Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Network

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

***Abstract*—Recent studies on cellular network measurement have provided the evidence that significant geospatial correlations, in terms of traffic volume and application access, exist in cellular network usage. Such geospatial correlation patterns provide local optimization opportunities to cellular network operators for handling the explosive growth in the traffic volume observed in recent years. To the best of our knowledge, in this paper, we provide the first fine-grained characterization of the geospatial dynamics of application usage in a 3G cellular data network. Our analysis is based on two simultaneously collected traces from the radio access network (containing location records) and the core network (containing traffic records) of a tier-1 cellular network in the United States. To better understand the application usage in our data, we first cluster cell locations based on their application distributions and then study the geospatial dynamics of application usage across different geographical regions. The results of our measurement study present cellular network operators with fine-grained insights that can be leveraged to tune network parameter settings.**

I. Introduction

A. Background and Problem Statement

Cellular network operators have globally observed an explosive increase in the volume of data traffic in recent years. Cisco has reported that the volume of global cellular data traffic has tripled (year-over-year) for three years in a row, reaching up to 237 petabytes per month in 2010 [1]. This unprecedented increase in the volume of cellular data traffic is attributed to the increase in the subscriber base, improving network connection speeds, and improving hardware and software capabilities of modern smartphones. In contrast to the traditional wired networks, cellular network operators are faced with the constraint of limited radio frequency spectrum at their disposal. As the communication technologies evolve beyond 3G to long term evolution (LTE), the competition for the limited radio frequency spectrum is becoming even more intense. Therefore, cellular network operators are increasingly focusing on optimizing different aspects of the network by customized design and management to improve key performance indicators (KPIs).

Two important aspects of a cellular network that present significant optimization potential to the network operators are: (1) *diverse application mix constituting the data traffic* and (2) *variations in the traffic depending upon the geo-location of users*. It has been shown that the performance of different applications constituting the data traffic in cellular networks is sensitive to various network KPIs [8], [13]. Tso *et al.* also showed that the network performance perceived by users is strongly related to their geolocation and mobility patterns [19]. Combining the above-mentioned two aspects, cellular network operators can potentially find even better opportunities for network optimization. However, to the best of our knowledge, no prior work has jointly studied the relationship between application usage and users' geospatial movement patterns.

B. Limitations of Prior Art

Trestian *et al.* conducted a study that provided the first evidence of geographic correlation of users' "interests" in a cellular network [18]. They showed that users in different geographical regions have different interests; for example, people mostly access mail URLs from office locations and access more music URLs from residential locations. However, cellular network operators not only need to know that there is geographic correlation of interests, but also how those interests translate into different types of application traffic. This is because it is the type of traffic (bursty, bulk transfers, streaming, *etc.*) that determines how an operator can best optimize each geographic area. Furthermore, cellular network operators would like to be able to map the above-mentioned coarse-grained geographic correlation to a more fine-grained cell sector correlation, as this is typically the smallest unit that operators can configure. Paul *et al.* separately studied application usage and geospatial patterns of aggregate traffic volume; however, they did not study correlation between them [13]. Other prior studies that either study application usage or geospatial patterns (but not both simultaneously) include but are not limited to [5], [8], [12], [16], [19], [21]. Further details of prior art are provided in Section V.

C. Major Contributions

To the best of our knowledge, this paper presents the first fine-grained characterization of the geospatial dynamics of application usage in a 3G cellular data network. We summarize the key contributions our research as follows:

1) **Data Collection:** For our study, we collected two traces from the cellular network: (1) periodically collected cell sector records of devices from the radio network and (2) data traffic records of IP flows passing through the core network. Due to the massive size of the collected traces, our data set is limited to 32 hours worth of data

in December 2010 covering a large metropolitan area spanning more than $1,200$ km^2 in the United States.

2) **Methodology:** We study application usage characteristics of users across more than two thousand 3G cell locations. For systematic analysis of application usage across these cell locations, we first cluster cells based on their application distribution. The results of our clustering experiments show that cells can be robustly categorized into a small number of clusters using traffic volume in terms of byte, packet, flow, and unique user count distributions. Using the clustering results, we analyze the geospatial patterns of application usage across different geographical regions, *e.g.* downtown, university, and suburban areas. To extract geospatial dependence patterns, we utilize basic cluster composition analysis and intensity function analysis in this paper.
3) **Findings and Implications:** The results of our geospatial analysis experiments reveal new insights that have important implications for network optimization. A major finding of our measurement study is that cell clustering results are significantly different for traffic volume in terms of byte, packet, flow count, and unique user count distributions across different geographical regions. These results present operators with an opportunity to fine-tune network parameter settings for different applications. However, they also suggest that operators should not optimize cells solely by traffic volume in terms of byte, packet, or flow counts because this may negatively impact the performance of other low volume–but popular–applications. Furthermore, we find that there is differentiation between the application mix of different cells even *within* a close neighborhood such as a university, downtown, or suburb. Consequently, there are opportunities for fine-grained network optimization within close neighborhoods.

Paper Organization: The rest of the paper proceeds as follows. We describe the details of our collected data set in Section II. In Section III, we provide the results of our measurement analysis for characterizing geospatial dynamics of application usage in cellular networks. We summarize the major findings of our study in Section IV and also provide their implications on cellular network optimization. We provide an overview of the related work in Section V. Finally, we conclude the paper in Section VI with an outlook to our future work.

