Geospatial and Temporal Dynamics of Application Usage in Cellular Data Networks

M. Zubair Shafiq Lusheng Ji Alex X. Liu Jeffrey Pang Jia Wang

Abstract—Significant geospatial and temporal correlations, in terms of traffic volume and application access, exist in cellular network usage as shown in recent studies on cellular network measurement. Such geospatial and temporal 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 joint characterization of the geospatial and temporal 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 and temporal 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 for better network performance and user experience.

Index Terms—Apps; Cellular Networks

1 Introduction

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 [2]. 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 increasingly focus on optimizing different aspects of the network by customized design and management to improve key performance indicators (KPIs).

There are three aspects of a cellular network that present significant optimization potential to the network operators: (1) *diverse application mix constituting the data traffic*, (2) *variations in the traffic depending upon the geo-location of users*, and (3) *temporal variations in the traffic*. It has been shown that the performance of different applications constituting the data traffic in cellular networks is sensitive to various network KPIs [11], [16]. Tso *et al.* also showed that the network performance perceived by users is strongly related to their geolocation and mobility patterns [25]. Furthermore, Xu *et al.* showed that different applications have different temporal dynamics [29]. Combining the aforementioned 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, temporal dynamics, and users' geospatial movement patterns.

Trestian *et al.* conducted a study that provided the first evidence of geographic correlation of users' "interests" in a cellular network [24]. 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 aforementioned coarse-grained geographic correlation to a more fine-grained cell sector correlation, as this is typically the smallest unit

- *The preliminary version of this paper titled "Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Network" was published in the proceedings of the 31th Annual IEEE Conference on Computer Communications (INFOCOM), Orlando, Florida, 2012.*
- *Alex X. Liu is the corresponding author of this paper.*
- *M. Zubair Shafiq is with the Department of Computer Science, The University of Iowa, Iowa City, IA, USA. Email:zubairshafiq@uiowa.edu*
- *Alex X. Liu is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA. Email:alexliu@cse.msu.edu*
- *Lusheng Ji, Jeffrey Pang, and Jia Wang are with AT&T Labs – Research, Bedminster, NJ, USA. Email:{lji, jeffpang, jiawang}@research.att.com*

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 [16]. Other prior studies that either study application usage or geospatial patterns (but not both simultaneously) include but are not limited to [6], [11], [15], [22], [25], [27]. Further details of prior art are in Section 8.

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

1) **Methodology:** 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.
   We study application usage characteristics of users across more than two thousand 3G cell locations. For the 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, intensity function analysis, and point-pattern based co-location analysis in this paper. To study the temporal dynamics of geospatial dependencies in application usage, we apply the aforementioned geospatial analysis techniques on the updated cell clustering results for different time intervals such as 0000hrs-0300hrs, 0300hrs-0600hrs.
2) **Findings and Implications:** The results of our geospatial and temporal 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 region such as a university, downtown, or suburb. Consequently, there are opportunities for fine-grained network optimization within close regions. Finally, we find that cell locations with particular application mixes have tendency to be co-located. This information can be used by operators to optimize frequency planning and management of transmission power and handovers in co-located cells.

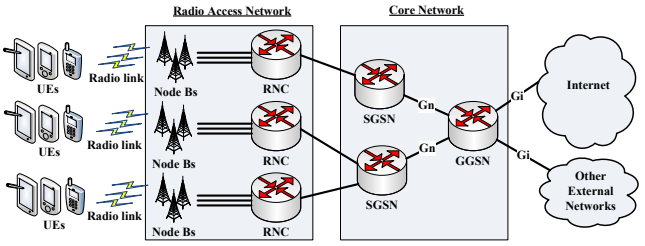


Fig. 1. 3G UMTS cellular data network.

2 Background and Data

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.

2.1 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 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* (RNCs), *Serving General Packet Radio Service [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 a physical base station. The data traffic generated by a cellular device is first sent to a NodeB and then 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 that 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.

2.2 Data Sets

In this paper, we use two anonymized data sets from a tier-1 cellular network operator for our study. The first data set contains flow-level information of IP traffic carried in PDP Context tunnels (*i.e.*, all data traffic sent to and from cellular devices). This data set is collected from all the links between SGSNs and GGSNs, called *Gn* links, in the core network and covers 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 set does not permit the reversal of the anonymization or re-identification of users. For proprietary reasons, the results presented in this paper are sum-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 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 [5]. 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 (more than 95% of all data traffic in terms of byte volume), so understanding the remainder is not critical for the purpose of traffic engineering [29]. Furthermore, we categorize applications into the following application realms, in no particular order, based on their functionality and traffic type (streaming, interactive, *etc.*) [21]. (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`. For example, `mixed HTTP streaming` includes apps like YouTube, `radio audio` includes apps like TuneIn Radio, `social network` includes apps like Facebook, and `music audio` includes apps like Pandora. Note that the application realms are non-overlapping. Applications can belong to multiple categories because of their dual functionality and traffic type. For example, YouTube can be classified as `mixed HTTP streaming` or `streaming video`; however, we classify YouTube as `mixed HTTP streaming` because it mainly uses HTTP streaming on smartphones [19].

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 [28]. 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 based on their timestamps, we get accurate cell locations at a 2 second granularity for IP flows in the first data set. In practice, a device may be connected to multiple cell sectors at the same time to allow transmission of uplink data from HSPA devices to the most suitable sector based on factors such as signal strength, load, interference, *etc.* 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 [20]. 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 from 1600hrs on December $6, 2010$ till 2400hrs December $7, 2010$.

The data sets cover a large metropolitan area spanning more than $1,200$ km^2 in the United States. The metropolitan area had complete 3G coverage; therefore, it would be rare for a device to handoff to any neighboring 2.5G cells. Thus, the data sets cover more than two thousand 3G cells in the metropolitan area, 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.

3 Aggregate 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 temporal dynamics of aggregate traffic, and then study application usage distributions in the traffic, and finally investigate the relative popularity of individual applications across different cell locations.

