Understanding Traffic Dynamics in Cellular Data Networks

Utpal Paul*, Anand Prabhu Subramanian†, Milind Madhav Buddhikot‡, Samir R. Das*
*Computer Science Department, Stony Brook University, Stony Brook, NY 11794-4400, U.S.A.
† Alcatel-Lucent USA Inc., 600 Mountain Avenue, Murray Hill, NJ 07974-0636, U.S.A.
‡ Alcatel-Lucent Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974-0636, U.S.A.

***Abstract*—We conduct the first detailed measurement analysis of network resource usage and subscriber behavior using a large-scale data set collected inside a nationwide 3G cellular data network. The data set tracks close to a million subscribers over thousands of base stations. We analyze individual subscriber behaviors and observe a significant variation in network usage among subscribers. We characterize subscriber mobility and temporal activity patterns and identify their relation to traffic volume. We then investigate how efficiently radio resources are used by different subscribers as well as by different applications. We also analyze the network traffic from the point of view of the base stations and find significant temporal and spatial variations in different parts of the network, while the aggregated behavior appears predictable. Broadly, our observations deliver important insights into network-wide resource usage. We describe implications in pricing, protocol design and resource and spectrum management.**

I. Introduction

Broadband cellular networks are emerging to be the most common means for mobile data access world-wide. Specifically, 3G networks such as WCDMA based HSPA (High Speed Packet Access) networks and CDMA based EVDO (Evolution-Data Optimized) networks are quite common. The popularity of broadband cellular networks is also fueled by the introduction of user-friendly smart phones, netbooks and tablet devices with a plethora of innovative mobile applications. Expectation is that the volume of data through cellular data networks will increase exponentially in near future. In order to support such increases, it is important to understand the traffic dynamics and its impact on resource allocation on the service provider's network. This will lead to better resource planning and network designs that finally benefit the end users.

There have been several works in the past that study spectrum usage and application characteristics in cellular data networks (see, e.g, [1], [2], [3]). Most of these prior studies try to understand wireless spectrum usage and characterize network performance and capacity using small scale measurements using a few mobile clients. To understand the network usage pattern and subscriber[1] behavior, a large scale comprehensive measurement and analysis of network-wide data traffic must be performed. Though there have been a few studies recently based on network-wide data collected 'in-network' such as [4], [5], these studies consider voice traffic [4] or users' browsing behavior [5]. A detailed network-wide study of data traffic is still lacking.

Our focus in this paper is to address this limitation and provide a measurement-driven analysis of the data traffic collected *at the core* of a nation-wide 3G network. Our goal is to provide answers to important questions regarding subscriber traffic patterns, subscriber mobility, and spatio-temporal behavior of network resource usage. Our data set spans one week in 2007 and consists of all data traffic associated with the entire subscriber base (in the order of hundreds of thousands) in a nation-wide network with thousands of base stations. All generated data packet headers (but not including user payloads) and various signalling and accounting packets are captured, archived and later post processed using a tool we have developed.[2]

In our study, we focus on the spatial and temporal dynamics of data traffic from both the subscriber's (Section II) and the network's (Section III) perspectives. We examine individual subscriber behavior and usage patterns. We also characterize subscriber mobility and temporal activity patterns, and analyze their relationships to subscriber traffic. From the network's perspective, we study traffic patterns at different parts of the network (base stations) and understand spatial and temporal dynamics. Finally, we describe the implications of our observations related to traffic spread, mobility and efficiency in connection to subscriber pricing, protocol design, spectrum allocation and energy savings (Section IV).

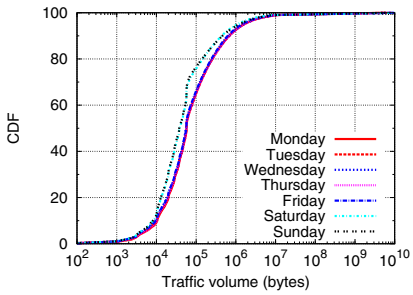
II. Subscriber Traffic Dynamics

We study the behavior of mobile subscribers in terms of the traffic they generate, their mobility and their activity on the temporal scale. We draw relations between traffic generated by subscribers to their mobility as well as activity level. Finally, we present important implications of subscriber traffic dynamics on resource planning in cellular data networks.

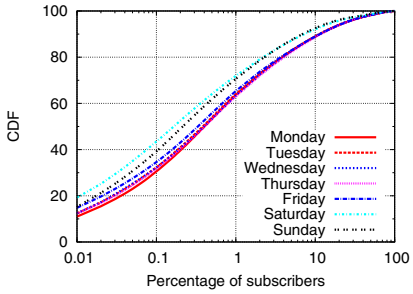
This work was supported in part by National Science Foundation (NSF) Grants CNS0831762 and CNS0831791.

[1]The terms 'client', 'user' and 'subscriber' are used interchangeably in this paper.

[2]For proprietary reasons, we are unable to provide further details about the nature of the 3G network, network location, data set, packet capture and post-processing techniques. This is not unusual in recent published network-wide studies [4]. In any case, the missing details are not relevant to understanding our analysis for commercially operated networks.



(a)



(b)

Fig. 1. (a) CDF of total traffic volume (in bytes) per subscriber per day. (b) CDF of normalized traffic over the percentage of subscriber per day.

