data set is a summary of hundreds of terabytes of traffic from millions of cellular devices over the duration of a week.

2. **Behavior of Device Types:** We study a wide range of mobile devices in cellular networks. Our studies, with detailed analysis and characterization, show that different types of devices exhibit different traffic patterns. There are two main reasons. First, different devices have different capabilities. Second, different mobile devices are generally designed for attracting different population segments which often exhibit different usage behaviors.

3. **Behavior of Applications:** We study cellular network traffic characteristics against the wide range of applications that generated such traffic because different applications impose different demands on network resources and have different requirements on reliability and performance. Using application type as an additional dimension for characterizing dynamics of cellular network traffic offers finer granularity insights for network operators to understand how mobile devices demand network resources.

4. **Modeling Dynamics of Network Traffic:** We utilize results from measurement analysis to develop models for aggregate spatial and temporal dynamics of traffic in cellular networks. Since different types of devices show different traffic behaviors, we extend the aggregate model by customizing it for different types of devices to improve its prediction accuracy.

1.4 Our Findings

The results of our study reveal several interesting insights. We summarize the major findings of our study as follows: (1) The distribution of network traffic with respect to both individual devices and constituent applications is highly skewed. Only 5% of the devices are responsible for 90% of the total network traffic. Moreover, the top 10% applications account for more than 99% of the flows. Further, the distribution of traffic volume with respect to applications varies for different device types. These distributions can be modeled using Zipf-like models. (2) The aggregate volume of Internet traffic flowing on the network shows strong diurnal patterns. These diurnal patterns differ across weekdays and weekends. Moreover, the diurnal patterns of different cellular device types show subtle variations. The time-series of aggregate Internet traffic volume can be modeled using a multi-order discrete time Markov chain. (3) Finally, the behavior of different device types can be clustered into distinct subgroups. An unsupervised clustering algorithm such as the k-means algorithm can be utilized with spatial and temporal feature sets to effectively cluster device types. Using the identified subgroups, the model developed for aggregate traffic can be further extended to a more insightful and accurate multi-class model.

The rest of the paper proceeds as follows. In Section 2, we provide an overview of the cellular network architecture and describe the data set used in our study. In Section 3, we present measurement results of a week-long Internet traffic trace from a cellular network containing millions of devices. In Section 4, we develop a stochastic model to capture the spatial and temporal dynamics of aggregate network traffic. We then extend this model to a multi-class model by applying unsupervised clustering to identify subgroups of device types. We provide a review of the related work in Section 5 and conclude in Section 6.

2. BACKGROUND

2.1 Overview of Cellular Network Architecture

The cellular network that we study employs both second generation (2G) and third generation (3G) mobile data communication technologies that are part of 3rd Generation Partnership Project (3GPP) lineage. Figure 1 illustrates the architecture of the cellular network used for this study, in particular the components that are related to carrying IP data traffic. Such a cellular network can be visualized as consisting of three major segments: (1) the mobile cellular device; (2) the Radio Access Network (RAN), and (3) the Core Network (CN). The radio access network consists of base stations (named Base Transceiver Stations or BTS in 2G terms or Node B in 3G terms) and controllers (Base Station Controllers or Radio Network Controllers). The RAN controllers connect to the core network at nodes known as the Serving GPRS Support Nodes (SGSNs). In the core network, the mobile-facing SGSNs connect to the external-facing Gateway GPRS Support Nodes (GGSNs), which are responsible for providing connectivity to external networks such as the Internet and other private networks.

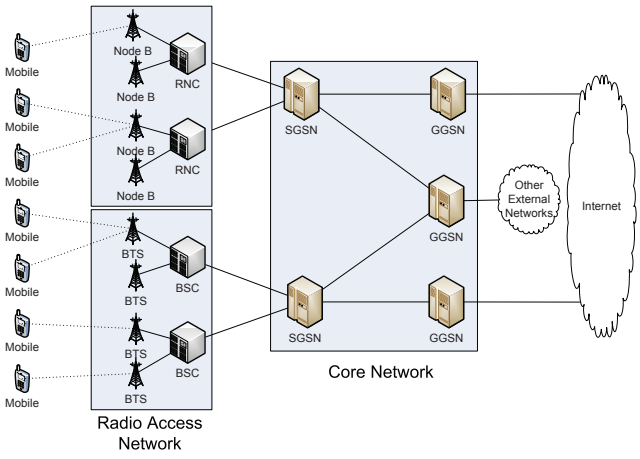


Figure 1: Architecture of a cellular network

2.2 Data Set Description

Our study is based on flow level mobile device traffic data collected from the cellular operator's core network. This

allows us to characterize the IP traffic patterns of mobile cellular devices and develop models that predict the bandwidth demands in the operator's core network over time. Due to the large volume of data and other limitations of our logging apparatus, we focus our study only in one particular state in the USA. This particular state was chosen because of log data availability, its geographical area, and population. That is, we only study the activities of mobile devices that are associated to base stations in that state. The data set covers activities during one whole week (18th to 24th) in January 2010. However, this data set does not contain complete temporal information due to some issues with the logging apparatus. Therefore, this data set is augmented with another aggregate data set only to study aggregate temporal traffic characteristics. The aggregate data set spans one whole week (14th to 20th) in June 2010. This data set also includes traffic data for two weekend days (12th and 13th June), which is only used for evaluation purposes. The aggregate temporal traffic results presented in this paper are from the second data set.

Each record contained in the aggregate data set is a summary report of activity during one particular flow by one mobile device. The records in the data set are indexed by a time stamp and a hashed mobile device identity. It is worth noting that we study traffic patterns of mobile devices instead of traffic patterns of users, which is also of more interest from operator's perspective. Each record in the data set also contains a cell identifier, which identifies the cell that serves the device, an application identifier, and data usage statistics for the flow, including total number of bytes, and total number of packets during that flow. A typical web-browsing activity, for example, may be represented by one flow record containing several packets of different sizes. These anonymous records were aggregated across all flow records and devices for analysis purposes. Different applications are identified using a combination of port information, payload signatures, and other heuristics. More details about application identification are provided in [5].

It is also worth noting that for privacy reasons the only device identifiers present in the data set are anonymized International Mobile Equipment Identifiers, or IMEI numbers. By design an original IMEI number uniquely identifies an individual mobile device. Such uniqueness is preserved by the anonymization process. Moreover, the anonymization preserves a portion of the IMEI number, known as the Type Allocation Code (TAC), which identifies the manufacturer and model of the device.

Our collected data set has two limitations that are mentioned below. First, the cell information in our data may not be accurate due to the fact that such information is obtained by monitoring GPRS Tunneling Protocol (GTP) message exchanges. Because GTP tunnel may remain intact despite device movements and handoffs, it is possible that a device initiates its data connection in a cell and thereafter moves across multiple cells [12] and such cell changes are not reflected in the data set as long as no GTP update is triggered by the device's movement. Partially due to these inaccuracies, user mobility characteristics are not part of this study. See reference [18] for quantification of the location inaccuracies in our data. Second, our data set, though covers complete population of one state with millions of users, only contains traffic information for one week time duration. This limitation is imposed due to huge volume of logged traffic records. Due to this, we cannot study long-term traffic patterns that span beyond one week time duration.