3.1 Temporal Analysis

As mentioned in Section 2, all traffic records in our data set are timestamped and are tagged with application and cell identifiers. Below, we analyze hourly variations in the traffic volume during 24 hours on December 7, 2012 for the sake of clarity. We first study the temporal dynamics of aggregate traffic volume. Figure 2 shows the temporal dynamics of aggregate traffic in terms of byte, packet, flow, and user counts. As reported in prior literature [16], [22], [29], we observe a strong diurnal behavior in aggregate traffic. However, we observe two daily peaks – instead of a daily peak observed in prior literature [22], [29] – and the second peak is around mid-night, which might reflect users' peculiar activity patterns in the metropolitan area studied in this paper. We note that the aggregate traffic volume during day time is significantly more than that during night time. Furthermore, the variations in traffic volume are different across bytes, packets, flows, and users. We also study the temporal dynamics of traffic volume for different applications. Figure 3 shows the temporal dynamics of traffic belonging to four applications for byte, packet, flow, and user counts. As observed for aggregate traffic, we observe strong diurnal characteristics in temporal dynamics across all applications. However, there are interesting differences across these applications. For instance, we note that traffic volumes of `dating` and `social network` applications peak around late night in terms of byte count. On the other hand, traffic volume of `web browsing` and `maps` applications peak around noon and afternoon. We also observe subtle differences in the temporal dynamics of traffic volume for byte, packet, flow, and user counts. We further elaborate on these observations in the rest of this section.

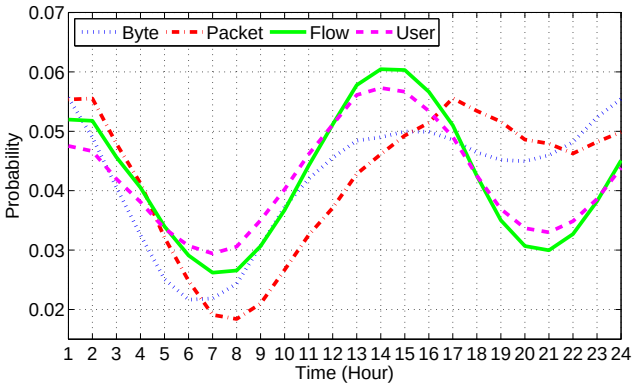
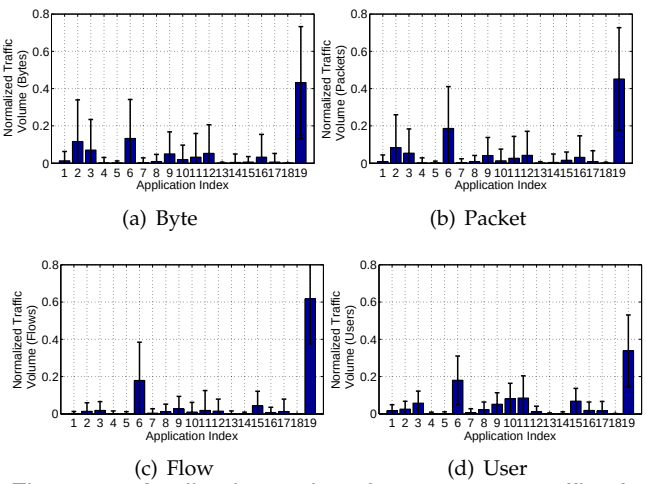


Fig. 2. Temporal dynamics of aggregate traffic in terms of bytes, packets, flows, and users.

3.2 Application Analysis

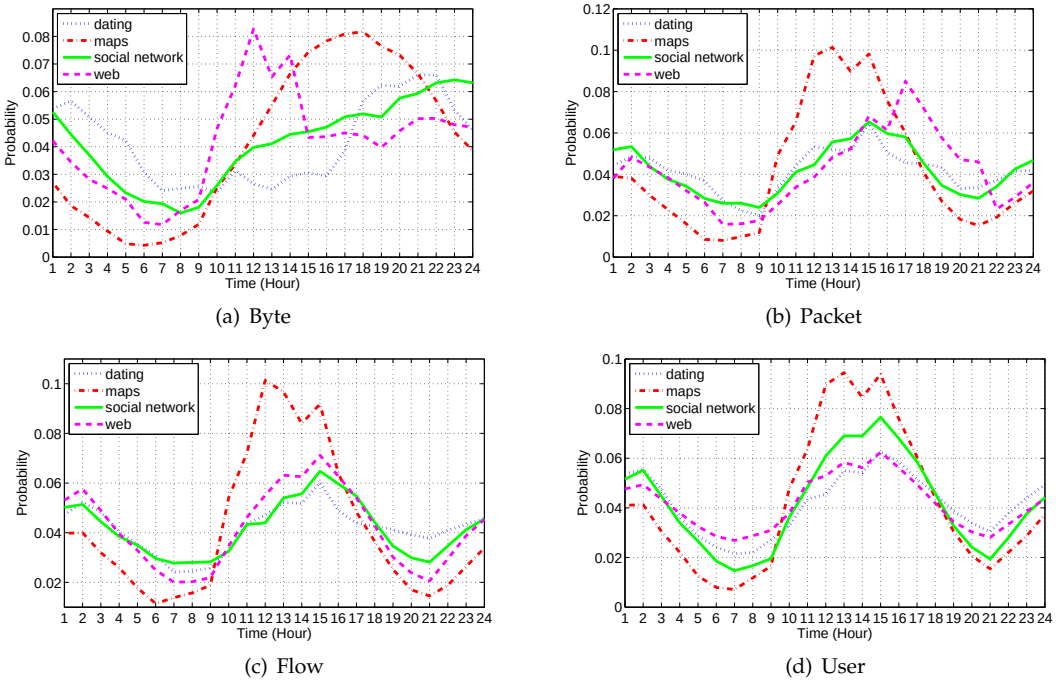
We now segregate all traffic records with respect to the application identifiers to study the application usage patterns. Towards this end, we construct application distributions using application identifiers as keys and byte, packet, flow, or user counts as values. In the rest of this paper, the terms 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. Figure 4 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.