A. Subscriber Traffic Distribution

We start with analyzing the amount of traffic generated by subscribers in the network. Figure 1(a) shows the cumulative distribution function (CDF) of traffic generated per subscriber. Each curve represents data for one day in a week. This figure shows a wide range of traffic generated by different subscribers in the network. The median traffic generated is close to 100 KB per day. However, there are heavy users who generate as high as 10 GB per day (10^5 $\times$ median) as well as light users generating less than 1 KB per day. We see the CDF slightly shifted towards the left for weekends (Saturday and Sunday) indicating less traffic relative to working days. We also present a normalized view of traffic over a percentage of subscribers in Figure 1 (b). It is interesting to see that only 1% of the subscribers (out of approximately about 500K unique subscribers who appear in each day) create more than 60% of the daily network traffic and less than 10% of the subscribers create 90% of the daily network traffic. *This points to a significant imbalance of network usage among subscribers with few subscribers hogging the much of the network resource.* Later, we will pay specific attention to mobility and network activity of these subscribers.

B. Subscriber Mobility

In our data set, we do not have access to precise location of subscribers. Our captured data set also does not have signal strength related information (as we capture packets at the IP layer in the core network) to help in radio localization. However, the signaling packets we capture provide enough information for us to track base station and the cell sector the mobile is associated to at all time instants. This provides us with a rich data set to study subscriber mobility based on the timestamped sequence of the base station he/she is connected to. We have this data in all times instants regardless of whether the subscriber is actually communicating. In this aspect, our data set is far richer than that used in some related literature [6], [4].

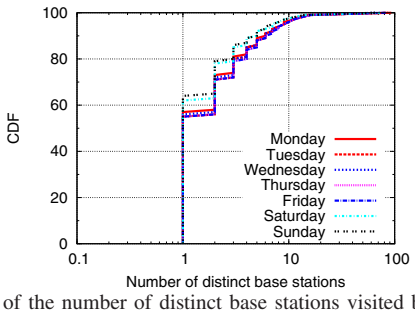


Fig. 2. CDF of the number of distinct base stations visited by a subscriber in each day.

1) Base Stations Visited: Figure 2 shows the CDF of the number of distinct base stations visited by each subscriber in a day. Note that the distribution is very similar in the weekdays, while the distribution in the weekends is somewhat different. Note the tendency of a lesser degree of mobility on weekends. *Overall, the mobility is low in terms of the number of distinct base stations visited.* Roughly, 60% of the users are mostly stationary (i.e., constrained within a cell) and over 95% of the users travel across less than 10 base stations in a day. On the other hand, the highest number of distinct base stations visited by a user in a day is 93. However, such highly mobile users who visit more than 50 distinct base stations in a day are very few, less than 0.01% of daily users. To understand the mobility of subscribers further, we study the extent of the distance they travel next.

2) Radius of Gyration: The above data only captures the number of base stations visited, but not the physical extent of travel. To capture physical distance traveled we use a concept called the *radius of gyration* [6]. The radius of gyration is the linear size occupied by a subscriber's trajectory. It is computed by averaging the displacement of the recorded locations of the subscriber from a central point. The central point is the center of mass of the entire trajectory. Note that this captures how widely the subscribers move as opposed to the actual distance traveled. For example, traveling in a circle continuously visiting the same sequence of base stations does not increase the radius of gyration but a long distance travel on a straight line does. Radius of gyration has traditionally been used to study human mobility, as in a recent influential study [6].

The radius of gyration [6] is defined as,

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\overrightarrow{r}_i - \overrightarrow{r}_{cm})^2},$$

where $\overrightarrow{r}_i$ represents the $i = 1, 2, \ldots, n$ locations recorded for a given user describing his/her trajectory. Recall that the locations are simply the locations of the base stations to which

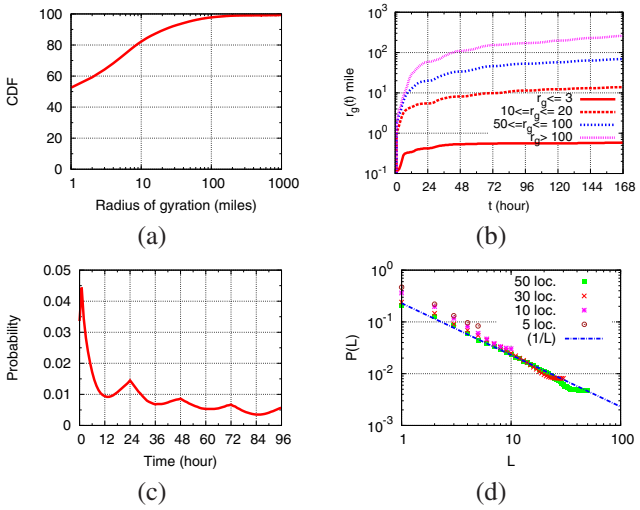


Fig. 3. (a) CDF of radius of gyration (r_g). (b) Radius of gyration versus duration of computation for subscribers categorized into 4 groups according to their final r_g at the end of the seven-day period. (c) Probability distribution of time to returning to the same location. (d) A Zipf distribution showing the probability of finding a subscriber at different locations that are ranked on the basis of their visit frequencies. The subscribers are categorized in terms of how many distinct locations they visit during the seven-day period.

the mobile is connected. $\overrightarrow{r}_{cm} = \frac{1}{n}\sum_{i=1}^{n}\overrightarrow{r}_i$ is the center of mass point of the user's trajectory.

Figure 3(a) shows the CDF of r_g, where r_g is calculated for each subscriber for the entire 7 day time period. We see approximately 53% of subscribers are practically static and almost 98% of subscribers have radius of gyration less than 100 miles. This reasserts the *low level of mobility for the majority of subscribers.* The probability distribution function of subscriber mobility represented by the radius of gyration can be well approximated with a truncated power-law :

$$P(r_g) = (r_g + r_g^0)^{-\beta_r} \exp(-r_g/\kappa),$$

with $r_g^0 = 2.8$ mile, $\beta_r = 1.7$ and $\kappa = 170$ mile. We note that a similar qualitative trend was observed in [6].