3. MEASURING INTERNET TRAFFIC DYNAMICS

In this section, we present the measurement results of the collected trace which spans a complete week and contains Internet traffic records of millions of cellular devices. As a first step, we study the distribution and temporal dynamics of aggregated Internet traffic. The insights gained by analyzing the distribution and temporal dynamics of Internet traffic are of significant importance for network management, traffic engineering, and capacity planning. Furthermore, we compare the traffic patterns of cellular devices from two popular mobile smart phone families and one cellular broadband modem family. The measurement results indicate significant differences in traffic patterns of different cellular device types.

3.1 Distribution and Temporal Dynamics of Aggregate Traffic

3.1.1 Traffic Volume Distribution

First we plot the distribution of traffic volume with respect to device identifier in Figure 2. Note that the curve approximately follows a straight line on a log-log scale across several orders of magnitude. We get a reasonably good fit for a Zipf model with index -0.57. This observation signifies that traffic volume in the cellular network is dominated by a small fraction of users.

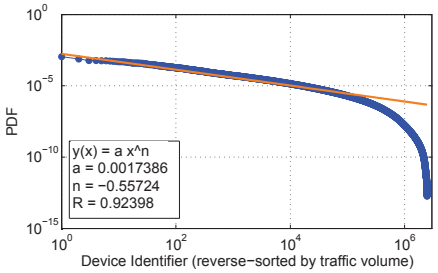
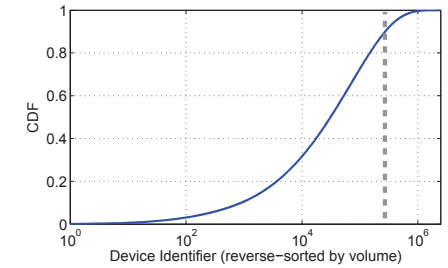


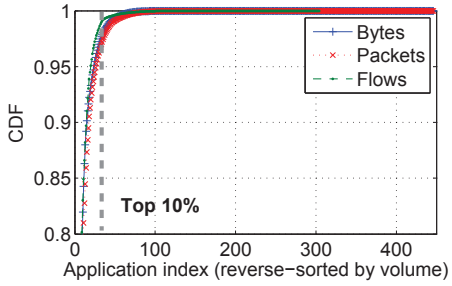
Figure 2: (Reverse-)Sorted distribution of traffic volume with respect to individual devices

Figure 3(a) shows the cumulative distribution function (CDF) plot of traffic volume with respect to device identifiers. In order to highlight the skewness in distribution, we have modified the x-axis to log-scale. It clearly shows that 5% of the devices are responsible for 90% of the total network traffic. The vertical dotted line partitions the top 5% devices on x-axis. A more careful look into the data reveals that in this data set the top-3 devices with respect to traffic volume belong to the family of wireless broadband modems. This observation is in accordance with our intuition as wireless modems are mostly plugged into desktop and laptop machines which provide more liberty to applications to utilize network resources. Moreover, desktop or laptop users tend to connect to the broadband network longer than handheld devices because the former has abundant power and storage resources, as well as more convenient user interfaces.

Traffic volume distribution can also be studied from a different perspective. Figure 3(b) shows the CDF of traffic

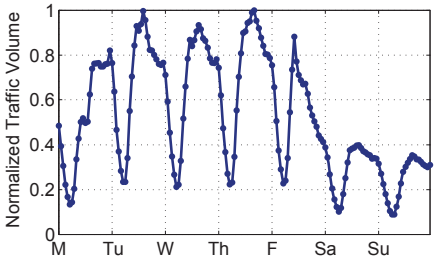


(a) Individual devices

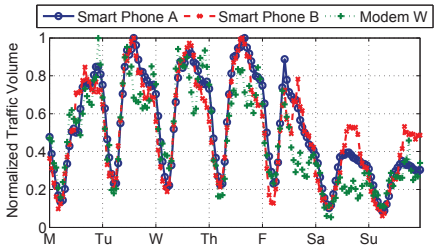


(b) Constituent applications

Figure 3: CDF plot of traffic volume



(a) Aggregate Traffic



(b) Separate Device Families

Figure 4: Diurnal characteristics of traffic volume over the duration of complete week

volume with respect to application identifiers. Just like the CDF plot of traffic volume with respect to device identifiers, it is evident that the distribution of traffic with respect to applications is highly skewed. The shape of the curve is similar for bytes, packets, and flows. However, the highest degree of skewness is observed for flows where the top 10% applications account for more than 99% flows.

3.1.2 *Temporal Dynamics*

It is also interesting to study the temporal dynamics of the logged traffic. In Figure 4(a), we plot time-series of the observed traffic volume at per hour granularity for the complete week. We clearly observe strong diurnal variations in aggregate traffic volume. This diurnality as well as several other features of the plot can all be reasonably explained by weekly working schedule of people. For instance, we observe a peak every day. The peak is centered around mid-day and lasts up to early evening. This indicates that people tend to vigorously use their cellular devices around lunch time and evening time compared to the rest of the working day – when they are busy at meetings, or are using office computers, and so forth. More insights regarding these peaks are further revealed in our analysis on the traffic patterns for different families of mobile devices later this section. In addition, the daily peaks observed on the weekdays are higher than those observed on weekends. This can be explained by less usage of wireless modem devices, some of which are likely the traffic heavy hitters, during the weekends.

3.2 Differentiating Cellular Devices

One intuitive way of dissecting the aggregate measurements is to separate out different types of devices. Different devices have different features and specifications, which may affect their traffic patterns. Moreover, different types of devices attract different groups of users, who may also use the cellular network in different ways. In this subsection, we attempt to differentiate the traffic patterns of different types of devices.

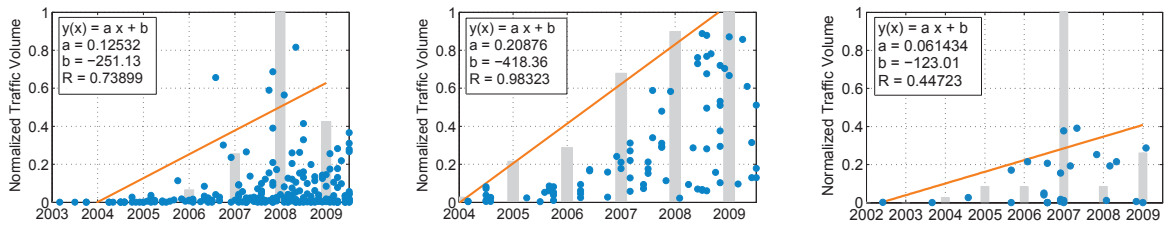
3.2.1 *Identifying Cellular Device Types*

As mentioned before, the TAC numbers of the device IMEI numbers are preserved by the hashed device identifiers in our data set. Such information can be used to identify the type, or more precisely the maker, model, and sometimes even version, of a cellular device by retrieving the corresponding TAC registration record from the GSM Association's TAC database. For the data set used in this study, we encountered approximately two thousand different TAC numbers which map to several hundred different types of devices.