II. Background and Data Sets

In this section, we first provide a brief overview of 3G Universal Mobile Telecommunications System (UMTS) cellular data network architecture and then provide information about the data set used in our study.

A. Network Architecture

Figure 1 shows the architecture of a typical 3G UMTS cellular data network. A UMTS cellular data network consists of two separate networks: radio access network and a core

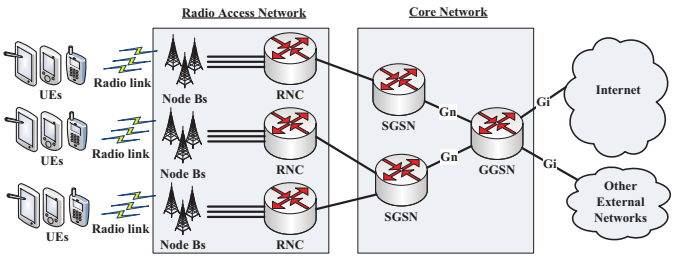


Fig. 1. Architecture of a typical 3G UMTS cellular data network.

network. The network elements in these networks are logically connected to each other in a tree topology. The following list orders the elements from the leaves to the root of the tree: *user equipment* (UE), *cell sectors*, *NodeBs*, *Radio Network Controllers*, *Serving GPRS Support Nodes* (SGSNs), and *Gateway GPRS Support Nodes* (GGSNs). A UE, or cellular device, connects to one or more cell sectors in the radio access network. Each sector is distinguished by a different antenna on a NodeB, or physical base station. The data traffic of a cellular device is passed by the NodeB to a RNC, which manages radio access network control signalling such as transmission scheduling and handovers. Each RNC typically sends and receives traffic to/from several NodeBs that cover hundreds of cell sectors, each of which in turn serves many users in its coverage area. The core network consists of SGSNs facing cellular devices and GGSNs that connect to external networks. RNCs send data traffic to SGSNs, which then send it to GGSNs. Finally, GGSNs send data traffic to external networks, such as the Internet. In order to support mobility without disrupting a cellular device's IP network connections, the IP address of the device is anchored at the GGSN. The IP address association is formed when the device connects to the network and establishes a Packet Data Protocol (PDP) Context which facilitates tunnelling of IP traffic from the device to the GGSN. These tunnels, implemented using the GPRS Tunneling Protocol (GTP), carry IP packets between the cellular devices and their peering GGSNs.

B. Data Sets

In this paper, we use two anonymized data sets from a tier-1 cellular network carrier for our study. The first data set contains flow-level information about IP flows carried in PDP Context tunnels (*i.e.*, all data traffic sent to and from cellular devices). This data set is collected from all *Gn* links between SGSNs and GGSNs in the core network. For a 3% random sample of devices, the data contains the following information for each IP flow per minute: start and end timestamps, per-flow traffic volume in terms of bytes and packets, device identifiers, user identifiers, and application identifiers. All device and user identifiers (*e.g.*, IMEI, IMSI) are anonymized to protect privacy without affecting the usefulness of our analysis. The data sets do not permit the reversal of the anonymization or re-identification of users. For proprietary reasons, the results presented in this paper are normalized. However, normalization does not change the range of the metrics used in this study.

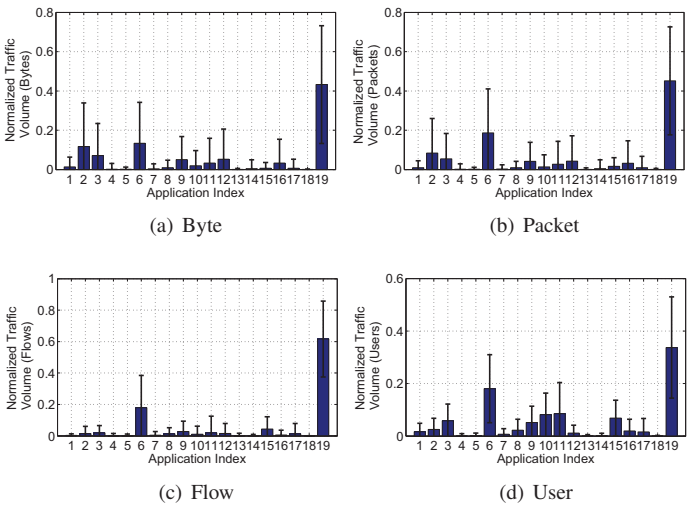
Furthermore, the missing information due to normalization does not affect the understanding of our analysis.

Application identifiers include information about application protocol (*e.g.*, HTTP, DNS, SIP), class (*e.g.*, streaming audio, streaming video, web, email), and, in the case of applications registered in popular "App Stores," the unique name of the application. Applications are identified using a combination of port information, HTTP host and user-agent information, and other heuristics [4]. Since we encounter tens of thousands of applications in the data, we only examine the top 100 by traffic volume. These top applications comprise the vast majority of all data traffic, so understanding the remainder is not critical for the purpose of network engineering [23]. Furthermore, we categorize the top applications into the following 19 application realms by function and traffic type (streaming, interactive, *etc.*): (1) `ads`, (2) `mixed HTTP streaming`, (3) `app store`, (4) `media optimization`, (5) `dating`, (6) `email`, (7) `games`, (8) `news info image media`, (9) `maps`, (10) `misc`, (11) `mms`, (12) `music audio`, (13) `p2p`, (14) `radio audio`, (15) `social network`, (16) `streaming video`, (17) `voip`, (18) `vpn`, (19) `web browsing/other http`.