Note that the radius of gyration computation requires use of certain duration of time (t) during which the subscriber trajectory is used for the computation. It is expected that the longer the duration t the larger is the radius of gyration $r_g(t)$. A saturation would indicate that some sort of boundary of the movement area has been reached. To study this, we plot 'average' $r_g(t)$ with increasing t until the entire seven-day period (168 hours) is exhausted [6]. Subscribers are categorized into four different groups based on their final r_g value at the end of the seven-day period. See Figure 3(b). Note that the radius of gyration on average comes to a saturation point relatively quickly, in just a few days. Also, users with larger radius of gyration need longer time to saturate.

We further investigate the reason for the quick saturation of the radius of gyration by measuring the 'return probability' for each subscriber as the probability that a subscriber returns after t hours to the same position [6]. Figure 3 (c) shows the distribution having relative peaks at 24th, 48th and 72nd hours.

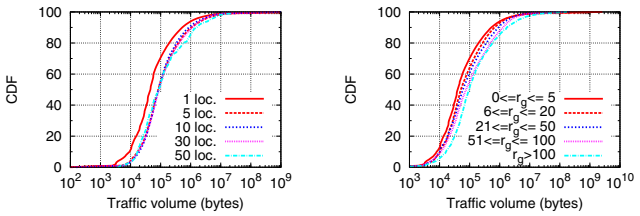


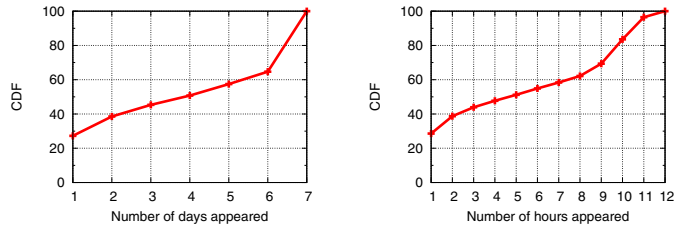
Fig. 4. (a) CDF of traffic generated per day by subscribers of different category based on number of locations (base stations) visited in a day. (b) CDF of traffic generated per day by subscribers of different category based on radius of gyration.

It indicates the periodic nature of human mobility with a 24-hour period and tendency of returning to the same location periodically. This is also the inherent reason for radius of gyration saturating after a few days.

To understand how predictable the subscriber location is, we rank each location a subscriber visits on the basis of the number of times he/she is found there [6]. For example, a location with rank $L = 1$ indicates the most-visited location of the selected subscriber. For each subscriber we create the list of locations where he/she is found in the ascending order of the rank. Figure 3 (d) is the Zipf distribution showing the probability distribution of the visit frequency of locations ranked L. The figure shows the results for four categories of subscribers that visit 5, 10, 30 or 50 distinct locations. It also shows that the distribution can be well approximated by $\sim \frac{1}{L}$ irrespective of the category. Note also that people spend roughly 30% of their time in their top two preferred locations. This clearly shows that *even when subscribers move between multiple locations, they can be found in their 'favorite' location with high probability.*

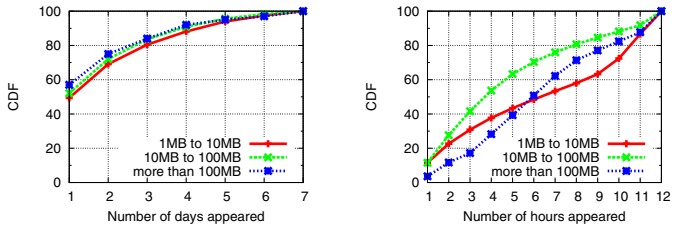
C. Relating Subscriber Mobility and Traffic

Now, a natural question is to relate the subscriber mobility and the volume of traffic they generate. We categorize subscribers based on the two mobility metrics used in the previous section: (i) number of locations (base stations) visited and (ii) radius of gyration. This simply categorizes subscribers based on their degree of mobility. For each category of subscribers, we plot the CDF of traffic volume generated per day in Figure 4. A careful reader will note that while the plot lines appear similar, due to the log-scale of the horizontal axis, there is actually significant difference in traffic volume for different categories. *The trend is that more mobile subscribers generate more traffic, with the median traffic generated by subscribers in the highest mobility category being roughly twice that of the subscribers in the lowest mobility category.* This correlation of mobility and traffic has implications in resource planning and spectrum management. While our current results describe only aggregated behavior, our future work will consider finer grain behavior based on timings of movement and timings of traffic generated.



(a) (b)

Fig. 5. (a) CDF of number of days in a week subscribers generate traffic. (b) CDF of number of hours among peak hours (8 AM to 8 PM) subscribers generate traffic.



(a) (b)

Fig. 7. (a) CDF of occurrence for the heavy users in days in a week. (b) CDF of occurrence for the heavy users in hours among peak hours.

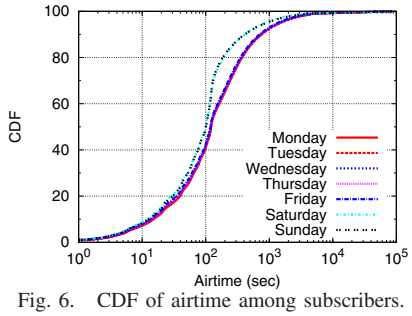


Fig. 6. CDF of airtime among subscribers.