Because of the large number of device types and the typically short lifespan of individual cellular device models, it makes more sense to compare cellular device families, for example the Nokia N series, instead of individual device types. Thus it is important to identify the lineage in devices of the same family. Moreover, it also offers a historical perspective into how data usage patterns change along the evolution path of cellular devices of the same lineage.

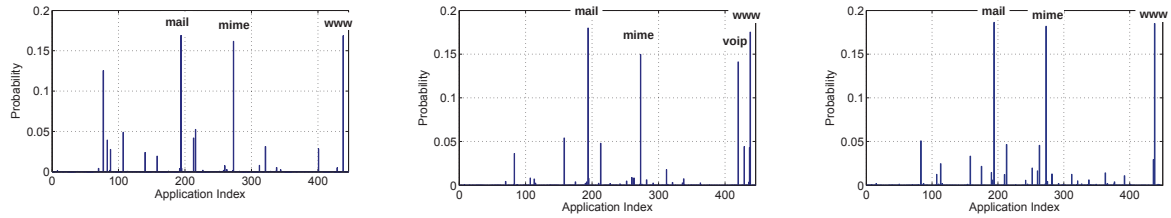
Normally the manufacturing time of a particular device or even a particular model is difficult to determine from public domain knowledge. In our study, we tackle this problem by using a simple heuristic for estimating the manufacturing time of a device. Because the TAC numbers are specific to particular device models and there are only limited IMEI numbers under each TAC lot, it is reasonable to assume that manufacturers apply for TAC numbers from the GSM Association according to their production plans. Thus, there is a correlation between the registration time of a TAC number and the manufacturing time of cellular devices with that TAC number. Hence, we use the TAC registration time for classifying devices when we want to study how device data usage pattern changes as device specification and configuration may change over time.

In the discussions below, our analysis will focus on the comparison between statistics of smart phone devices from



(a) Smart Phone A (b) Smart Phone B (c) Wireless Modem W

Figure 5: Variation in traffic volume for smart phone A, smart phone B and wireless modem W devices manufactured in recent years



(a) Smart Phone A (b) Smart Phone B (c) Wireless Modem W

Figure 6: Volume distributions of applications constituting network traffic from different device families

two popular families, denoted as smart phone A and smart phone B. The choice of studying smart phones instead of traditional phones is relatively easy because smart phones are generally more capable and user-friendly for Internet usage. We have selected the two particular smart phone families because both are popular in different user markets – smart phone A models are popular more among general consumers whereas smart phone B models are largely adopted by business customers. The contrast in usage patterns between these two product lines will provide important insights into the behavioral differences between these two distinct classes of customers.

We will also compare statistics of smart phone A and smart phone B with those of a wireless modem cards family (denoted by W). These wireless modem cards provide cellular broadband connectivity to traditional desktops, laptops, or netbooks. As shown previously, this class of devices is also a major contributor of cellular Internet traffic. In addition, it is reasonable to believe that the traffic patterns of these modem devices resemble more traffic patterns seen on wired Internet because the equipment behind these modems is similar to those on the wired Internet. Thus, they form a baseline for comparing Internet traffic patterns and dynamics.

3.2.2 *Traffic Temporal Dynamics of Different Device Families*

We first revisit the traffic temporal dynamics of different device families. Previously, Figure 4(a) showed the aggregate Internet traffic volume over time. Here we separate out traffic volumes for the three cellular device families, smart phone A, smart phone B, and wireless modem W, and plot them individually in Figure 4(b). Note that we normalize the traffic volume of each device family by the maximum value for the respective device family.

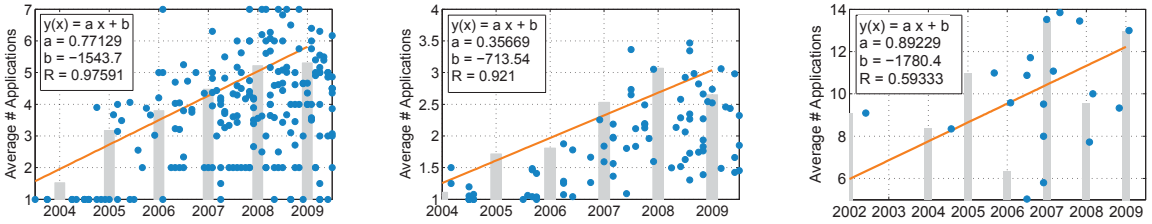
The differences in plots of different device types can be explained if we restate the common impression that smart phone B devices are favored more by business users and smart phone A devices are popular among general consumers. For example, on weekdays, the peak around mid-day is higher for smart phone B devices as compared to smart phone A devices whereas the peak at night is relatively higher for smart phone A devices as compared to smart phone B devices. However, note that this trend is reversed on weekends when smart phone B devices have higher peak in afternoons. This observation can be explained by the reasoning that on weekends business customers rely heavily on their smart phone B to remain updated about business-related activities whereas on weekdays they usually have access to their office desktops or laptops.

3.2.3 *Traffic Volume*

Figure 5 shows the variation in average normalized traffic volume from devices manufactured in different years. Note that each dot represents the result for a particular model which is identified by its TAC registration date. The x-axes of the figures for each device family start from the year when TAC was registered for its first model. The grey bars represent the average for a year. The regression line is plotted for the average yearly values. It is apparent that for both smart phone families, later models tend to generate more traffic. However, there is an outlier peak for smart phone A at 2008 and this trend is not obvious for wireless broadband modem family, which is indicated by the relatively small slope of its regression line and lower goodness of fit value (R). This is reasonable because later models tend to support newer communication technologies, with more powerful computing engines and friendlier user interfaces. All of the above-mentioned factors encourage more data usage from users.

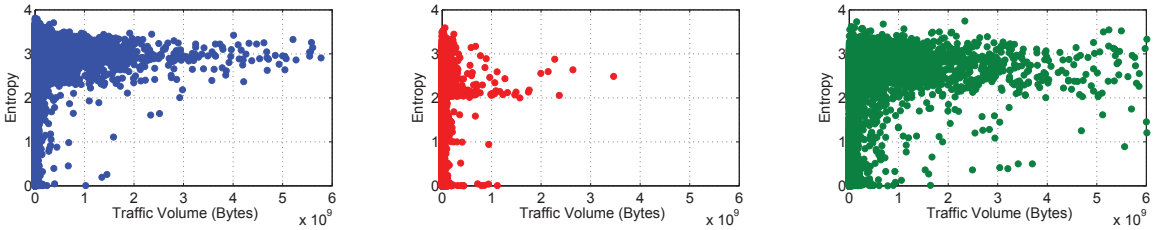
3.2.4 *Volume Distribution of Applications*

Figure 6 provides the traffic volume distributions with respect to constituent applications for different device types. It is clear that each device family has *different* traffic behaviors. An interesting finding is that, for each device family, most top peaks in the volume distribution are for same applications. These peaks correspond to e-mail and web traffic, which are prevalent on all device families.