(a) Byte (b) Packet

(c) Flow (d) User

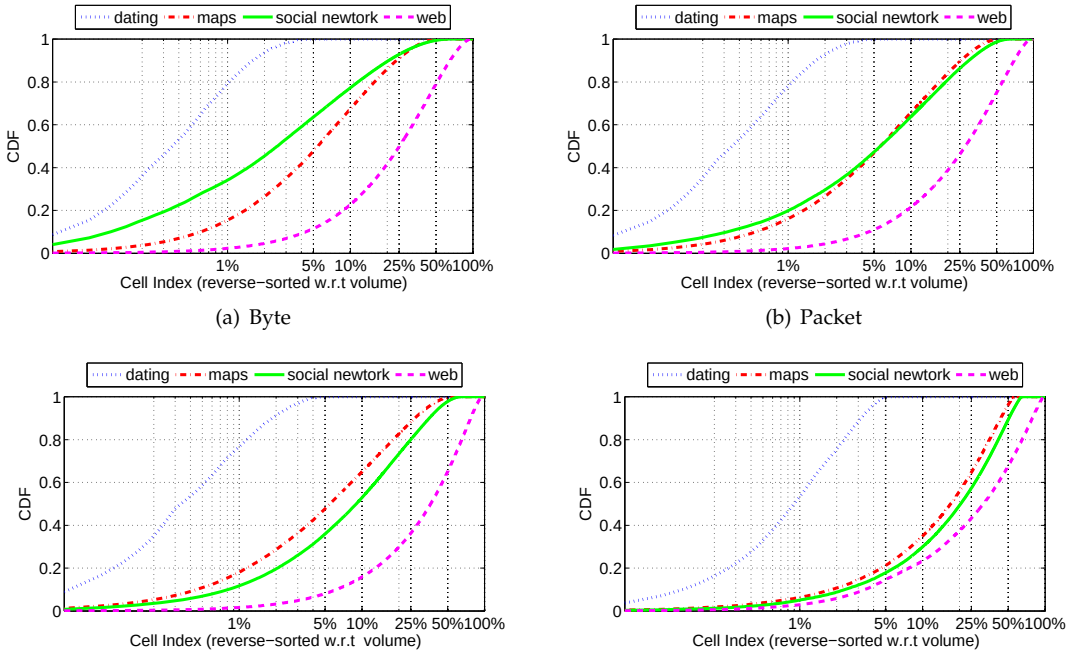
Fig. 4. Application mix of aggregate traffic for byte, packet, flow, and user distributions. The mapping of application indices is as follows: (1) `ads`, (2) `mixed HTTP streaming`, (3) `app store`, (4) `media optimization`, (5) `dating`, (6) `email`, (7) `games`, (8) `news info image media`, (9) `maps`, (10) `misc`, (11) `mms`, (12) `music audio`, (13) `p2p`, (14) `radio audio`, (15) `social network`, (16) `streaming video`, (17) `voip`, (18) `vpn`, and (19) `web browsing`.



(a) Byte (b) Packet

(c) Flow (d) User

Fig. 3. Temporal dynamics of applications in terms of bytes, packets, flows, and users.



(a) Byte (b) Packet

(c) Flow (d) User

Fig. 5. Distributions of traffic volume with respect to byte, packet, flow, and user counts across all cell sector locations.

3.3 Geospatial Analysis

We now study the relative popularity of a given application across different cell locations in our data set. Figure 5 shows the cumulative distribution function (CDF) of traffic volume of `dating`, `maps`, `social network`, and `web browsing` applications 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 three major findings: (1) traffic volumes of applications exhibit strong diurnal characteristics, (2) the popularity of a given application realm varies across different cell locations, and (3) the traffic volume of a few application realms dominate others overall. These findings suggest strong dependence of application usage on geospatial and temporal dynamics. We follow a three 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. Third, we analyze the temporal dynamics of geospatial dependencies in application usage. 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 aforementioned steps in the following sections.

4 Cell Clustering

To study the application usage patterns for any given cell, we now segregate all traffic records with respect to the application and cell identifiers. 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. 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 from Figure 4.

We utilize a well-known unsupervised clustering algorithm called k-means to cluster application distributions of cells. The k-means algorithm is a simple yet effective technique to cluster feature vectors into a predefined k number of groups [13]. The selection of appropriate value of k is crucial and is an open research problem [4]. Several heuristics have been proposed in prior literature, which primarily focus on the change in intra-cluster dissimilarity for increasing values of k [8], [12], [14]. A well-known heuristic, called gap statistic, can be used to compare the change in intra-cluster dissimilarity W_k for given data and that for a reference null distribution [23]. Gap statistic provides a statistical method to find the elbow of intra-cluster dissimilarity W_k as the values of k is varied over B iterations. 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. In this work, we set the value of $B = 1000$ and the initial centroid is randomly selected in each iteration to avoid any bias. Figure 6 shows the plot of gap statistic for varying values of k for byte distributions. 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. Similar results were obtained for packet, flow, and user distributions.

To gain insights into the clustering results, we plot four cluster centroids of byte, packet, flow, and user distributions in Figure 7. The error bars represent the intra-cluster standard deviation for each application. While the size of error bars may be affected by the number of applications and traffic volume for each category, we observe that traffic volume generally correlates with standard deviation rather than the number of applications. We label 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 7, we also provide the percentage distribution of cells across all cluster types. As expected, we observe that `web browsing` and `email` are the common cluster