D. Subscriber Temporal Activity

We describe the temporal activity of subscribers by the number of days in a week or number of hours in a day that they generate traffic. This addresses basic questions such as whether the subscribers generate traffic frequently or only occasionally. Figure 5(a) shows the CDF of the number of days subscribers generate traffic. We see that about 34% of the subscribers generate traffic on all 7 days of a week. It is interesting to note that about 45% of total number of subscribers generate traffic only on three or less number of days in a week. To understand the hourly activity of subscribers, we plot the distribution of hours among peak hours (8 AM to 8 PM) in a work day (i.e., Mon-Fri) the subscribers generate traffic. Figure 5(b) shows about 28% of subscribers generate traffic only in a single hour among this set of peak hours. A typical subscriber (median) is active in 4 different hours during the peak hours in a day. The high level conclusion here is that *a large fraction of subscribers generate traffic only in few days a week and only in a few hours within the day.*

To understand the temporal activity of subscribers at a much finer granularity, we study the distribution of 'airtime' used by each subscriber. This term requires some explanation. In the commonly used 3G standards (3GPP or 3GPP2), a subscriber requests and is in turn allocated a radio channel[3] whenever it has data to send. The allocated radio channel is revoked by the network when the subscriber is dormant for certain period known as the dormancy period (typically about 10 seconds) [7] that is configurable for different networks. A subscriber can go between active (with a channel allocated) and dormant state multiple times within a single mobile IP session. We refer the amount of time a subscriber holds onto a radio channel (regardless of whether it actually communicates) as the *airtime*. Effectively, the airtime gives us the amount of time a subscriber uses radio and spectrum resources.

Figure 6 shows the CDF of airtime among all subscribers. We see a significant variation in the amount of airtime used by different subscribers. The median usage is about 100 sec. in a day. Interestingly, there are few subscribers (less than 1%) that use almost 24 hours of airtime in a day. About 90% of subscribers use less than 1000 sec. of airtime. The median is about 100 sec. Weekend usage is typically lower compared to weekday usage. *In general, we see that a typical subscriber occupies the radio channel only for a short duration in the entire day. This is consistent with our previous observation that the median traffic volume per subscriber per day is not significant while there are a small number of 'heavy hitters' that consume a significant amount of network resource.* Such statistics can help providers develop effective pricing structures.

E. Relating Subscriber Activity and Traffic

In this section, we draw relation between the traffic generated by subscribers and how frequently they appear in the trace. We particularly focus on the 'heavy users' as they are the ones that transmit bulk of the traffic. Here, the heavy users are the subset of subscribers that are within the top 5000 in at least one day in the week based on the traffic volume. Recall from Section II-A that about 1% of subscribers send about 60% of traffic. The number 5000 forms roughly 1% of the number of subscribers that generate traffic in a typical day.

Figure 7 (a) shows the number of days these heavy users generate traffic. Interestingly, we see that almost 50-60% of the heavy users generate traffic only on one day in the entire week. This result is different from the percentage of subscribers (about 28%) generating traffic only on one day considering all subscribers as shown in II-D. *This shows that most heavy users are not habitual, but actually quite sporadic.* In Figure 7 (b), we show the distribution of hourly activity of heavy user during peak hours (8 AM to 8 PM). This plot shows that a typical heavy user appear in 4 to 6 different hours during the peak hours in the days they generate traffic. This distribution is not significantly different from the distribution of the entire set of subscribers.

[3]We use the term 'radio channel' to refer to any radio resource allocated to the mobile such as code, frequency or time slot.

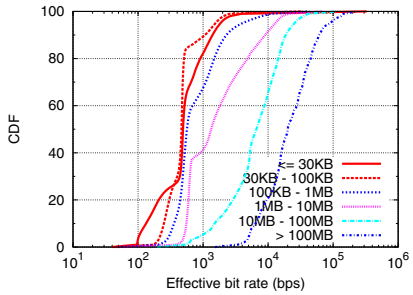


Fig. 8. CDF of effective bit rate for subscribers categorized by traffic generated per day.

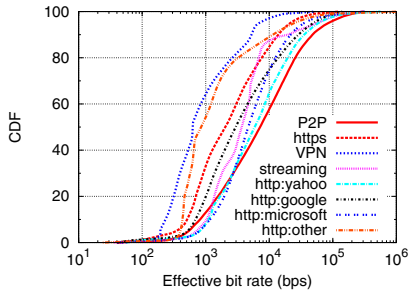


Fig. 9. Effective bit rate for popular applications.

It is also interesting to look at how efficiently subscribers use radio resources, and whether there is any difference between the low and high volume users. To do this, we define a metric called 'effective bit rate'. This is the ratio between the amount of traffic generated by subscribers to the airtime (time actually occupying the radio channel irrespective of traffic generated) used by them. This metric tells us how efficiently the allocated radio channel is used for sending traffic. Figure 8 shows the CDF of effective bit rate with subscribers categorized based on the amount of daily traffic they generate. We can clearly see that *subscribers generating less traffic have progressively poorer effective bit rate.* This may be due to the applications used by subscribers not fully utilizing the allocated channel bandwidth. Even the effective bit rate of a typical high volume subscriber ($\geq$ 100 MB) is approximately 20 Kbps which is much less compared to the maximum nominal bit rate that could be supported. For a low volume subscriber, it is roughly 0.5 Kbps.

To investigate the reason for poor efficiency, we identify the most popular applications (that account for 75% of total daily traffic among all subscribers) and study their channel efficiencies. See Figure 9. Only TCP based applications are considered so that the flow start and stop instants can be clearly identified. During the lifetime of each flow the number of bits transmitted and the total airtime consumed are used to compute the effective bit rate. Port numbers in the packet headers are used to identify the application type. For http we also track the server IP addresses to identify the sites visited. Statistics for a few popular sites (google, microsoft and yahoo) are shown separately.