(a) Smart Phone A (b) Smart Phone B (c) Wireless Modem W

Figure 7: Variation in number of applications for smart phone A, smart phone B, and wireless modem W device families



(a) Smart Phone A (b) Smart Phone B (c) Wireless Modem W

Figure 8: Entropy of application volume histogram for different device families

3.2.5 *Diversity of Applications*

Figure 7 provides the variation in average number of unique applications accessed by cellular devices manufactured in different years. First, we note that, for both smart phone A and smart phone B devices, the average number of unique applications accessed by a device shows an increasing trend across device manufacturing years. However, this trend is not obvious for wireless modem W. Second, it is clear that the average numbers of unique applications accessed by smart phone A devices and wireless modem W devices are significantly more than that by smart phone B devices. The number of unique applications accessed by a cellular device, which we refer to as *application diversity*, is an indicator of the device's versatility.

To quantitatively compare the diversity of applications constituting devices' traffic, we calculate the entropy of their application volume distributions. Entropy quantifies the spread of probability distribution of a random variable. For a given random variable X, its entropy $H(X)$ is given as: $H(X) = \sum_{\forall x_i \in X} x_i \log_2(x_i)$. Figure 8 shows the scatter plot of entropy of application histogram versus total volume. Note that in these plots each dot represents a unique device. For the baseline comparison, we also provide a scatter plot for all wireless modem W devices (as they are usually plugged into powerful desktop machines or laptops). As per our expectations, the entropy and total volume for smart phone A devices is significantly more than those of smart phone B devices. This is essentially indicated by the size of the *bulge* towards the top-right in scatter plots. The wireless modem W devices tend to have the highest entropy and total volume.

3.3 Summary

In this section, we have presented measurement and analysis for the distribution and the temporal dynamics of aggregated Internet traffic. We have also separately analyzed the traffic from different cellular families. We have shown that the aggregate traffic distribution is highly skewed both across different kinds of applications and different cellular devices. Furthermore, our study reveals that different groups of cellular devices indeed behave differently in terms of their Internet usage. Such differences are not only present between different kinds of cellular devices, *i.e.* smart phones vs. modem cards, but also are obvious among different groups of cellular devices of the same kind but favored by different market segments and user groups. Based on the findings stated above, we will now formally model the distributions and the temporal dynamics of Internet traffic from cellular devices. Similar to the measurement study in this section, we begin our modeling with aggregate traffic and then refine the models by taking cellular device population composition and sub-group characteristics into consideration.

4. MODELING INTERNET TRAFFIC DYNAMICS

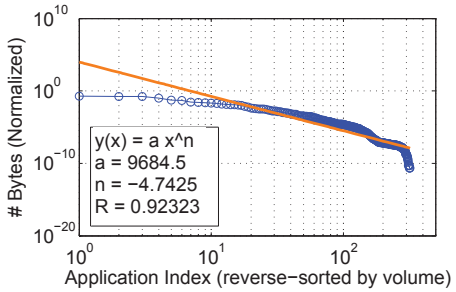
In this section, we first use a Zipf-like distribution to model the long term distribution of Internet traffic volume versus constituent applications. Second, we use a Markov chain model to capture the temporal dynamics of aggregated Internet traffic volume. Then, we enhance the models with a multi-class approach by applying unsupervised clustering on different types of devices. The multi-class model can more accurately capture the distribution and temporal dynamics of Internet traffic. At the end of this section, we evaluate the improvement provided by the proposed multi-class model with respect to the aggregate traffic model.

4.1 Aggregate Traffic Model

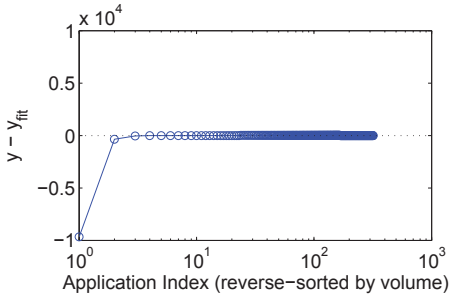
4.1.1 *Modeling Long Term Distribution of Traffic*

It has been shown that the popularity distribution in World Wide Web (WWW), User Generated Content (UGC), and channel popularity in IPTV systems is scale-free [10]. From our observations in Section 3, we know that the distribution of Internet traffic (in terms of bytes, packets, and flows) is highly skewed. It can be observed in Figure 3(b) that top 10% of the applications constitute about 99% of the flows.

This observation naturally leads to a Zipf-like model. In a Zipf model, an object of rank x has probability p: $p \sim x^{-b}$. Figure 9(a) shows the distribution plot of volume versus application index averaged for the complete week. The residual plot in Figure 9(b) demonstrates that this Zipf-like model has reasonable accuracy.



(a) Zipf model



(b) Residual plot for Zipf model

Figure 9: The Zipf model for long term average distribution of traffic volume patterns

4.1.2 Modeling Temporal Dynamics of Traffic

The temporal dynamics of traffic volume can be represented as a random process V. So, let its vector representation be $V =< V_1, V_2, ..., V_i, ... >$, where V_i denotes the traffic volume at time index i. Note that we can analyze the traffic volume at different time resolutions; however, in the rest of this paper we will only consider the traffic volume at hourly time resolution. Without loss of generality, we can aggregate consecutive n entries in V as a single element. For example, if $V =< V_1, V_2, V_3, V_4, V_5 >$, and we aggregate two consecutive entries as a single element (*i.e.* $n = 2$), we produce a new sequence as $< V_1V_2, V_2V_3, V_3V_4, V_4V_5 >$. This up-scaling, however, increases the dimensionality of the distribution from k to k^n, where k is the dimensionality of the original time series. It not only increases the underlying information of our process but may also result in sparse distributions due to requirement of large training data. Therefore, an inherent tradeoff exists between the amount of information – characterized by entropy – and the minimum training data required to build a model.

It is important to note that the up-scaled sequence with $n = 2$ is in fact a simple joint distribution of two sequences with $n = 1$, and so on. The joint distribution may contain some redundant information which is not relevant for a given problem. Therefore, we choose to remove the redundancy by using the conditional distribution for a more accurate analysis. The use of conditional distribution, instead of joint distribution, reduces the size of the underlying sample space which corresponds to removing the redundant information from the joint distribution. Using conditional distribution also enables us to model the traffic volume time series as a discrete time Markov chain. Here we do not evaluate other well-known statistical time series modeling approaches such as Box-Jenkins methodology due to limited available training data (only one week) [2]. Such time series modeling approaches require large run of time series training data and may be used if enough training data is available.