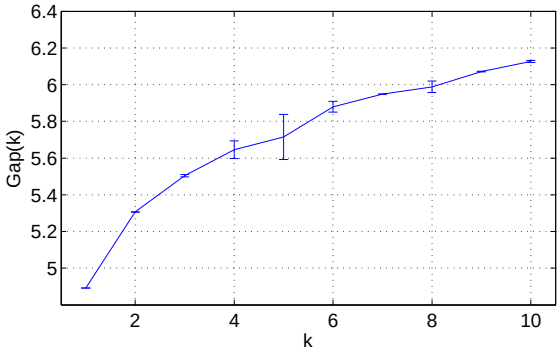
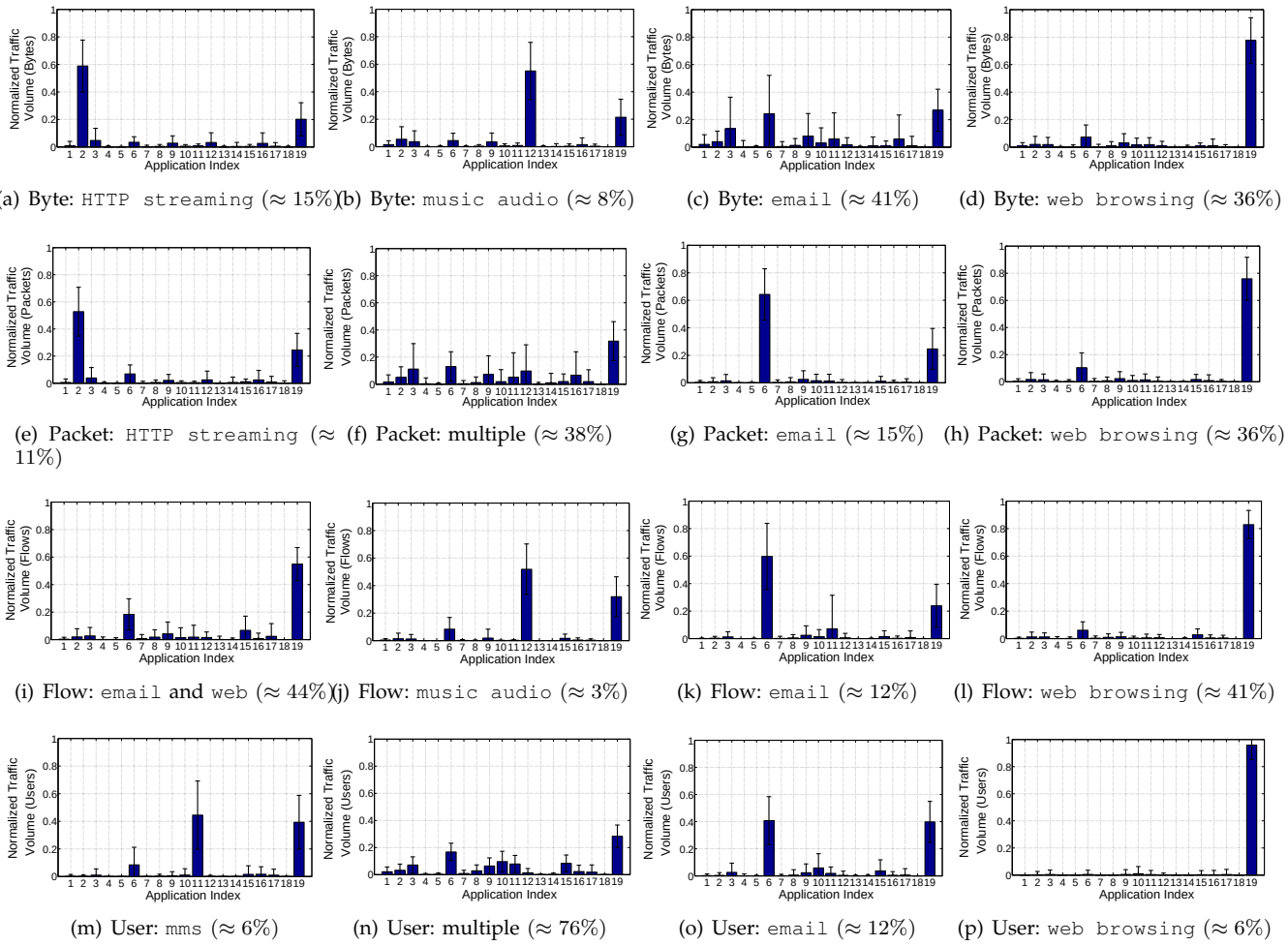


Fig. 6. Gap statistic for finding the suitable number of clusters for traffic distributions of cells.



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

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

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

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

Fig. 7. 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 as follows: (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`, and (19) `web browsing`.

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 7 highlight the important differences across byte, packet, flow, and user distributions. For instance, we observe that only one or two applications (*e.g.* `email`, `web browsing` 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 7(n) shows that 76% of cells fall into the `multiple` realm for user distributions. This implies 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, approximately 40% cells belong to this cluster for byte, packet, and flow distributions. As we discuss later in Section 7, these differences have important implications in terms of cellular network planning and optimization.

5 Geospatial Analysis

Using the clustering methodology presented in the previous section, we uniquely label each cell location for byte, packet, flow, and user application distribution clusters. For geospatial analysis, we apply the following three techniques, (1) basic cluster composition analysis, (2) intensity function analysis, and (3) co-location 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. These regions were differentiated based on manual analysis of the points

of interest. For example, the region around a university campus was classified as university, whereas the regions surrounding co-located housing areas were classified as suburbs.

5.1 Cluster Composition Analysis

In the cluster composition analysis, we study the distribution of cells belonging to different clusters in various geographical regions. This analysis aims to uncover the cases where cells belonging to a particular cluster type are more prevalent in certain geographical regions.

Table 1 shows the distribution of cells belonging to different clusters across all geographical regions. We observe important differences in application usage across different geographical regions 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, while 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 7. Table 1 also indicates that a majority of cells belong to `multiple` cluster for user count distributions across all geographical regions. For instance, Table 1 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 most users in these cells.

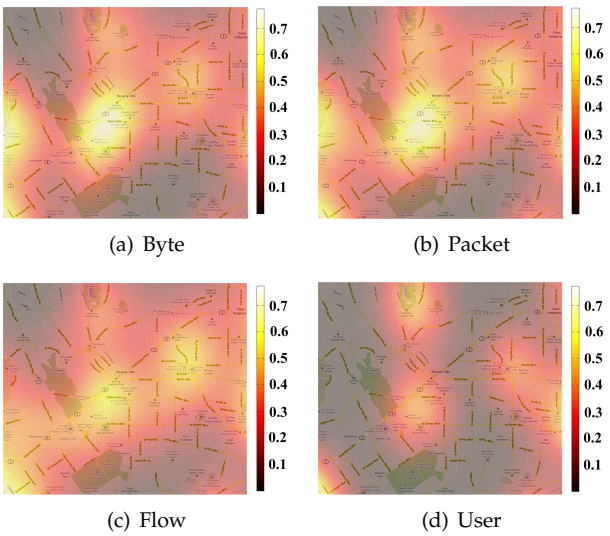
5.2 Intensity Function Analysis

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 [7]. 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 [26]. A typical kernel estimated intensity function takes the form:

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

TABLE 1
Cluster composition analysis results

	mixed HTTP streaming	music audio	email	web browsing
Byte (%)				
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



(a) Byte (b) Packet

(c) Flow (d) User

Fig. 8. Kernel estimated intensity function for `web browsing` cluster types in a suburban region for byte, packet, flow, and user distributions.

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 8. We can visually observe similarity among the intensity functions for byte, packet, and flow distributions; whereas, the intensity function for user distribution is 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 [18]. The magnitude of one signifies perfect

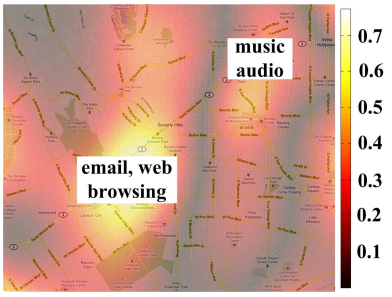


Fig. 9. Difference between intensity functions of `music audio` clusters and `email` + `web browsing` clusters for byte distribution.

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 region 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 regions, which are discussed later in Section 7. 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 for specific performance metrics. In Figure 9, 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.