Although this data set also contains the cell locations associated with each PDP context, these locations are often inaccurate because they are typically only recorded when PDP contexts are established and may not be updated for hours or days even when users are mobile [22]. Therefore, we cannot study fine-grained geospatial dynamics of application usage using the location information collected only from the core network. To get accurate location information, we collect a second data set at RNCs in the radio access network. The second data set contains fine-grained logs of signaling events at the RNCs, which include handover events. By joining the PDP sessions in the first data set with complete handover information in the second data set, we get accurate cell locations at a 2 second granularity for IP flows in the first data set.[1] It is important to note that the second data set cannot be continuously collected over long durations of time because its collection can introduce non-trivial additional overheads at the RNCs.

For this study, we simultaneously collected both data sets over a weekday period of 32 hours in December 2010. The data sets cover a large metropolitan area spanning more than $1,200$ km^2 in the United States. The data sets cover more than two thousand 3G cells, but do not cover any 2.5G cells. It accounts for hundreds of gigabytes of IP traffic, consisting of hundreds of millions of packets and tens of millions of flows, and covers tens of thousands of devices. Although we cannot study long-term application usage patterns due to the significant overheads of collecting the second data set over longer timescales, we believe our results still provide generalizable insights due to the volume of data and number of devices studied.

[1]In practice, a device may be connected to multiple cell sectors at the same time. For the purposes of our study, we use the *primary* or *serving cell*, which is the sector that actually transmits downlink data to HSPA devices.



(a) Byte (b) Packet

(c) Flow (d) User

Fig. 2. Application mix of aggregate traffic for byte, packet, flow, and user distributions. The mapping of application indices is provided in Section II-B.

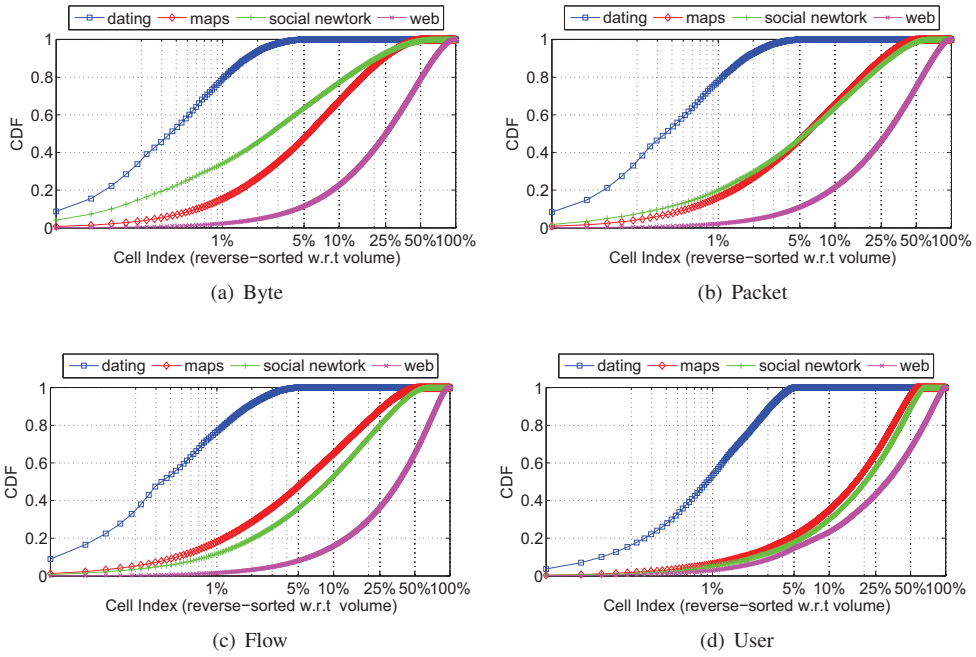
III. Measurement Analysis

In this section, we explain the details of our measurement analysis conducted on the two data sets collected from the cellular networks to study the geospatial dynamics of application usage. Towards this end, we start by examining the application usage distributions in the data traffic and then investigate the relative popularity of individual applications across different cell locations.

As mentioned in Section II, all data traffic records in our data set are tagged with application and cell identifiers. For initial analysis, we first segregate all traffic records with respect to the application identifiers to study the application usage patterns. We then construct application distributions using application identifiers as keys and byte, packet, flow, or user counts as values.[2] Figure 2 shows the byte, packet, flow, and user distributions for the collected data set. We note that application popularity in the complete data set is highly skewed, where `web browsing` and `email` realms dominate with respect to byte, packet, flow, and user counts. We also note some differences in the popularity of applications across byte, packet, flow, and user distributions. Specifically, `maps` and `social network` have higher volume with respect to user counts as compared to byte, packet, and flow counts. This observation shows that these applications are relatively low volume (with respect to byte, packet, and flows) but are accessed by relatively more number of users. This finding will be further highlighted later in our analysis when we cluster application distributions of different cells.

We now study the relative popularity of a given application across different cell locations in our data set. Figure 3 shows the cumulative distribution of traffic volume of `dating`, `maps`, `social network`, and `web browsing` applica-

[2]In the rest of this paper, the terms byte, packet, flow, and user distributions refer to the traffic volume distributions in terms of byte count, packet count, flow count, and unique user count, respectively.