Note that applications like VPN, https (used for secure connection) and http (for sites other than the popular ones such as google, microsoft and yahoo) have the poorest efficiency, while P2P and http for certain popular sites (yahoo) have the best efficiency. The median difference between VPN and P2P is over an order of magnitude (note the log nature of the horizontal axis). Broadly, it appears that *the enterprise applications generate much less traffic compared to other applications for the same airtime consumed.* The likely reason for this is that such applications tend to use the network sporadically (e.g., frequent use of keep-alive messages in VPN) and/or typically are not high throughput applications. Considering the nature of the dormancy period in 3G networks it it easy to see that channel usage will be inefficient in such applications. On the other hand, high throughput applications like P2P downloads or http browsing on certain popular sites tend to send more data during their allocated airtime. Overall again, all applications have significantly poorer effective bit rate compared to nominal bit rates of the underlying physical layer technology, implying significant scope of protocol improvements across layers. More will be discussed on this in Section IV.

III. Base Station Traffic Dynamics

In this section we turn our attention to the network behavior as a whole or in terms of network components (base stations) instead of focusing on subscribers.

A. Aggregate Load

First, we characterize the aggregate load in the entire network considered. Figure 10(a) presents the total traffic split into upload and download for each day of the week. As expected, weekends see a lesser load. Also, downloads dominate relative to uploads with more than 75% of daily load coming from download traffic. We also break down the traffic load on the network in a single day into 4 hour periods, as shown in Figure 10(b). We can see that the load on the network is relatively low in the early morning hours, and roughly similar during the day and the evening.

B. Base Station Load Distribution

Next, we analyze the volume of daily traffic load for each base station. Figure 11(a) shows the CDF of the daily load (in bytes) of each base station for each day. It shows that roughly about 80% of the base stations are loaded in the range of 1-100MB per day and 10% of the base stations are highly loaded (more than 100MB per day). Figure 11(b) shows the CDF of daily base station loads normalized by the total network load. It shows that *10% of the base stations experience roughly about 50-60% of the aggregate traffic load.* In both cases, weekend behavior is slightly different than weekday behavior. The load imbalance seems more pronounced in weekends. This great imbalance of the base station loads indicates that a more careful cell planning is possibly needed. Network providers may use smaller cells or microcells at the hotspots to even out the imbalance.

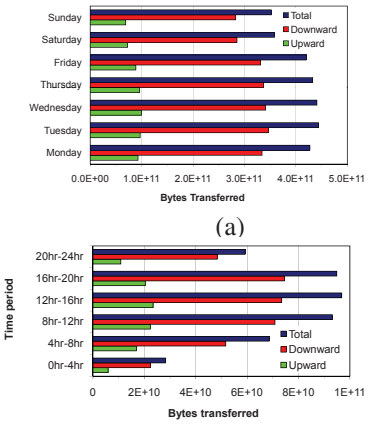


Fig. 10. (a) Aggregate load on the network on each day of the week. (b) Breakdown of total load in a single day in 4 hour periods.

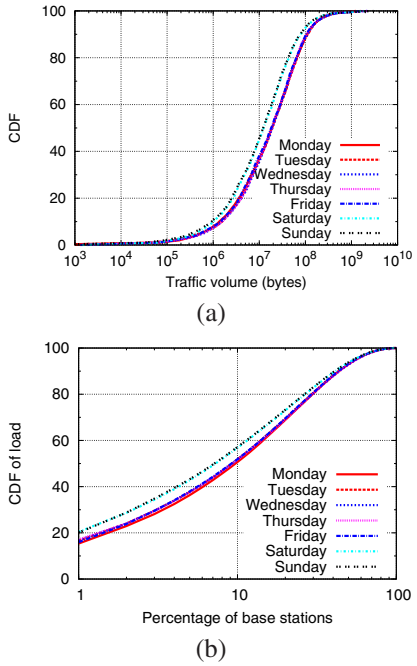


Fig. 11. (a) CDF of actual daily traffic (in bytes) per base station. (b) CDF of normalized traffic per base station.

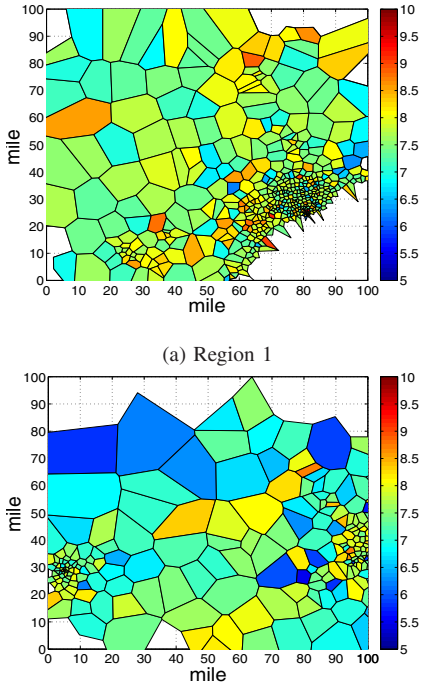


Fig. 12. Total load of each cell in a typical day in two geographically separated regions. The partition in terms of Voronoi cells defines the coverage of each base station. The color bar on the right hand side of each figure indicate to the total load per cell in bytes in $\log_{10}$ scale.

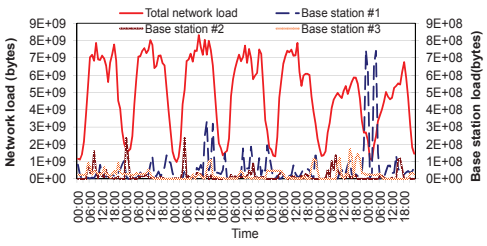


Fig. 13. Aggregated network load in each hour and hourly load of three top loaded base stations. Note that they use two different scales.