In this paper, we use a discrete time Markov chain to model the traffic time series. An important parameter to determine when modeling a stochastic process with a Markovian model is the order of the Markov chain. The order is equivalent to the level of up-scaling n mentioned above. The order represents the extent to which past states determine the present state, *i.e.*, how many lags should be examined when analyzing higher orders. The rationale behind this argument is that if we take into account more past states, less surprises or the uncertainties are expected in the present state. Towards this end, we have analyzed a number of statistical properties of the traffic volume time-series. A relevant property that has provided us interesting insights into the statistical characteristics of traffic time-series is the *autocorrelation* [4]. Another relevant property that can be helpful in determining the suitable value of n is the *relative mutual information* [8]. We discuss both of these properties for our data below.

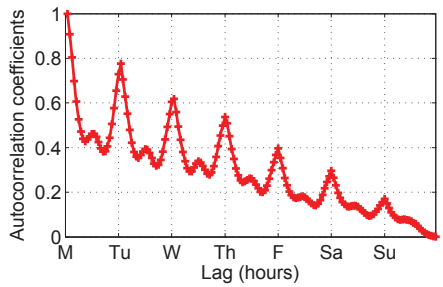
(1) Autocorrelation: Autocorrelation is an important statistic for determining the order of a sequence of states. Autocorrelation describes the correlation between the random variables in a stochastic process at different points in time or space. For a given lag t, the autocorrelation function of a stochastic process, V_m (V denotes traffic volume process and m is the time index), is defined as:

$$\rho[t] = \frac{E\{V_0V_t\} - E\{V_0\}E\{V_t\}}{\sigma_{V_0}\sigma_{V_t}}, \quad (1)$$

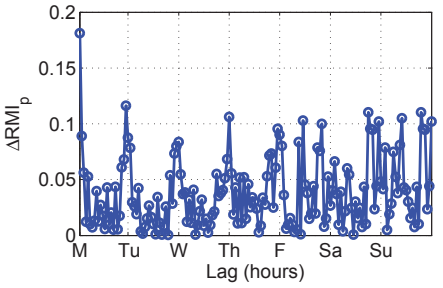
where $E\{.\}$ represents the expectation operation and σ_{V_m} is the standard deviation of the random variable (representing traffic volume) at time lag m. The value of the autocorrelation function lies in the range $[-1, 1]$, where $\rho[t] = 1$ means perfect correlation at lag t, and $\rho[t] = 0$ means no correlation at all at lag t.

To observe the dependency level in a sequence of traffic volume V, we calculate sample autocorrelation functions for the one week aggregate volume trace. Figure 10(a) shows the sample autocorrelation functions plotted versus the lag. First, we note that the value of the autocorrelation function steadily decays over the week. Clearly, the dependency of traffic volume at a given time instance on time-lagged traffic volumes should decrease as the time lag increases. Second, the traffic volume at a given time instance shows the strongest dependence on the previous states that lag by multiples of 24 hours. This is indicated by the autocorrelation peaks at $n \approx 24, 48, 72, ...$ This effect is due to the diurnal (non-stationary) nature of the patterns observed in our data. These observations will be helpful to select the appropriate order for the Markov chain model.

(2) Relative Mutual Information: Another interesting statistic that provides insight to determine order of a stochastic process is called relative mutual information. Relative mutual information quantifies the amount of information that a random variable V_t provides about V_{t+1} (separated by one unit of time lag) while providing a measure of



(a) Autocorrelation



(b) Relative mutual information

Figure 10: Analysis techniques to determine temporal dependency in traffic volume time-series

the remaining uncertainty about V_{t+1} [8]. Mathematically,

$$RMI(V_{t+1}, V_t) = \frac{I(V_{t+1}; V_t)}{H(V_{t+1})}$$

where $I(V_{t+1}; V_t)$ is information gain and $H(V_{t+1})$ is entropy. Clearly, RMI is a non-symmetric measure and it is bounded in the range $[0, 1]$. The values of RMI approaching one indicate high dependency and the values approaching zero indicate low dependency. Note that an arbitrary number m of previous states can be included.

$$RMI(V_{t+1}, ..., V_2, V_1) = \frac{I(V_{t+1}; V_t, ..., V_2, V_1)}{H(V_{t+1})}$$

However, the computation complexity of RMI increases exponentially with respect to the number of previous states under consideration. A variant of RMI is called pair-wise relative mutual information RMI_p which is computed only between a random process and its lagged version. The maximum lag for which $\Delta RMI_p = |RMI_p(m-1) - RMI_p(m)|$ remains greater than ϵ defines the order of underlying stochastic process [8]. With pair-wise relative mutual information, the order of underlying stochastic process is determined as:

$$M(\epsilon) = max\big(|RMI_p(m-1) - RMI_p(m)|\big) \geq \epsilon,\ \forall m \in [1, \infty)$$

Figure 10(b) shows the plot of ΔRMI_p for aggregate traffic time-series. We note that the dependency between two time lags shows a repetitive pattern. Using the methodology described above, the order of this process is determined to be 24. In other words, there is an obvious redundancy beyond time difference of 24 hours.

The results of autocorrelation and relative mutual information measures highlight the dependency of traffic volume on the previous 24 hours; therefore, we use a 23rd order discrete time Markov chain. A nth order discrete time Markov chain can be visualized by considering all possible values of states at previous n lags. The state space of our Markov chain model represents discretized traffic volume. For an nth order discrete time Markov chain with q elements in state space, we have the transition probability matrix $\mathbf{T}$ with q^n rows and columns. Notice that each row has the transition probabilities of going out from the respective state. Consequently, the probabilities in a row sum up to 1.

4.1.3 *Forecasting Internet Traffic Dynamics*

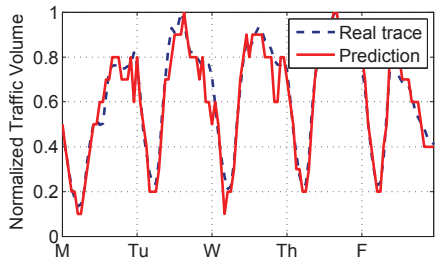
Note that for a given nth order Markov chain with q possible values of states, the total number of probability parameters denoted by $|P|$ is $(q-1)q^n$. For the present case where $n = 23$ and $q = 10$ (if we quantify traffic volume into 10 discrete levels) this will result in 9 x 10^{23} probability parameters. Clearly, we need to significantly reduce the number of probability parameters in our multi-order Markov model. Towards this end, we limit the number of probability parameters by using a many-to-one mapping. This mapping is essentially determined by the amount of data samples available to train the model. For each training sample, we can update the value of at most one probability parameter.

Once we have trained our model, we can use it to forecast future traffic volume. More specifically, given previous n states of this process (V_1, V_2, ..., V_n), can we predict the next state, *i.e.* V_{n+1} with reasonable accuracy? To make sure that with our choice of the Markovian order and the reduction of states the model can still accurately describe the data set, we now evaluate our proposed model using the collected traffic trace.