(a) Byte (b) Packet

(c) Flow (d) User

Fig. 3. Distributions of traffic volume with respect to byte, packet, flow, and user counts across all cell sector locations.

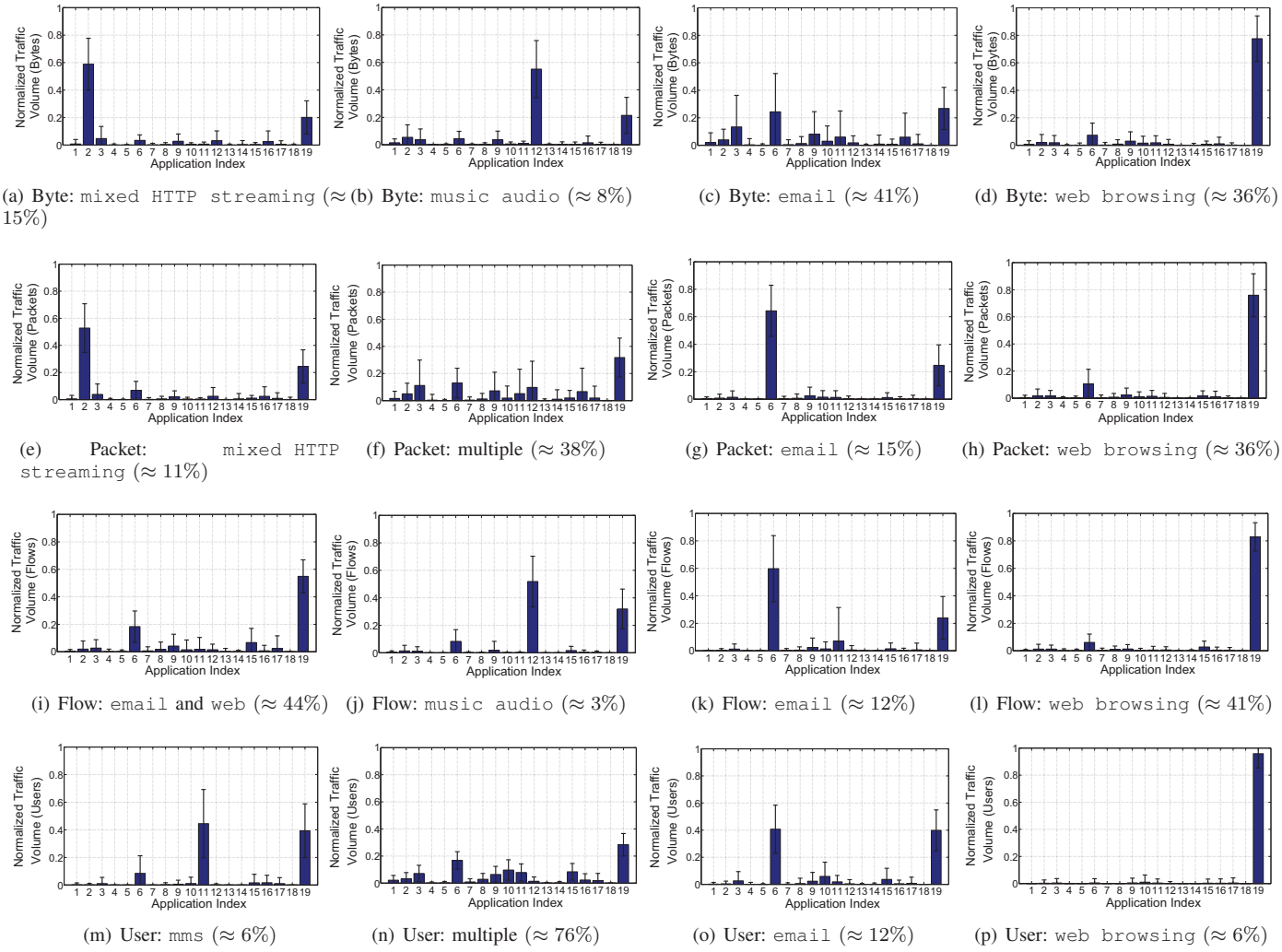
tions with respect to byte, packet, flow, and user counts across all cells in our data set. Our first observation is that applications are not equally popular across all cells in our data set. Furthermore, the popularity of some applications is more skewed than others across cells. For instance, all traffic volume of `dating` application is generated from less than 5% of all cells. On the other hand, `web browsing` is the most ubiquitous application realm. However, even for `web browsing` 80% of the byte traffic volume is generated from 50% of all cells. It is also interesting to note the differences in the byte, packet, flow, and user volume of applications across cells. For instance, the distribution of byte volume of `social network` is more skewed than `maps` across cells; however, this trend is reversed for flow and user volume distributions. This observation indicates that flows and users in a fraction of cells dominate byte volume for `social network` applications.

Until now we have established two major findings: (1) the traffic volume of a few application realms dominate others overall and (2) the popularity of a given application realm varies across different cell locations. These findings suggest strong dependence of application usage on geospatial dynamics. To do more useful fine-grained analysis, in the rest of this section, we first introduce the analytical approaches used for characterizing the geospatial dynamics of application usage in a cellular network. We then present the results of our analysis on our collected data set. We follow a two step methodology to systematically conduct our analysis. First, we group the application usage distributions of cells using an unsupervised clustering algorithm. Second, we conduct a comprehensive analysis of geospatial dynamics of application usage across clusters using geospatial analysis techniques. The goal of our analysis is to identify patterns in our data and to formulate new hypotheses about the underlying processes that gave rise to the data. We now separately discuss the above-mentioned steps in the following text.