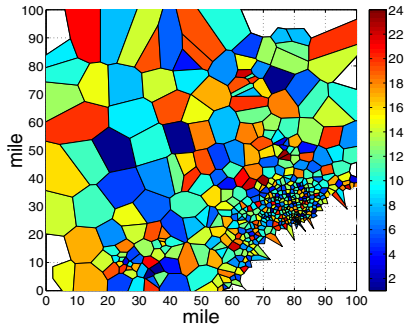
C. Spatial Characteristics

Our main goal here is to identify whether or how much spatially correlated the network load is. Such estimates can potentially help the provider to allocate resources appropriately. This can also be helpful in predicting the load of a spatially separated region given the load of another region.

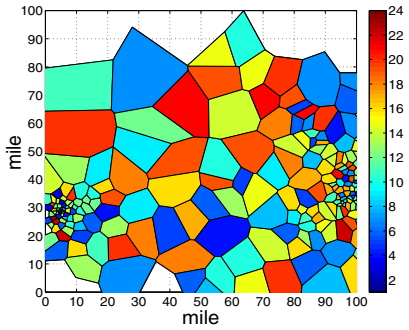
We did preliminary tests and data exploration to investigate the spatial charateristics of network load using Voronoi cells. Each Voronoi cell corresponds to the geographic region of each base station's coverage. Figure 12 shows the aggregate load in bytes for each cell in a typical day in $\log_{10}$ scale for two geographically separated regions in the studied network. Each region is 100 mile $\times$ 100 mile and includes major city centers as well as suburban areas. Note higher density of Voronoi cells in certain areas (city centers) signifying some degree of cell planning. We can readily see again that *the cells are not uniformly loaded in space. The load differentials can extend several orders of magnitude.* There does appear to be some degree of negative correlation between the Voronoi cell size and load. This is expected as large Voronoi cells mean sparsely located base stations, implying sparer population density. No significant spatial correlation between adjacent cells is observed via visual inspection of similar plots for all days.

D. Temporal Characteristics

1) Load: We summarize the hourly load of each base station for the whole 7 day period. Figure 13 shows the hourly aggregate load of the entire network and then top three highly loaded base stations. The aggregate network load exhibits a nice periodic behavior with relatively high loads



(a) Region 1



(b) Region 2

Fig. 14. Peak hour of each cell in a single day in two geographically separated regions.

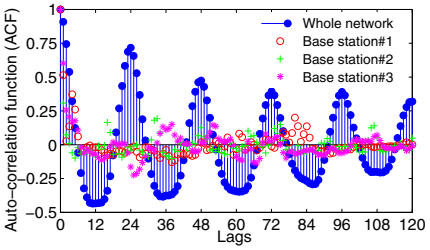


Fig. 15. Auto-correlation Function on the network load in time series.

during the day and the lowest load during midnight. On the contrary, individual base station loads do not show that much periodicity. Also, the load curve varies significantly among individual base stations with their peaks occurring at different times of the day.

We then investigate how the network load varies temporally as well as spatially. To do this, we determine the peak hour of each base station in the day. The peak hour of a base station is the hour in which the given base station has the highest load among its own hourly loads of the day. Figure 14 shows the peak hour of each cell for the same two geographic regions described in Section III-C for a typical day. It shows that base stations have widely different peak hours that without further analysis appear somewhat random. Once again note the lack of spatial correlation.

2) Auto-correlation: For rigorous analysis of the periodic behavior describing the network load we evaluate temporal correlation for a load metric. This will enable us understand the underlying trends and seasonal variations better. We

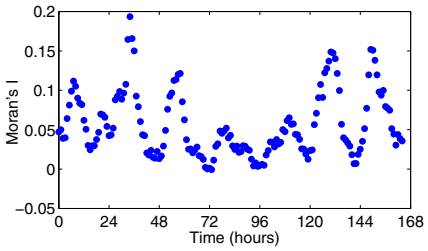


Fig. 16. Moran's I evaluated based on hourly loads on all base stations and plotted on a temporal scale.

represent the hourly aggregate load as time series for the whole network and also for top three base stations used in Section III-D1. Each time series thus has 168 data points for the 7 day period. Figure 15 shows the auto-correlation function (e.g., cross-correlation of the time series with itself) of these time series at different lags. Note that the *plot shows a high degree of temporal correlation.* Again the high peaks occur at 24 hour intervals and low peaks at 12 hour intervals. This is consistent with diurnal human activity patterns. Note that the positive peaks are very pronounced relative to the negative peaks and also the slow decreasing trend of the peaks with increasing lag. The high degree of correlation of network load at the same time of day can have tremendous implication in network resource management techniques. On the other hand, the individual base station loads do not show good temporal correlation (neither positive not negative) and the periodicity is also missing.

E. Spatiotemporal Characteristics

To further follow up on the our observations in Section III-C we use the *Moran's I* statistic [8] but on a temporal scale. Moran's I is a popularly used measure of spatial autocorrelation. It measures how correlated a spatial phenomenon is along space similarly as temporal autocorrelation measures correlation along time. Several earlier works (e.g., [9]) use Moran's I to investigate spatial behavior. A concept of distance is used to indicate proximity and is used as 'weights' in the formula. Moran's I is defined as:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2},$$

where x is the random variable studied, $\overline{x}$ being the sample mean, x_i's are the observations. w_{ij} is the weight associated with each pair (x_i, x_j) and N is the number of observations. Here, the random variable x is the hourly load on a base station. As common with Moran's I studies, we use binary weights: $w_{ij} = 1$, when the base stations are in close proximity (a threshold of 2 miles is used), else $w_{ij} = 0$. We then plot the Moran's I metric for hourly loads of all base stations in the network on a temporal scale. See Figure 16. The plot has been smoothened to remove noise by using a sliding window averaging with a window size of 4 hours. Note that the overall correlation remains small. However, the periodic behavior with a diurnal cycle is interesting. It appears that while temporal usage patterns of base stations may be very different

and might even miss periodicity (Section III-D2), *there is a general tendency for proximate base station loads to be more correlated when the loads are high.* However, the correlation is fairly small, rarely exceeding 0.15. The minimum is often very close to zero, showing almost independent loading behavior around midnights when generally the loads are small.