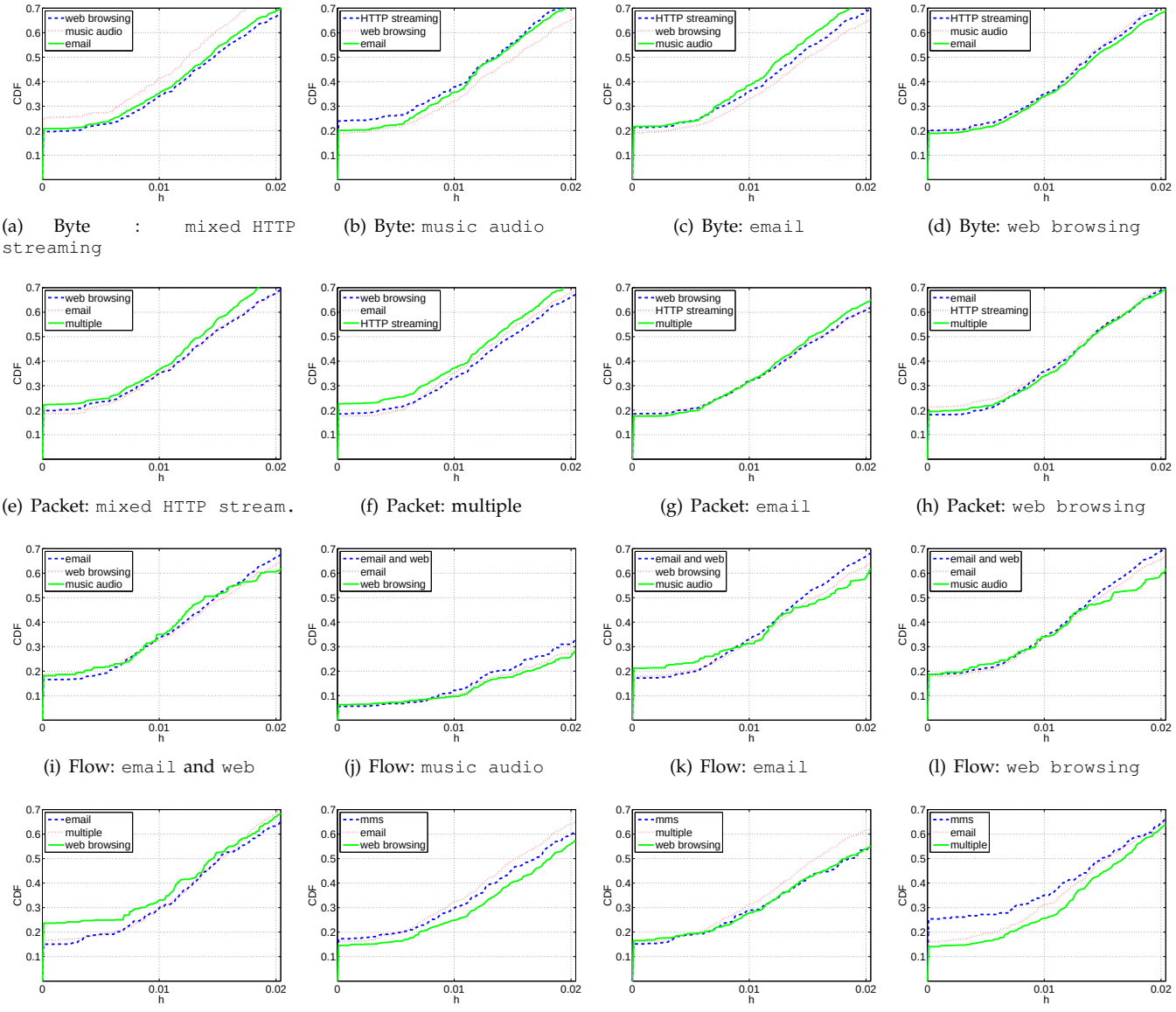
5.3 Co-location Analysis

The intensity function is designed for univariate geospatial analysis to identify the geographical regions where an application is popular. As we show in Section 5.2, it can also be extended for multivariate geospatial analysis, by utilizing the differences among individual intensity functions, to identify the geographical regions where an application is more popular than the rest. However, the multivariate intensity function analysis does not convey precise information about co-location of cells belonging to different clusters. Such information is useful to cellular network operators for frequency planning and management of handovers among co-located cells.

To characterize the co-location characteristics of cells belonging to different clusters, there are two candidate point pattern interaction analysis techniques: (1) Ripley's cross-K function, and (2) nearest neighbor function [3]. Ripley's cross-K is a mean based statistic that is defined as a function of distance between two point sets, which represent cluster cell locations for the present problem. On the other hand, the nearest neighbor function specifically considers the nearest neighbors of one point set from another point set as a function of distance. Among these two techniques, we choose the nearest neighbor function because network operators are primarily interested in findings about immediately co-located cell locations. To define the nearest neighbor function for two sets of points i and j, let $G_{ij}(h)$ be the probability that the distance from a randomly selected point i to the nearest point j is less then or equal to h. Note that G is a non-symmetric measure. Given i and j represent the cell locations of two clusters, Figure 10 plots the CDFs of $G_{ij}(h)$ for all byte, packet, flow, and user clusters over a range of h. The ordering of G CDFs indicates the relative attraction between cell locations of different clusters. Specifically, $G_{ij}(h) > G_{ik}(h)$ shows that cell locations of cluster i are closer to cell locations of cluster j as compared those of cluster k. We provide the most significant observations for byte, packet, flow, and user clusters below. For byte clusters, we observe that `mixed HTTP streaming` and `music audio` cells tend to be co-located. For packet clusters, we observe that `mixed HTTP streaming` and `multiple` cells tend to be co-located. For flow clusters, we observe that `email` and `music audio` cells tend to be co-located. For user clusters, a major observation is that `mms` and `web browsing` cells are mostly co-located.

6 Temporal Analysis

In this section, we analyze the temporal dynamics of geospatial dependencies in application usage. We separately analyze each of the geospatial analysis techniques used in the previous section, including cluster composition, intensity function, and co-location analyses. Recall that these geospatial analysis techniques operate on the clustering results obtained in Section 4. For temporal analysis of these geospatial analysis techniques, we follow a three step process. First, we separately construct application distributions of all cells for different time intervals, *e.g.* 0000hrs-0300hrs, 0300hrs-0600hrs, *etc.* Second, we re-label application distributions of cells with the centroids identified in



(a) Byte : `mixed HTTP streaming` (b) Byte: `music audio` (c) Byte: `email` (d) Byte: `web browsing`

(e) Packet: `mixed HTTP stream.` (f) Packet: multiple (g) Packet: `email` (h) Packet: `web browsing`

(i) Flow: `email` and `web` (j) Flow: `music audio` (k) Flow: `email` (l) Flow: `web browsing`

(m) User: `mms` (n) User: multiple (o) User: `email` (p) User: `web browsing`

Fig. 10. Nearest neighbor metric for co-location analysis.