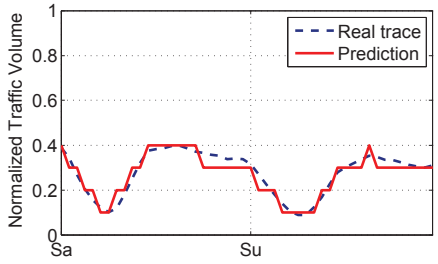
Recall from Section 3 that traffic time-series shows different behavior for weekdays and weekend. Therefore, we separate the proposed Markov model for aggregate traffic volume into two independent sub-models – one for weekday and one for weekend. For weekday traffic, we initially train our model using Monday's traffic data. The testing is then carried out for the remaining weekdays, comparing the model produced data with the actual data in the traffic data set. To evaluate the performance of our model on weekend traffic, we obtained additional data records for the previous weekend and train our model with them. The testing is then carried out for the next weekend similarly to weekday testing by comparing model produced volume with actual volume in data set. We further improve the accuracy of our stochastic model by utilizing online feedback to update the underlying probability parameters.

The result of our experiment shows that our model successfully captures the dynamics of Internet traffic volume with a reasonably small mean squared error (MSE) value (= 1.7 x 10^{-4}). Figure 11 shows the plot of our model's forecast values along with the actual trace values. It is evident that our model successfully reproduces most of the diurnal behavior observed in the aggregate traffic volume trace.

It is worth noting that not only the models we have developed can be used to formally describe cellular devices's Internet traffic distribution and dynamics, they are more valuable in forecasting future traffic. More specifically, given previous n states of this process (V_1, V_2, ..., V_n), we can predict the next state, i.e. V_{n+1} with reasonable accuracy, assuming the underlying fundamentals such as device usage behavior and device population composition are not changed. We have catered to the changing device usage behavior by using online feedback. However, device population composition



(a) Weekday



(b) Weekend

Figure 11: Traffic volume forecast based on the proposed Markov model

slowly changes over time resulting in degraded model accuracy. To overcome this issue and to further improve the accuracy of our proposed model, we now refine our model for different devices as they may exhibit vastly different behaviors and traffic patterns.

4.2 Multi-class Model

Previously we have developed a Zipf-like model to capture the traffic volume distribution for constituent applications and a multi-order Markov model to capture the temporal dynamics of cellular devices' Internet traffic. Both models are for aggregate Internet traffic of cellular devices. However, as we have shown in the Section 3, different devices may exhibit vastly different behaviors and traffic patterns. A naive extension of this model will be to develop a specialized model for every device type. However, we have several hundred different device types and having a separate model for each device type is not feasible. Hence, the natural next step is to further identify groups in device population with similar characteristics and refine the models.

We follow a two step methodology to develop such grouping. First, we study different feature sets that can be utilized to cluster the devices. Second, we examine the outcome of clustering using different feature sets to determine the suitable grouping methodology. This examination provides interesting insights which may help determine the reasons which lead to such grouping. Once we have the final grouping, we extend our model for aggregate traffic to a multi-class model of traffic distribution and temporal dynamics.

4.2.1 *Grouping Strategies*

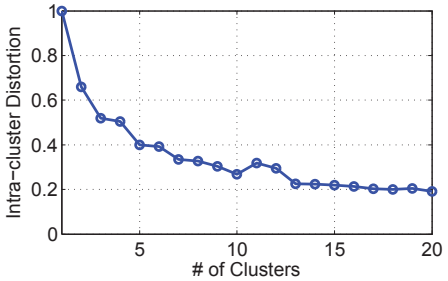
We now take a look at different ways using which we can group device population. Note that the objective of our grouping methodology is to combine the devices with similar traffic characteristics into a handful number of clusters so that we can train separate and independent models for each of these groups. Towards this end, we propose the following simple yet effective feature sets for clustering device types.

(1) Average Traffic Volume per Application: It is a 100 element tuple which represents normalized average traffic volume for top 100 applications with highest aggregate volume for a given device type.
(2) Average Traffic Volume per Hour: It is a 24 element tuple which represents normalized average traffic volume at each hour of the day for a given device type.

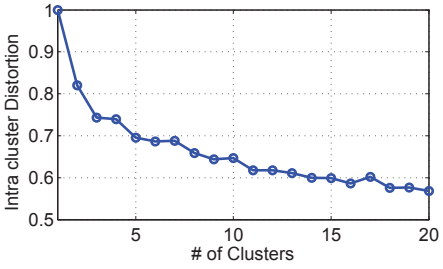
We utilize an unsupervised clustering algorithm to cluster the device types into groups. Towards this end, we have selected the well-known k-means clustering algorithm which has definite advantages over other clustering techniques especially for large number of variables and large data sets [9]. It is important to set an appropriate value of k in k-means clustering algorithm. Note that our goal is to obtain multiple representative models of our data that can be used later to extend our single aggregate model to the multi-class model. To limit the number of classes in the multi-class model, we are interested in finding the minimum number of clusters that can capture distinct underlying behaviors in our data. We use intra-cluster dissimilarity D_k measure to select the appropriate value of k. We calculate the value of D_k for increasing values of k starting from $k = 2$. Intra-cluster dissimilarity is defined as:

$$D_k = \sum_{j=1}^{k} \sum_{i \in C(j)} |x_i - \hat{x}_j|,$$

where x_i is a data point residing in j-th cluster, $\hat{x}_j$ is the centroid point of j-th cluster. Figure 12 shows the variation in the values of D_k for increasing values of k. We expect the values of D_k to mostly decrease for increasing values of k. We select the value of k to be the least value for which either $D_k - D_{k+1} \to 0^+$ or $D_k - D_{k+1} < 0$ [13]. For both spatial and temporal features, in Figures 12(a) and 12(b), $D_3 - D_4 \to 0^+$; thus, $k = 3$ for both cases.

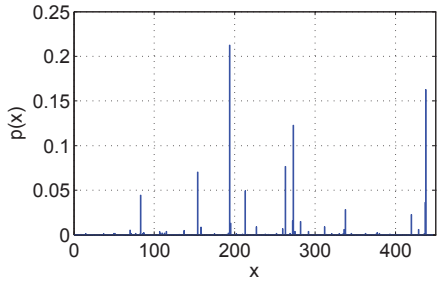


(a) Average Traffic Volume per Application

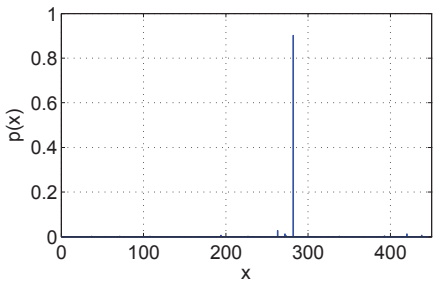


(b) Average Traffic Volume per Hour

Figure 12: Variation in intra-cluster dissimilarity with respect to increasing number of clusters



(a) High Diversity (HD)



(b) Low Diversity (LD)

Figure 13: Cluster centroids for spatial features

4.2.2 *Explaining Internet Traffic Dynamics for Identified Clusters*

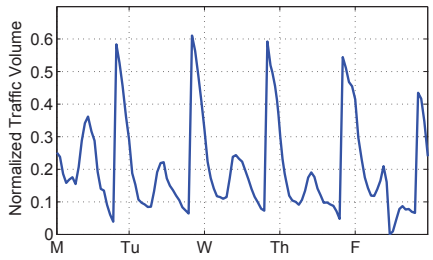
In Section 3.1, we studied traffic volume distribution across different applications and temporal dynamics of aggregate Internet traffic. Now, we want to study the behaviors characterized by the identified clusters. We have used two feature sets to cluster device population into distinct groups. Here we discuss the clustering results of both feature sets separately in the following text. We will then use these results to explain the characteristics of traffic from two popular mobile smart phone families and one cellular broadband modem family.