IV. Summary of Observations and Practical Implications

We now summarize our key observations and identify important practical implications on network planning and protocol design in cellular data networks.

A. Key Observations

1) Traffic Load: There is a significant traffic imbalance both from individual subscriber's and base station's view point. Few subscribers and also few base stations carry a significant fraction of the total load. Less than 10% of subscribers generate 90% of the load, while 10% of base stations carry 50-60% of the load. The subscribers appear to be sporadic users of the network, the heavy users being more so. A typical heavy user only appears occasionally, but generates a large amount of traffic.

2) Mobility: A large fraction of subscribers have limited mobility (roughly half of them being practically static moving within just one mile). The mobility also exhibits periodic behavior with high probability of returning to the same location at the same time of the day. Overall, the mobility is highly predictable. Interestingly, the more mobile subscribers tend to generate more traffic.

3) Efficiency: Effective bit rate is poor due to the intermittency of data transfers and channel dormancy effects. Efficiency is poorer for low volume users relative to high volume users. This could be tied to the types of applications they use. For example, enterprise applications appear to have much poorer effective bit rate relative to P2P.

4) Correlations: Aggregate network load exhibits excellent periodic behavior and temporal correlation, but individual base stations do not exhibit such properties in any significant extent. Spatial correlation among base station loads appear to be small, increasing only when the loads are high (during middays) and remaining almost uncorrelated when the loads are low (around midnights).

B. Implications

1) Subscriber Pricing and Usage Pattern: An unlimited data plan with flat rate pricing is not efficient both from the carrier's perspective as well as the majority of subscribers' perspective. The CDF shown in Figure 1 can be used as a guidance to create 'tiered' rate plans. The idea of tiered rate plans are becoming popular [10] to provide different pricing options based on data usage. One of our future works is to devise optimal pricing schemes based on subscriber usage and available network capacity. Also, sporadic network use by high volume subscribers can create poor experience for other subscribers if such usage occurs during peak periods. This can be alleviated by providing high volume subscribers some incentives (e.g., lower pricing during off-peak hours) to move their load to other times.

2)'Wireless-Friendly' Protocol Design: The highly predictable nature of the mobility pattern can be exploited by innovative cloud-based content delivery applications. The idea is to cache the content of particular interest to a subscriber close to the edge of the network where the subscriber can be found with high probability [11]. This reduces the latency in accessing the content to a large extent. Location based services and targeted ad-services can exploit such highly predictable mobility pattern to optimize their performance. Further, it is clear that the network protocols and applications designed for general wired Internet usage are not very 'wireless-friendly,' using valuable channel air time very poorly. This inefficiency is much higher in enterprise applications. Innovative protocols that make use of the wireless channel more efficiently need to designed. We note that some recent research targeting energy usage addresses a similar issue (see, e.g., [12]).

3) Spectrum Allocation and Energy Savings: The high degree of variability in base station loads has important implication on spectrum allocation and energy saving schemes in the network. New energy saving schemes such as adaptively turning on/off certain carriers or radios in base stations based on the load experienced need to be developed. In Section III-D, we noted that the peak hours of different cells vary a lot, which advocates dynamic allocation of spectrum resources to highly loaded cells during their peak hours. One of our future work will be to model the demand characteristics on different cells in cellular data networks based on measurements for a long period of time and feed the model as inputs to dynamic spectrum allocation algorithms such as [13].

V. Related Work

There have been field measurement studies on 3G networks mainly focusing on the performance of data traffic, but only from the point of view of individual client devices. Representative works in this space are measurement studies on commercial WCDMA 3G/UMTS networks in [1], performance evaluation of GPRS and UMTS networks in [2], various forms of TCP performance evaluation in [14], [3] and [15], and cross layer studies in [16]. In addition, Joyce *et al.* [17] have presented single cell and network capacity measurements using a commercial network in UK. Yao *et al.* [18] have evaluated bandwidth predictability for HSDPA networks. Tan *et al.* [19] have studied the capacity of 3G networks in terms of throughput, latency, video and voice call handling ability. The authors in [20] have evaluated multimedia streaming through measurements taken from real networks (GSM, GPRS and UMTS). Performance of push-to-talk applications have been evaluated in [21] on 3G networks. The above studies do not use the global view of the network as a whole and a broader analysis of the subscriber behaviors are missing.

Such global views have been pursued only in a limited number of papers. The authors in [22] have carried out spectrum measurements in 2G and 3G bands during the 2006 Soccer World Cup in two German cities. They have shown that the change of spectrum usage is related to specific events. In [9], a measurement-based spectrum modeling approach has been developed using spatial statistics and random fields. The authors in [23] have shown the distribution of voice call duration analyzing the call logs from a cellular GSM provider. The authors in [4] have presented a large scale measurement analysis to characterize the primary usage in cellular voice network. In [5] the browsing behavior of mobile users in a large scale 3G data network has been analyzed. In contrast to these papers, our focus is purely on data traffic behavior in the context of resource usage.