A. Cell Clustering

1) Methodology: We segregate all traffic records with respect to the application and cell identifiers to study the application usage patterns for any given cell. Our goal is to cluster cells into a manageable number of groups based on their application usage distributions. It is important to cluster cells by byte, packet, and flow distributions to understand which sectors have similar traffic distributions. But it is also important to understand how cells cluster by user distributions because the applications that are used widely but infrequently by many users will not be well represented relative to the byte, packet, or flow counts of higher volume applications, even if those applications are not as popular. This argument follows our earlier observation in this section from Figure 2.

We utilize a well-known unsupervised clustering algorithm called k-means to cluster application distributions of cells. k-means algorithm is a simple yet effective technique to cluster feature vectors into a predefined k number of groups [10]. The selection of appropriate value of k is crucial and is an open research problem [2]. Several heuristics have been proposed in prior literature, which primarily focus on the change in intra-cluster dissimilarity for increasing values of k [7], [9], [11]. One of the most well-known heuristic, called gap statistic, involves comparing the change in intra-cluster dissimilarity W_k for given data and that for a reference null distribution



(a) Byte: `mixed HTTP streaming` (≈ 15%) (b) Byte: `music audio` (≈ 8%) (c) Byte: `email` (≈ 41%) (d) Byte: `web browsing` (≈ 36%)

(e) Packet: `mixed HTTP streaming` (≈ 11%) (f) Packet: multiple (≈ 38%) (g) Packet: `email` (≈ 15%) (h) Packet: `web browsing` (≈ 36%)

(i) Flow: `email` and `web` (≈ 44%) (j) Flow: `music audio` (≈ 3%) (k) Flow: `email` (≈ 12%) (l) Flow: `web browsing` (≈ 41%)

(m) User: `mms` (≈ 6%) (n) User: multiple (≈ 76%) (o) User: `email` (≈ 12%) (p) User: `web browsing` (≈ 6%)

Fig. 5. Centroids of application distributions of cells identified using k-means clustering. Clustering results (centroids and composition distribution) are separately provided for byte, packet, flow, and user distributions. The mapping of application indices is provided in Section II-B.

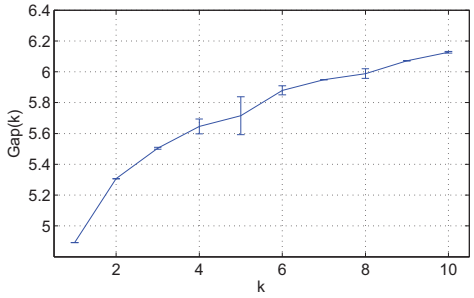


Fig. 4. Gap statistic to find the optimal number of clusters for traffic distributions of cells.

[17]. Gap statistic provides a statistical method to find the elbow of intra-cluster dissimilarity W_k as the values of k is varied. Gap statistic is defined as:

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}) - log(W_k),$$

where W_{kb} denotes the within-cluster dispersion of a reference data set from a uniform distribution over the range of the observed data. Using gap statistic, the optimal value of k is chosen to be the smallest one for which:

$$Gap(k) \geq Gap(k+1) - \sigma_{k+1},$$

where σ denotes the standard deviation of within-cluster dispersions in reference data sets. Figure 4 shows the plot of gap statistic for varying values of k. We observe that $Gap(4) \geq Gap(5) - \sigma_5$, so we select the optimal value of $k = 4$. After selecting the value of $k = 4$ using gap statistic, we apply k-means clustering algorithm to cluster application distributions of cells into four groups.

To gain insights into the clustering results, we plot four cluster centroids of byte, packet, flow, and user distributions in Figure 5. We have labeled the cluster centroids using their popular application types. The cluster centroids that do not have any outright popular application are labeled as `multiple`. In Figure 5, we also provide the percentage

TABLE I
CLUSTER COMPOSITION ANALYSIS RESULTS

Byte (%)				
	mixed HTTP streaming	music audio	email	web browsing
Downtown	12	4	37	47
University	11	11	22	**55**
Suburb 1	19	**17**	39	25
Suburb 2	**29**	0	**42**	29
Packet (%)				
	mixed HTTP streaming	multiple	email	web browsing
Downtown	11	34	7	48
University	11	22	11	**55**
Suburb 1	**14**	**56**	0	31
Suburb 2	7	50	**14**	28
Flow (%)				
	email, web browsing	music audio	email	web browsing
Downtown	42	0	5	**51**
University	33	0	**22**	45
Suburb 1	47	**3**	6	44
Suburb 2	**64**	0	7	28
User (%)				
	mms	multiple	email	web browsing
Downtown	5	74	**15**	5
University	**11**	78	11	0
Suburb 1	8	86	0	**6**
Suburb 2	0	**93**	7	0

distribution of cells across all cluster types. As expected, we observe that web browsing and email are the common cluster centroids for byte, packet, flow, and user distributions. Other cluster centroids include mixed HTTP streaming, music audio, and mms. The plots of cluster centroids in Figure 5 highlight important differences across byte, packet, flow, and user distributions. For instance, we observe that only one or two applications (*e.g.* email, web, and mixed HTTP streaming) make up a predominant percentage of the traffic volume in terms of bytes for a majority of cells. However, the application distributions are relatively even in terms of users for most cells. For example, Figure 5(n) shows that 76% of cells fall into the multiple realm for user distributions, implying that most cells have users that access a diverse set of applications. Whereas, the percentage of cells with relatively balanced application traffic is much lesser for byte, packet, and flow distributions. Another important difference is that the percentage of cells belonging to dominant applications, *e.g.* web browsing and email, significantly vary across byte, packet, flow, and user distributions. For example, only 6% cells belong to web browsing cluster for user distributions; whereas, $\approx 40\%$ cells belong this cluster for byte, packet, and flow distributions. As we discuss later in Section IV, these differences have important implications in terms of cellular network planning and optimization.