We can label the identified centroids using spatial features as High Diversity (HD), Medium Diversity (MD), and Low Diversity (LD). In Figure 13, we plot centroids of two of the three clusters. By diversity, we are referring to the variation in traffic application distribution, which in turn is quantified using entropy. It is clear that the centroid model plotted in Figure 13(a) has higher entropy as compared to the one plotted in Figure 13(b) which is mostly dominated by traffic of one particular application.

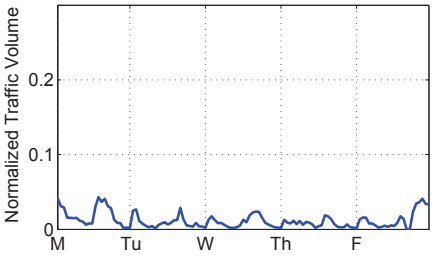
It is interesting to see how cellular devices belonging to different device families are distributed among different clusters based on the above clustering technique. These results will enhance our understanding of device behavior from different manufacturers. Again we list the same three device families as in Section 3. Table 1 shows the percentage distribution of cellular devices made by different device families over different cluster groups, which portraits a more detailed image than Figure 6.

Table 1: Population distribution of device families based on clustering using spatial features

	Wireless Modem W	Smart Phone A	Smart Phone B
HD (%)	**79.3**	**94.4**	**76.8**
MD (%)	0.0	5.2	0.0
LD (%)	20.7	0.4	23.2



(a) High Volume (HV)



(b) Low Volume (LV)

Figure 14: Cluster centroids for temporal features

The analysis of cluster centroids obtained from temporal features also provide interesting insights about distinct traffic behavior of different device groups. Figure 14 shows the plots for 2 of the cluster centroids from k-means clustering. We have labeled the cluster centroids based on their volume characteristics as high/medium/low volume. The traffic volume is normalized by the maximum observed value for every device type. We define the volume category of a centroid to be high, medium, or low by taking the average of peak values for weekdays only. We only consider weekday peak values because traffic volume on weekdays is significantly higher than weekends for aggregate traffic time series in Figure 4(a). If the average normalized volume for weekdays is more than 0.5 then the assigned volume category is *high*. Else if average normalized volume is less than 0.5 and more than 0.1 then it is categorized as *medium*. Finally, if the normalized volume is less than 0.1 then it is categorized as *low* The thresholds for such volume partitioning are selected after manually analyzing all centroids. There are 3 cluster centroids based on temporal features, high volume HV, medium volume MV, and low volume LV. Two of the temporal cluster centroids are shown in Figures 14(a) and 14(b).

We again analyze the distribution of devices from different device families across these clusters. First, we note that almost 70% of Smart Phone A devices fall into HV cluster indicating that the owners of these devices tend to use them heavily throughout the week. On the other hand, the Smart Phone B devices spread more into LV cluster indicating that Smart Phone B owners use them less rigorously as compared to Smart Phone A devices. Wireless Modem W devices are more evenly spread across all clusters as compared to Smart Phone A and Smart Phone B.

To conclude, our clustering results highlight that different groups of devices do have distinct traffic behaviors and using our clustering method these different groups can be partitioned out of the device population. Because the distinctions between different groups are concealed by the aggregate traffic model, as a next step we extend our aggregate

Table 2: Population distribution of device families based on clustering using temporal features

	Wireless Modem W	Smart Phone A	Smart Phone B
HV (%)	**48.3**	**69.0**	15.9
MV (%)	31.0	16.5	27.1
LV (%)	20.7	14.5	**57.0**

traffic model proposed in Section 4.1 to a multi-class model. Such multi-class model can describe the traffic patterns and dynamics in a better way.

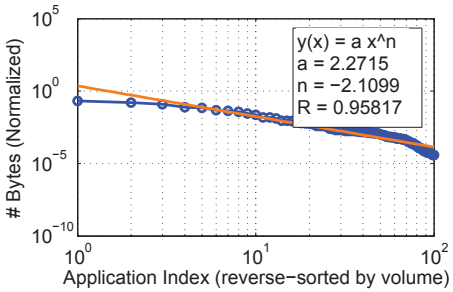
4.2.3 *Evaluation of the Multi-class Model*

We now use the clustering results to extend the aggregate traffic model to a multi-class model. Note that we are primarily interested in accurately describing the volume distribution across different applications and temporal dynamics of cellular devices' Internet traffic. We follow a three-step methodology in this regard. First, we aggregate the traffic from all types of devices that fall into the same cluster. Second, we normalize the cluster aggregated traffic with respect to its relative proportion in the aggregate traffic which is determined empirically. Finally, we model each of the aggregated and normalized traffic traces separately. Note that we model the spatial and temporal dynamics of traffic separately. Remember that we have three clusters for both spatial features and temporal features. So, in the eventual multi-class model we obtain three Zipf-like characterizations for the distribution of Internet traffic and three Markov chain based models to capture the temporal dynamics of the traffic.

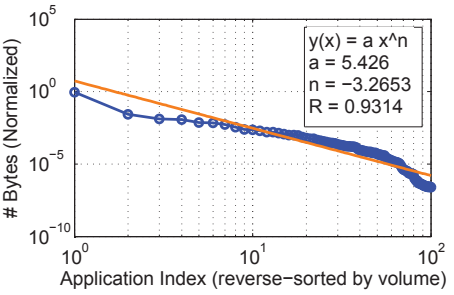
Figure 15 shows the plots of Zipf-like distribution models for HD and LD classes. To evaluate the improvement in accuracy for the multi-class model as compared to the aggregate model, we compare both to the real trace. We note that the average value of R (which quantifies goodness of fit) improves to 0.96 for multi-class models as compared to 0.92 for the aggregate model.

Figure 16 shows the plot of predictions from multi-order Markov models trained for two of the classes (HV and LV). It is evident that the predictions of Markov models are reasonably accurate. The value of average MSE for all three classes is 9.2 x 10^{-5} which is lower than the value achieved by the aggregate model. To conclude, our multi-class model improves on the single-class (aggregate) model in terms of prediction accuracy.

Once again, the multi-class extended models can also be used for predicting future traffic patterns just like the models for aggregate traffic. Recall that device population composition slowly changes over time which degraded the accuracy of aggregate model. However, we can update the device population composition by periodically refreshing clustering results used by the multi-class model. Therefore, we can successfully eliminate the root-cause of accuracy degradation from multi-class model which results from changing device population composition.