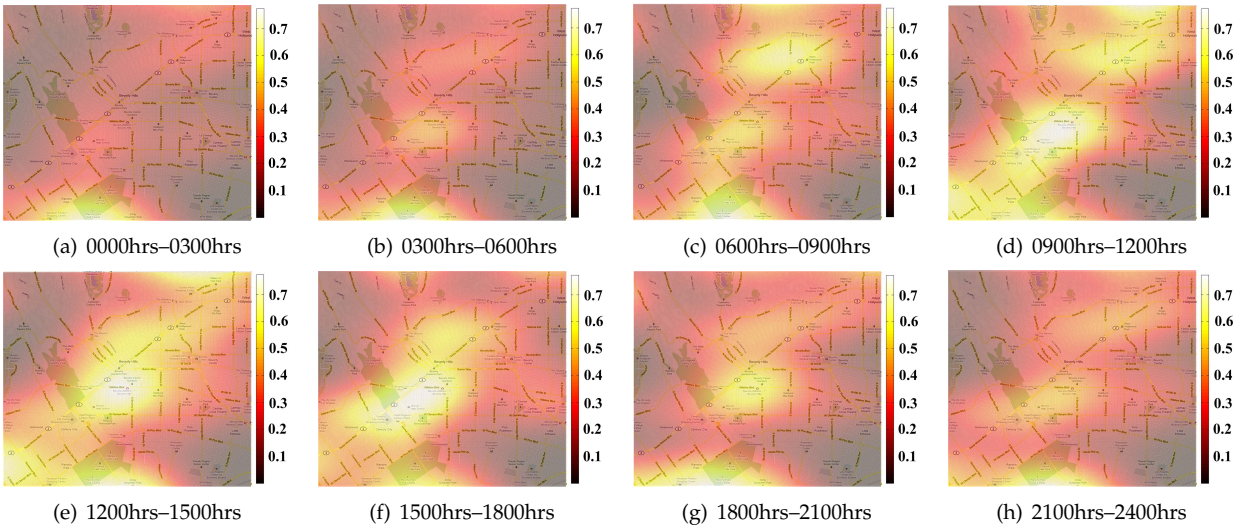
Section 4. Finally, using the new labels, we recompute and analyze results for the three geospatial analysis techniques. Next we separately present the temporal analysis for all of them.

6.1 Cluster Composition Analysis

The cluster composition results of the byte distribution for different time intervals and geographical regions are provided in Figure 11. Overall, we observe some fluctuations in the cluster composition results for different time intervals. For instance, in downtown region, the percentage of cells belonging to `web browsing` cluster is more than `email` cluster for 0600hrs-0900hrs, 0900hrs-1200hrs, 1200hrs-1500hrs, 1500hrs-1800hrs, and 1800hrs-2100hrs time intervals. However, this trend is reversed for 0000hrs-0300hrs, 0300hrs-0600hrs, and 2100hrs-2400hrs time intervals. Moreover, `mixed HTTP streaming` accounts for more traffic in suburbs as compared to university area. We observe several of these variations across applications and geographical regions. These observations indicate that cellular network operators should take into account such temporal variations while deploying pre-dominant application specific network optimization strategies.

6.2 Intensity Function Analysis

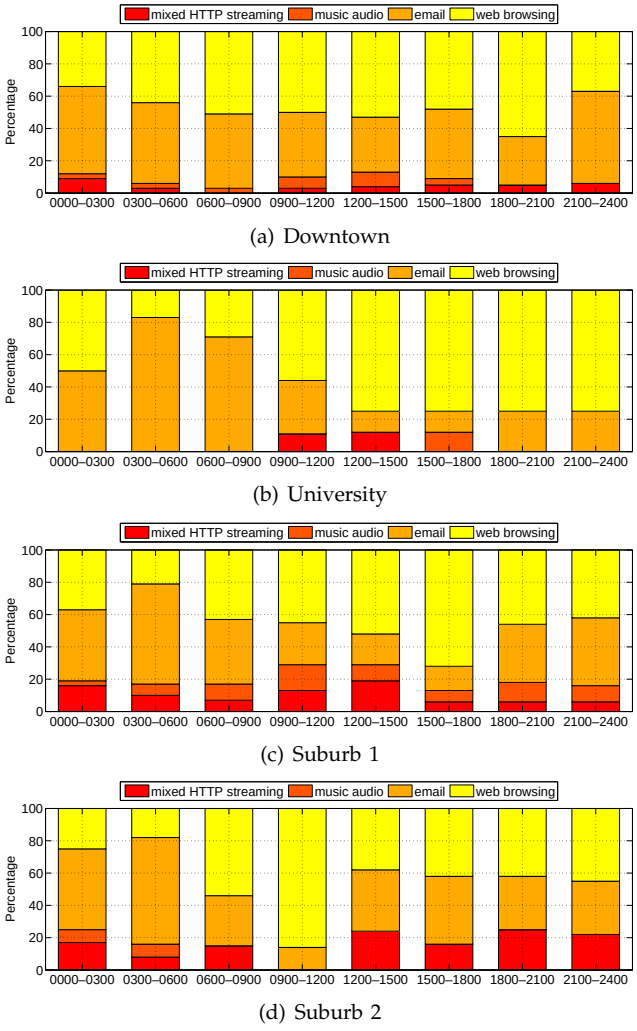
Figure 12 plots kernel estimated intensity function of `web browsing` cluster for different time intervals in a suburban region. Similar to the temporal variations observed in cluster composition analysis results, we observe some fluctuations in the shape of intensity function plots. We can identify three major high intensity regions at top-right, middle, and bottom-left of the plots. These regions have varying relative intensities at different time intervals with some distinct patterns. For example, the middle region



(a) 0000hrs–0300hrs (b) 0300hrs–0600hrs (c) 0600hrs–0900hrs (d) 0900hrs–1200hrs

(e) 1200hrs–1500hrs (f) 1500hrs–1800hrs (g) 1800hrs–2100hrs (h) 2100hrs–2400hrs

Fig. 12. Temporal dynamics of kernel estimated intensity function for `web browsing` byte distribution clusters in a suburban region.