B. Geospatial Analysis

Using the clustering methodology defined in the previous subsection, we uniquely label all cell locations for each of the byte, packet, flow, and user application distribution clusters. For geospatial analysis, we apply basic cluster composition analysis and intensity function analysis to the clustering results, which are separately discussed below. To gain interesting insights from the geospatial analysis, we also study different geographical regions, *e.g.* downtown, university, and suburban areas.

1) Cluster Composition Analysis: In cluster composition analysis, we study the distribution of cells belonging to different clusters in various geographical regions. This analysis aims to uncover the cases when cells belonging to a particular cluster type are more prevalent in certain geographical regions.

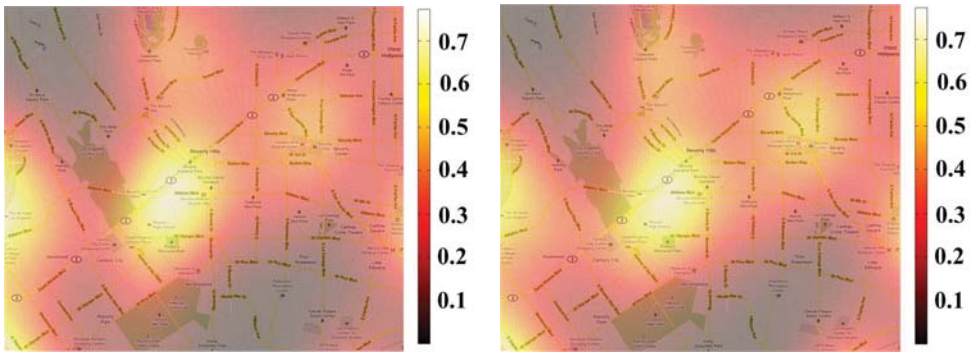
Table I shows the distribution of cells belonging to different clusters across all geographical regions. We observe important differences in application usage across different geographical regions for with respect to byte, packet, and flow distributions. For example, the cells belonging to web browsing cluster are typically less common in suburban areas as compared to downtown and university areas; whereas, the cells belonging to mixed HTTP streaming and music audio clusters are more popular in suburban areas than downtown and university areas. We also note that the cells belonging to mms and email clusters are more popular in the university area. These patterns show that the user interests in cellular data networks are dependent on location and have implications for cellular network optimization as discussed later in Section IV. Table I also indicates that a majority of cells belong to multiple cluster for user count distributions across all geographical regions. For instance, Table I shows that as few as 7% cells belong to clusters with a predominant application with respect to users for suburb 2. Therefore, cellular network operators can only optimize network parameters for specific applications in a minority of cells while satisfying a majority of users.

2) Intensity Function Analysis: The usefulness of basic cluster composition analysis is limited because it does not identify or quantify the patterns within a given geographical region due to its aggregate nature. This limitation of the cluster composition analysis is addressed by the intensity function. Intensity function quantifies the expected number of points (*i.e.* cells belonging to a particular cluster type) per unit area [6]. Intensity function is constant for uniformly distributed points and varies if points are non-uniformly distributed, with peaks in denser regions and troughs in sparse regions. To estimate the continuous intensity function using discrete geographical location information, nonparametric techniques such as Gaussian kernel smoothing are commonly utilized [20]. A typical kernel estimated intensity function takes the form:

$$\tilde{\lambda}(d) = e(d) \sum_{i=1}^{n} \kappa(d - x_i),$$

where $\tilde{\lambda}(d)$ is an unbiased estimator of the true intensity function $\lambda(d)$, $e(d)$ is an edge bias correction, $\kappa(d)$ is the kernel function (isotropic Gaussian kernels are most commonly used), n is the number of points, and d denotes geographical distance.

The intensity functions of web browsing clusters over a suburb area are shown for byte, packet, flow, and user distributions in Figure 6. We can visually observe similarity



(a) Byte (b) Packet

(c) Flow (d) User

Fig. 6. Kernel estimated intensity function for `web browsing` cluster types in a suburban region for byte, packet, flow, and user distributions.