Finally, studying human mobility from cellular network data is an important component of our work, as mobility directly impacts resource usage. Much of our analysis has been motivated by Barabasi and co-authors' influential work on this topic [6], [24]. They have studied human mobility patterns based on the voice call records over a six-month period of 100,000 anonymized mobile phone users. They have concluded that human trajectories show a high degree of temporal and spatial regularity. In [25] an analysis of user mobility patterns is presented based on data traffic traces from a major regional CDMA2000 cellular network. The overall mobility was found to be limited. Pathirana *et al.* [26] have presented a technique to predict the trajectory of a user in a variant of GSM network. Authors in [24] have investigated the human dynamics and social interactions, and focused on the occurrence of anomalous events. None of these works, however, directly relate the users' mobility to network access behavior and network usage patterns.

VI. Conclusions

In our knowledge, our work is the first major study in measurement analysis of subscriber and network behavior in a large scale 3G data network. We have made several important observations related to traffic load, mobility and resource efficiency. We have indicated the implications of these observations in pricing, protocol design and resource management. Our future work will target (i) analysis for longer periods of time as well as in finer grain, (ii) addressing the topics highlighted in the discussions about implications in Section IV.

References

[1] K. Pentikousis, M. Palola, M. Jurvansuu, and P. Perl, "Active goodput measurements from a public 3G/UMTS network," *IEEE Communications Letters*, vol. 9, pp. 802–804, 2005.

[2] P. Reichl, M. Umlauft, J. Fabini, R. Lauster, and G. Pospischil, "Project WISQY: A measurement-based end-to-end application-level performance comparison of 2.5G and 3G networks," in *Proc. Fourth Ann. Wireless Telecomm. Symp. (FTS)*, 2005.

[3] K. Mattar, A. Sridharan, H. Zang, I. Matta, and A. Bestavros, "TCP over CDMA2000 networks: A cross-layer measurement study," in *Proc. PAM*, 2007.

[4] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary users in cellular networks: A large-scale measurement study," in *Proc. DySPAN*, 2008.

[5] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao, "Profiling users in a 3G network using hourglass co-clustering," in *Proc. ACM MobiCom*, 2010.

[6] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 2008.

[7] M. Chuah and W. Luo, "Impacts of inactivity timer values on UMTS system capacity," in *Proc. of IEEE Wireless Communications and Networking Conference*, 2002, pp. 897–903.

[8] P. A. P. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, p. 1733, 1950.

[9] J. Riihijärvi, P. Mähönen, M. Wellens, and M. Gordziel, "Characterizing and modelling of spectrum for dynamic spectrum access with spatial statistics and random fields," in *Proc. PIMRC*, 2008.

[10] AT&T Wireless Data Plan Press Release. http://www.att.com/gen/press-room?pid=4800\&cdvn=news\&newsarticleid=30854.

[11] P. Deshpande, A. K. C. Sung, and S. R. Das, "Predictive methods for improved vehicular WiFi access," in *Proc. ACM Mobisys Conference*, 2009, pp. 263–276.

[12] C.-C. Lee, H. Yeh, and J.-C. Chen, "Impact of inactivity timer on energy consumption in WCDMA and CDMA2000," in *Proc. Wireless Telecommunications Symposium*, 2004.

[13] A. P. Subramanian, H. Gupta, S. R. Das, and M. M. Buddhikot, "Fast spectrum allocation in coordinated dynamic spectrum access based cellular networks," in *Proc. DySPAN*, 2007.

[14] Y. Le, "Measured TCP performance in CDMA 1x EV-DO networks," in *Proc. PAM*, 2005.

[15] M. Kohlwes, J. Riihijärvi, and P. Mähönen, "Measurements of TCP performance over UMTS networks in near-ideal conditions," in *Proc. VTC 2005-Spring*, 2005.

[16] A. S. X. Liu, S. Machiraju, M. Seshadri, and H. Zang, "Experiences in a 3G network: Interplay between the wireless channel and applications," in *Proc. ACM MobiCom*, 2008, pp. 211–222.

[17] R. Joyce, B. Graves, T. Griparis, I. Osborne, and T. Lee, "Case study: The capacity of a WCDMA network-Orange UK," in *Proc. 3G Mobile Communication Technologies*, 2004.

[18] J. Yao, S. Kanhere, and M. Hassan, "An empirical study of bandwidth predictability in mobile computing," in *Proc. ACM WiNTECH*, 2008, pp. 11–18.

[19] W. Tan, F. Lam, and W. Lau, "An empirical study on the capacity and performance of 3G networks," *IEEE Transactions on Mobile Computing*, vol. 7, pp. 737–750, 2008.

[20] J. Chesterfield, R. Chakravorty, J. Crowcroft, P. Rodriguez, and S. Banerjee, "Experiences with multimedia streaming over 2.5G and 3G networks," in *Proc. BROADNETS*, 2004.

[21] W. Chen, S. Licking, T. Ohno, S. Okuyama, and T. Hamada, "Performance measurement, evaluation and analysis of Push-to-Talk in 3G networks," in *Proc. IEEE International Conference on Communications*, 2007.

[22] O. Holland, P. Cordier, M. Muck, L. Mazet, C. Klock, and T. Renk, "Spectrum power measurements in 2G and 3G cellular phone bands during the 2006 Football World Cup in Germany," in *Proc. DySPAN*, 2007.

[23] J. Guo, F. Liu, and Z. Zhu, "Estimate the call duration distribution parameters in GSM system based on K-L divergence method," in *Proc. WiCom*, 2007.

[24] J. Candia, M. Gonzalez, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabasi, "Uncovering individual and collective human dynamics from mobile phone records," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, pp. 1–11, 2008.

[25] E. Halepovic and C. Williamson, "Characterizing and modeling user mobility in a cellular data network," in *Proc. Workshop on PE-WASUN*, 2005.

[26] P. Pathirana, A. Savkin, and S. Jha, "Mobility modelling and trajectory prediction for cellular networks with mobile base stations," in *Proc. ACM MobiHoc*, 2003, pp. 213–221.