(a) Downtown

(b) University

(c) Suburb 1

(d) Suburb 2

Fig. 11. Temporal dynamics of cluster composition analysis results for byte distributions.

has the highest intensity for 0900hrs-1200hrs, 1200hrs-1500hrs, and 1500hrs-1800hrs time intervals. Thus, fine-grained tuning of network parameters should incorporate temporal variations.

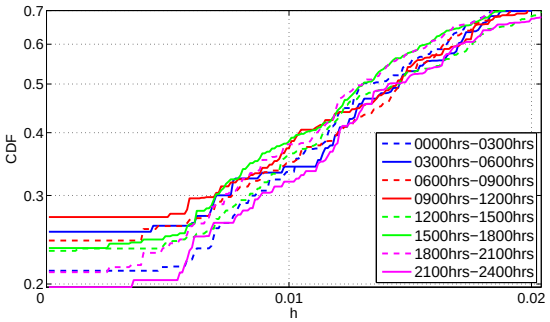


Fig. 13. Temporal dynamics of nearest neighbor metric for co-location between `mixed HTTP streaming` and `music audio` byte distribution clusters.

6.3 Co-location Analysis

Figure 13 plots the CDFs of $G_{ij}(h)$ for `mixed HTTP streaming` and `music audio` byte distribution clusters in different time intervals. Similar to the cluster composition and intensity function analyses results, we observe a pattern in co-location `mixed HTTP streaming` and `music audio` clusters. The ordering of CDFs in Figure 13 shows that the probability of their co-location is higher during 0300hrs-1800hrs and it drops substantially for other time intervals. Cellular network operators should take into account such temporal variations for network planning and management among neighboring cells.

7 Major Findings and Implications

In this section, we first provide a summary of major findings of our study and then provide an example of how operators can leverage the findings for network optimization.

7.1 Summary of Findings

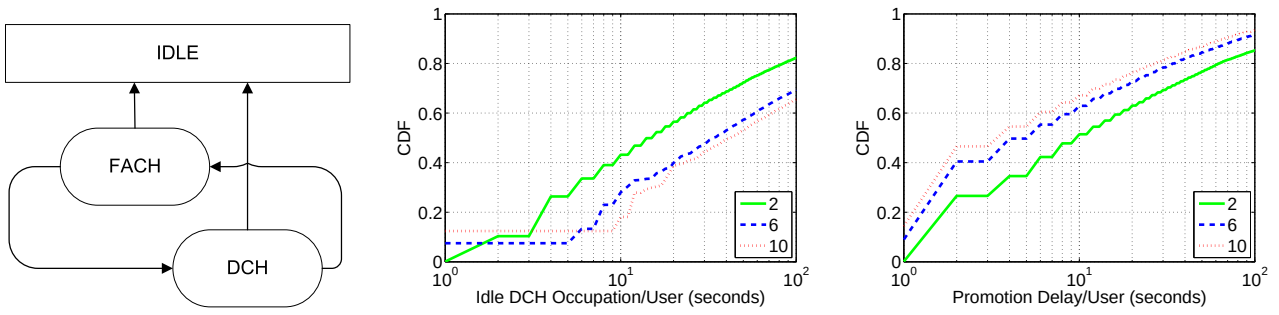
1) *A few application realms dominate others in our data set (Figure 4).* We observed that `web browsing` and `email` are overall the most popular applications in our data set. This observation presents

an optimization opportunity for cellular operators specifically for these applications.

2) *Any given application does not enjoy the same level of popularity across different cell locations (Figure 5).* This finding implies that cellular network operators cannot take "one size fits all" approach in optimizing network parameters for specific applications.
3) *Application usage has diurnal characteristics (Figure 3).* We observed that `web browsing` and `maps` applications have peak usage around noon and afternoon. On the other hand, `dating` and `social network` applications have peak usage around late night. Consequently, cellular network operators need to dynamically adapt various network parameters for optimal performance.
4) *Application mix significantly varies across different cell sectors (Table 1).* From cluster composition analysis, we observed that application mixes significantly vary across downtown, university, and suburban regions. Furthermore, application mix of two same type of regions (*e.g.* suburb 1 and suburb 2) show significant similarity. Therefore, cellular network operators can generalize their optimization strategies across regions 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 regions. In contrast to `web browsing` and `email` traffic, these applications are streaming in nature. Therefore, cellular network operators can again fine-tune radio network parameter settings for them.
5) *The popularity of different applications significantly varies even within a given region (Figures 8 and 9).* 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.
6) *Certain applications have a higher probability of co-location (Figure 10).* Network operators can use this co-location information for frequency planning and handover management among co-located cells. For instance, cells with more streaming traffic should handover users to the neighboring cells quicker compared to the cells with mostly best effort traffic like `email`.
7) *Geospatial dependencies also have temporal variations (Figures 11, 12, and 13).* Cellular network operators need to dynamically optimize network parameters and settings as discussed in above findings.
8) *Application distributions significantly vary for byte, packet, flow, and user counts (Figure 4 and Table 1).* This finding implies that cellular network operators may not 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.

7.2 Implications

Cellular network operators can leverage the aforementioned findings to optimize various network parameters to improve performance. Below, we show using trace-driven simulations that cellular network operators can adapt Radio Resource Control (RRC) state machine inactivity timers to improve performance [1]. UEs acquire and release radio resources by transitioning to different states in their RRC state machines, which are synchronously maintained by the UEs and network. Figure 14(a) shows the RRC state machine with three states: Idle, Forward Access Channel (FACH), and Dedicated Channel (DCH) – each with progressively more allocated radio resources. When a UE has some data to transfer, it is promoted to a higher energy state. Likewise, a UE is demoted to a lower energy state based on inactivity timeouts. Shorter inactivity timeouts result in more efficient radio resource utilization via more frequent state promotions. However, frequent state promotions can also result in degraded user experience especially for delay sensitive applications such as web browsing. Therefore, RRC inactivity timers can be increased in `web browsing` cell sectors to improve user experience. On the other hand, RRC inactivity timers can be reduced in cell sectors belonging to delay tolerant application clusters such as `music audio` for more efficient radio resource utilization. The RRC state machine of every user is simulated using the RNC logs while focusing on the DCH state, which has the highest allocated radio resources compared to all RRC states. We study the effect of varying DCH→FACH RRC timeout parameter (denoted by $T_{\text{DCH}\rightarrow\text{FACH}}$) on radio resource utilization efficiency for `music audio` cell sectors and state promotion delay for `web browsing` cell sectors in Figures 14(b) and 14(c). In Figure 14(b), we observe that idle DCH occupation time decreases in `music audio` cell sectors for decreasing values of $T_{\text{DCH}\rightarrow\text{FACH}}$. Due to the buffer-based streaming nature of traffic in `music audio` cell sectors, cellular network operators can reduce the values of RRC inactiv-