among the intensity functions for byte, packet, and flow distributions; whereas, the intensity function for user distribution is significantly different than the rest. To quantify this similarity, we compute the pair-wise Pearson product-moment correlation coefficient (denoted by $\rho, |\rho| \in [0, 1]$) between two intensity functions [15]. The magnitude of one signifies perfect correlation and zero signifies no correlation at all between the two given intensity functions. Pearson product-moment correlation coefficient is defined as:

$$\rho_{\tilde{\lambda}_1,\tilde{\lambda}_2} = \frac{E[(\tilde{\lambda}_1 - \mu_{\tilde{\lambda}_1})(\tilde{\lambda}_2 - \mu_{\tilde{\lambda}_2})]}{\sigma_{\tilde{\lambda}_1}\sigma_{\tilde{\lambda}_2}},$$

where E and σ respectively denote the expected value and standard deviation. As expected from visual observation, we find that $|\rho| \geq 0.9$ for all possible combinations of the intensity functions of byte, packet, and flow clusters; however, $|\rho| \approx 0.6$ among the intensity functions of user clusters and that of byte, packet, or flow clusters. The visual inspection of intensity functions also shows that even within a close neighborhood such as a university, downtown, or suburb, there is differentiation between the application mix of different cells. Consequently there are opportunities for fine-grained network optimization within close neighborhoods, which are discussed later in Section IV. Note that such detailed analysis is made possible in our study because the mobility information in our data set obtained from radio access network is fine-grained.

We can also identify the geographical areas where one type of traffic is more prevalent than others using the difference of the intensity functions. For such geographical areas, cellular network operators can optimize network parameters to optimize for specific performance metrics. In Figure 7, we add up the intensity functions of `email` and `web browsing` clusters and plot its difference to the intensity function of `music audio`. We observe two distinct geographical areas where either `email` and `web browsing` or `music audio` traffic is dominant. It is well-known that email/web browsing and music traffic have conflicting Quality of Service (QoS) requirements. This type of analysis provides more actionable insights as compared to the basic cluster composition analysis described earlier.

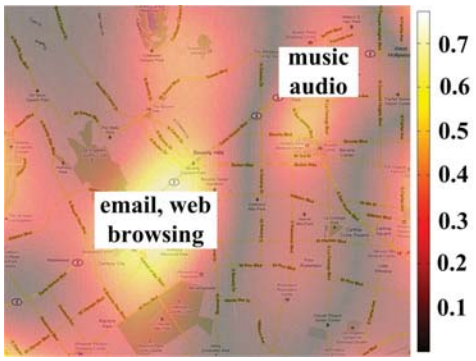


Fig. 7. Difference between intensity functions of `music audio` clusters and `email` + `web browsing` clusters for byte distribution.

IV. Major Findings and Implications

In this section, we provide a summary of major findings of our study and highlight their implications on network optimization.

1) *A few application realms dominate others in our data set (Figure 2).* We observed that `web browsing` and `email` are overall the most popular applications in our data set. This observation presents an optimization opportunity for cellular operators as it is known that web browsing and email traffic is typically bursty in nature. Therefore, cellular network operators can fine-tune radio network parameter settings. For instance, inactivity timers of radio resource control (RRC) state machine can be decreased for cells with more bursty-natured traffic to avoid wasteful occupation of radio channels that result in inefficient spectrum utilization [14].
2) *Any given application does not enjoy same level of popularity across different cell locations (Figure 3).* This finding implies that cellular network operators cannot take "one size fits all" approach in optimizing network parameters for specific applications.
3) *Application mix significantly varies across different neighborhoods (Table I).* From cluster composition analysis, we observed that application mixes significantly vary across downtown, university, and suburban neighborhoods. Furthermore, application mix of two same type of neighborhoods (*e.g.* suburb 1 and suburb 2) show significant similarity. Therefore, cellular network operators can generalize their optimization strategies across neighborhoods of the same type to some extent. In addition, we also observed that music and video applications are popular in a fraction of cells across all neighborhoods. In contrast to web browsing and email traffic, these applications are streaming in nature. Therefore, cellular network operators can fine-tune radio network parameter settings for them by increasing the inactivity timers of RRC state machine to avoid excessive state transitions that result in increased delays and packet losses [14].
4) *The popularity of different applications significantly varies even within a given neighborhood (Figures 6 and 7).* For more detailed optimization strategies, cellular network operators can utilize the difference of the intensity function of two applications to identify distinct cell locations where either of the applications dominant. Given the knowledge of the application preferences for a specific cell location, the cellular network operator may fine tune the QoS profile settings and the RNC admission control procedure when processing Radio Access Bearer (RAB) assignment requests for that specific cell. To the best of our knowledge, this finding represents the most fine-grained characterization of geospatial dynamics of application usage in a cellular network and provides actionable insights for network optimization.
5) *Application distributions significantly vary for byte, packet, flow, and user counts (Figure 2 and Table I).* This finding implies that cellular network operators should take care not to optimize cells solely by byte, packet, or flow volume as this may negatively impact other low volume–yet popular–applications that many users use in those cells. As a result, there is only a small set of cells where a specific application is popular with respect to all of the byte, packet, flow, and user counts. This leaves cellular network operators with a minority of cells where operators can optimize for specific applications while satisfying most users.

V. Related Work

In this section, we provide an overview of the prior research relevant to characterizing application usage in cellular data networks. The prior work that first provided evidence of geographic correlation of users' interests in a cellular network is by Trestian *et al.* in [18]. In their study, the authors categorized web requests into six groups: mail, social networking, trading, music, news, and dating; and categorized locations into 'home' and 'work'. Their study focused on differences in users' interests across different locations. They also identified hotspots – locations with large inflow of users – and studied users' interests across different hotspots. There are three important limitations of their work that we overcame in our study.