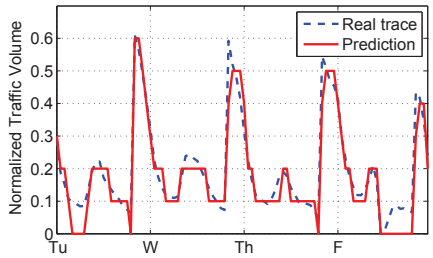


(a) HD

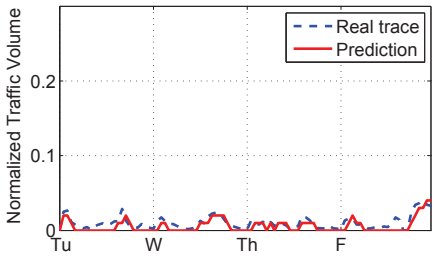


(b) LD

Figure 15: Separate Zipf-like characterizations for two of the classes (obtained by clustering using spatial features)



(a) HV



(b) LV

Figure 16: Prediction of multi-order Markov model for two of the classes (obtained by clustering using temporal features)

5. RELATED WORK

Several related works analyze usage data from cellular networks. In [17], Willkomm *et al.* perform measurement and modeling of voice call data collected from a CDMA-based cellular operator. In [16], the authors carry out a low level measurement analysis on a CDMA2000 cellular data network. The results of their experiments show that user data traffic is bursty and shows strong diurnal patterns. In [19], the authors perform a measurement study of Short Message Service (SMS) of a nationwide cellular network. In contrast to the above-mentioned studies, our work focuses on measurement and modeling of *distribution and temporal dynamics of data traffic in a cellular network.*

In [14], the authors analyze the relationship between the types of applications accessed and user mobility in a 3G cellular network. The results of their measurement studies show that there is a strong relationship between the types of applications accessed and mobility patterns of users. The content access patterns quantified in [14] are limited to six general categories, namely mail, music, social network, news, trading, and dating. On the other hand, in our work we analyze more than 400 fine-grained application categories. Moreover, in our paper we model the distribution and temporal dynamics of content access patterns. In a recent relevant work [6], Falaki *et al.* study traces from 255 users to study their interaction with smartphones. They collected data by deploying a custom logger on smartphones. The results of their experiments show that user interaction has diurnal patterns and that a few applications dominate the rest. In contrast to this work, our work focuses on data traffic analysis as seen by cellular network. Also, the scale of our study is significantly larger – containing data from millions of devices and several hundred unique device types.

Several additional related works use similar modeling methodologies. In [7] and [20], the authors perform measurement and modeling studies for YouTube traffic at different points in the network. In [7], the authors collect traffic between YouTube and an edge network. Relevant to our work, the authors model video popularity using Zipf distribution. This result is also verified by findings reported in [20]. In [20], the authors further show that the distribution of number of video requests per client follows power-law distribution. Relative to these studies, we have modeled the steady-state distribution of application in content access patterns using Zipf-like distribution. In [3], Cao *et al.* utilize stochastic models for source-level modeling of HTTP traffic. Likewise, the technique proposed in [11] accomplishes a similar task for flow-level traces. In [15], the authors have proposed a packet-level network traffic generator which utilizes a structural model to capture interactions of applications and users. The model trains itself on a given packet trace and then generates live packet traces using the trained models. In relation to these studies, our proposed technique also trains itself on a given trace capturing characteristic features of Internet traffic dynamics. Afterwards, the trained models are used to predict/generate live realistic traces.

6. CONCLUDING REMARKS

In this paper, we have presented an analysis of Internet traffic dynamics of cellular devices in a large cellular network. The results of our measurement and modeling experiments have important implications on cellular network design, troubleshooting, performance evaluation, and optimization. For example, the skewness of traffic distribution with respect constituent applications implies that only a few applications are popular. Therefore, cellular device manufacturers and software developers can focus on the specific characteristics of the popular applications for performance optimization. Furthermore, the diurnal variations observed in this paper imply that the network usage is strongly non-stationary. Cellular network operators typically do resource allocation based on peak usage requirements and these resources are wasted during non-peak time. To mitigate this resource wastage, cellular network operator can devise billing schemes to differentiate between peak and off-peak network usage.

Acknowledgements

We would like to thank Alexandre Gerber and Jeffrey Erman for providing technical comments on the paper, and Jeffrey Pang for helping us in general understanding of the traffic logging apparatus. We would also like to thank our shepherd, Alberto Lopez Toledo, and the anonymous reviewers for their helpful comments and suggestions.

7. REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010-2015. White Paper, February 2011.

[2] G. Box, G. M. Jenkins, and G. Reinsel. *Time Series Analysis: Forecasting & Control*. Wiley Series in Probability and Statistics, 4th edition, 2008.

[3] J. Cao, W. S. Cleveland, Y. Gao, K. Jeffay, E. D. Smith, and M. Weigle. Stochastic models for generating synthetic HTTP source traffic. In *IEEE INFOCOM*, 2004.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[5] J. Erman, A. Gerber, M. T. Hajiaghayi, D. Pei, and O. Spatscheck. Network-aware forward caching. In *WWW*, 2009.

[6] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. Diversity in smartphone usage. In *MobiSys*, 2010.

[7] P. Gill, M. Arlittz, Z. Li, and A. Mahantix. YouTube traffic characterization: A view from the edge. In *ACM SIGCOMM IMC*, 2007.

[8] M. Ilyas and H. Radha. On measuring memory length of the error rate process in wireless channels. In *Conference on Information Sciences and Systems (CISS)*, 2008.

[9] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Math Statistics and Probability*, 1967.

[10] T. Qiu, Z. Ge, S. Lee, J. Wang, Q. Zhao, and J. Xu. Modeling channel popularity dynamics in a large IPTV system. In *ACM SIGMETRICS*, 2009.

[11] J. Sommers and P. Barford. Self-configuring network traffic generation. In *ACM SIGCOMM IMC*, 2004.

[12] S. Tekinay and B. Jabbari. Handover and channel assignment in mobile cellular networks. In *IEEE Communications Magazine*, 1991.

[13] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63:411–423, 2001.

[14] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Measuring serendipity: Connecting people, locations and interests in a mobile 3G network. In *ACM SIGCOMM IMC*, 2009.

[15] K. V. Vishwanath and A. Vahdat. Realistic and responsive network traffic generation. In *ACM SIGCOMM*, 2006.

[16] C. Williamson, E. Halepovic, H. Sun, and Y. Wu. Characterization of CDMA2000 cellular data network traffic. In *IEEE Conference on Local Computer Networks*, 2005.

[17] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz. Primary users in cellular networks: A large-scale measurement study. In *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, 2008.

[18] Q. Xu, A. Gerber, Z. M. Mao, and J. Pang. AccuLoc: Practical localization of performance measurements in 3G networks. In *ACM MobiSys*, 2011.

[19] P. Zerfos, X. Meng, and S. H. Wong. A study of the short message service of a nationwide cellular network. In *ACM SIGCOMM IMC*, 2006.

[20] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch global, cache local: YouTube network traffic at a campus network – measurements and implications. In *Annual Multimedia Computing and Networking Conf*, 2008.