(a) RRC State Machine (b) Idle DCH state occupation time in `music audio` cell sectors for varying $T_{\text{DCH}\rightarrow\text{FACH}}$ (c) State promotion delay in `web browsing` cell sectors for varying $T_{\text{DCH}\rightarrow\text{FACH}}$

Fig. 14. Effect of tuning DCH→FACH timeout parameter ($T_{\text{DCH}\rightarrow\text{FACH}}$) on performance metrics.

ity timeouts to free up radio channels to accommodate additional users without affecting user experience. In Figure 14(c), we observe that state promotion delay increases in `web browsing` cell sectors for decreasing values of $T_{\text{DCH}\rightarrow\text{FACH}}$. Therefore, cellular network operators can increase the values of RRC inactivity timeouts to reduce the delays and improve users' web browsing experience.

Although the findings and implications presented here are based on traffic traces collected 3 years ago from a 3G cellular network, we argue they would translate to current 4G LTE networks. For instance, the findings about geospatial and temporal dynamics of application usage are not specific to a particular radio access technology. Moreover, LTE networks use RRC state machines similar to 3G networks and are subject to similar performance tradeoffs based on RRC parameters [10]. For example, the promotion delay is reported to be 260ms (compared to 2 seconds in 3G) and the inactivity timer is reported to be 12 seconds (compared to 12 seconds in 3G) in an operational LTE network [9], [17]. 4G LTE cellular network operators can tune these parameters to optimize performance for different applications in various cell sectors.

8 Related Work

Several studies have examined cellular network data traffic, but do not study the relationship between application usage and location as we do in this paper [6], [11], [15], [16], [22], [24], [27], [29]. The seminal work that provided first evidence of geographic correlation of users' interests in a cellular network is by Trestian *et al.* in [24]. 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. Later, Xu *et al.* also provided evidence of variation in app usage across different US states [29]. There are two major limitations of these studies that we overcame in this paper. First, they only examined web requests (HTTP URLs) or smartphone apps, but traffic in modern cellular networks can be differentiated with respect to application protocol (*e.g.*, HTTP, DNS, SIP) and class (*e.g.*, streaming audio, streaming video, web, email). On the other hand, our data set is 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 [24] or app name as in [29]. Second, they showed differentiation in application interests at the macro-scale but not at the micro-scale (cell sectors), this leaves the open 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 [28]. This accuracy enables us to detect distinct differences in application usage among cell sectors very close to each other.

9 Conclusion

In this paper, we jointly characterized the geospatial and temporal 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 and temporal 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 and temporal dynamics of application usage in cellular networks.

Potential extensions of this work include analyzing geospatial and temporal dynamics of application usage in cellular data networks using: (1) multiple data sets collected from the same location at different

times, (2) multiple data sets collected from different locations at the same time, and (3) data sets collected over a long time duration. The key challenge in obtaining these traces is that data collection at the RNCs introduces non-trivial overheads, which can jeopardize network operation particularly during peak hours.

REFERENCES

[1] Radio Resource Control (RRC) Protocol specification. Technical Report TS 25.331, 3GPP.

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

[3] R. S. Bivand, E. J. Pebesma, and V. Gomez-Rubio. *Applied Spatial Data Analysis with R*. Springer, 2008.

[4] 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.

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

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

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

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

[9] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. A Close Examination of Performance and Power Characteristics of 4G LTE Networks. In *ACM MobiSys*, 2012.

[10] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck. An In-depth Study of LTE: Effect of Network Protocol and Application Behavior on Performance. In *ACM SIGCOMM*, 2013.

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

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

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

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

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

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

[17] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. Characterizing Radio Resource Allocation for 3G Networks. In *ACM IMC*, 2010.

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

[19] M. Z. Shafiq, J. Erman, L. Ji, A. X. Liu, J. Pang, and J. Wang. Understanding the Impact of Network Dynamics on Mobile Video User Engagement. In *ACM Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, 2014.

[20] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang. A First Look at Cellular Network Performance during Crowded Events. In *ACM Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, 2013.

[21] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang. A First Look at Cellular Machine-to-Machine Traffic - Large Scale Measurement and Characterization. In *ACM Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, 2012.

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

[23] 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.

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

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

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

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

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

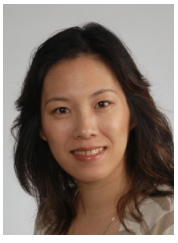
[29] Q. Xu, A. Gerber, Z. M. Mao, J. Pang, and S. Venkataraman. Identifying diverse usage behaviors of smartphone apps. In *ACM IMC*, 2011.

M. Zubair Shafiq is an assistant professor in the Department of Computer Science at the University of Iowa. He received his Ph.D. in Computer Science from Michigan State University in 2014. He received the 2013 Fitch-Beach Outstanding Graduate Research Award at Michigan State University. He received the 2012 IEEE ICNP Best Paper Award. His research interests include networking, security, Internet measurement, and performance evaluation.

Lusheng Ji is a Principal Member of Technical Staff - Research at the AT&T Shannon Laboratory, Florham Park, New Jersey. He received his Ph.D. in Computer Science from the University of Maryland, College Park in 2001. His research interests include wireless networking, mobile computing, wireless sensor networks, and networking security. He is a Senior Member of the IEEE.

Alex X. Liu received his Ph.D. degree from the University of Texas at Austin in 2006. He is an associate professor in the Department of Computer Science and Engineering at Michigan State University. He received the IEEE & IFIP William C. Carter Award in 2004 and an NSF CAREER award in 2009. He received the Withrow Distinguished Scholar Award in 2011 at Michigan State University. His research interests focus on networking, security, and dependable systems.

Jeffrey Pang is a researcher at AT&T Labs - Research. He received his Ph.D. in Computer Science from Carnegie Mellon University in 2009. He currently builds systems to measure and optimize cellular networks. His research interests include networking, mobile systems, distributed systems, and privacy.



Jia Wang received her Ph.D. from Cornell University in January 2001. She is a Principal Technical Staff Member AT&T Labs - Research. Her research interests focus on network measurement and management, network security, performance analysis and troubleshooting, IPTV, and cellular networks. She was co-recipient of the ACM SIGMETRICS 2004 Best Student Paper Award and the ACM CoNext 2011 Best Paper Award.