1) They only examined web requests (HTTP URLs), but traffic in modern cellular networks can be differentiated with respect to application protocol (*e.g.*, HTTP, DNS, SIP), class (*e.g.*, streaming audio, streaming video, web, email), and distinct applications downloaded from "App Stores." On the other hand, our data set more representative of mobile data usage as we identify and analyze 19 application realms in all IP traffic, not only in HTTP URLs as in [18]. This is important because a dominant mode of application usage on smartphones is through individual "apps," not only via traditional web browsers.
2) They showed differentiation in application interests at the macro-scale (neighborhoods) but not at the micro-scale (cell sectors) — this leaves open the question how granular geospatial differentiation actually is. On the other hand, cell sector locations in our traces are accurate to a finer timescales because they are collected directly from a UMTS radio network, not from core network servers, which do not record all cell changes due to handovers [22]. This accuracy enables us to detect distinct differences in application usage among cell sectors very close to each other within both 'home' and 'work' areas.
3) They only studied user interests with respect to session counts, whereas network operators are also interested in understanding application usage distributions with respect to traffic volume, flow counts, or unique user counts as they may yield completely different estimates of application popularity. On the other hand, we analyze

application usage on all of the above-mentioned four dimensions important to network operators (volume in terms of bytes and packets, flows, and users). We find that the 'top' applications and their prevalence in different areas does differ depending on the dimension used to analyze them.

Several other studies have also examined cellular network data traffic, but do not study the relationship between application usage and location as we do in this paper [5], [8], [12], [13], [16], [21]. The authors in [13] study traffic volume dynamics in cellular data networks. In particular, the authors study effective bit rate for different applications. The results of their experiments show that `P2P` and `http` traffic of certain popular sites have better effective bit rate than that of `VPN` and `https` traffic. In [16], the authors studied the distribution of applications across different cellular devices. The results of their measurement analysis showed that application volume distribution is highly skewed. They used a Zipf-like distribution further modeled aggregate and device-specific application volume distributions. In [5], Falaki *et al.* studied application usage patterns in data collected from 255 smartphone users. The results of their experiments highlighted strong diversity in the applications usage, in terms of number of applications and interaction time across user population. Huang *et al.* [8] studied data from a cross-platform measurement tool. They studied key factors that impact network and web browsing performance of applications for different carrier networks, device capabilities, and sever configurations. In a similar work, the authors in [12] studied end-to-end key performance indicators in cellular networks. In [21], the authors developed a measurement platform to collect end-to-end latency, throughput, and timeout interval statistics between cellular devices and the designated servers.

VI. Conclusions and Future Work

In this paper, we characterized the geospatial dynamics of application usage in a 3G cellular data network. Using traces collected from the network of a tier-1 cellular operator in the United States, we first clustered cell locations based on their application usage and then conducted the geospatial analysis of cells belonging to different clusters. The results of our empirical study revealed that the cell clustering results are significantly different for byte, packet, flow, and user distributions across different geographical regions. However, our results also suggested that care should be exercised so that cells are not optimized solely with respect to traffic volume based on byte, packet, or flow counts because this may negatively impact other low volume applications used by most users in those cells. These and other findings of our measurement analysis have important implications in terms of network design and optimization. To our best knowledge, this paper presents the first attempt to conduct fine-grained analysis of the geospatial dynamics of application usage in cellular networks. In future, we plan to extend the analysis presented in this study by collecting a data set over longer time duration. We also plan to utilize other more rigorous techniques such as Ripley's k-cross function and variogram for analyzing geospatial dynamics of application usage in cellular networks [3].

References

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

[2] M. M.-T. Chiang and B. Mirkin. Experiments for the number of clusters in k-means. In *Lecture Notes in Computer Science, Progress in Artificial Intelligence*, 2007.

[3] N. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1993.

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

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

[6] R. Haining. *Spatial Data Analysis: Theory and Practice*. Cambridge University Press, 2003.

[7] J. Hartigan. *Clustering Algorithms*. J. Wiley & Sons, 1975.

[8] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and V. Bahl. Anatomizing application performance differences on smartphones. In *ACM MobiSys*, 2010.

[9] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

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

[11] B. Mirkin. *Clustering for Data Mining: A Data Recovery Approach.* Chapman and Hall, 2005.

[12] B. M. Orstad and E. Reizer. End-to-end key performance indicators in cellular networks. Master's thesis, Agder University College, Norway, 2006.

[13] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das. Understanding traffic dynamics in cellular data networks. In *IEEE Infocom*, 2011.

[14] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. Characterizing radio resource allocation for 3g networks. In *ACM SIGCOMM IMC*, pages 137–150, New York, NY, USA, 2010.

[15] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[16] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang. Characterizing and modeling internet traffic dynamics of cellular devices. In *ACM SIGMETRICS*, 2011.

[17] 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.

[18] 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.

[19] F. P. Tso, J. Teng, W. Jia, and D. Xuan. Mobility: A double-edged sword for HSPA networks. In *ACM MobiHoc*, 2010.

[20] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Monographs on Statistics and Applied Probability. Chapman & Hall, 1995.

[21] M. P. Wittie, B. Stone-Gross, K. Almeroth, and E. Belding. MIST: Cellular data network measurement for mobile applications. In *IEEE BROADNETS*, 2007.

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

[23] Q. Xu, A. Gerber, Z. M. Mao, J. Pang, and S. Venkataraman. Identifying diverse usage behaviors of smartphone apps. In *ACM IMC*, 2011 (to appear).