Mobile Netw Appl
DOI 10.1007/s11036-015-0648-6

Understanding and Predicting Data Hotspots in Cellular Networks

Ana Nika · Asad Ismail · Ben Y. Zhao · Sabrina Gaito · Gian Paolo Rossi · Haitao Zheng

Abstract The unprecedented growth in mobile data usage is posing significant challenges to cellular operators. One key challenge is how to provide quality of service to subscribers when their residing cell is experiencing a significant amount of traffic, *i.e.* becoming a traffic hotspot. In this paper, we perform an empirical study on data hotspots in today's cellular networks using a 9-week cellular dataset with 734K+ users and 5327 cell sites. Our analysis examines in details static and dynamic characteristics, predictability, and causes of data hotspots, and their correlation with call hotspots. We show that using standard machine learning methods, future hotspots can be accurately predicted from past observations. We believe the understanding of these key issues will lead to more efficient and responsive resource management and thus better QoS provision in cellular networks. To the best of our knowledge, our work is the first to empirically characterize traffic hotspots in today's cellular networks.

The final publication is available at http://link.springer.com/article/10.1007/s11036-015-0648-6

A. Nika · A. Ismail · B. Y. Zhao · H. Zheng
Department of Computer Science, University of California, Santa Barbara, Santa Barbara, CA 93106-5110, USA
E-mail: anika@cs.ucsb.edu

A. Ismail
E-mail: asad@cs.ucsb.edu

B. Y. Zhao
E-mail: ravenben@cs.ucsb.edu

H. Zheng
E-mail: htzheng@cs.ucsb.edu

S. Gaito · G. P. Rossi
Università degli Studi di Milano, Milan, Italy
E-mail: gaito@di.unimi.it

G. P. Rossi
E-mail: rossi@di.unimi.it

Keywords Data hotspots · Cellular networks · Machine learning

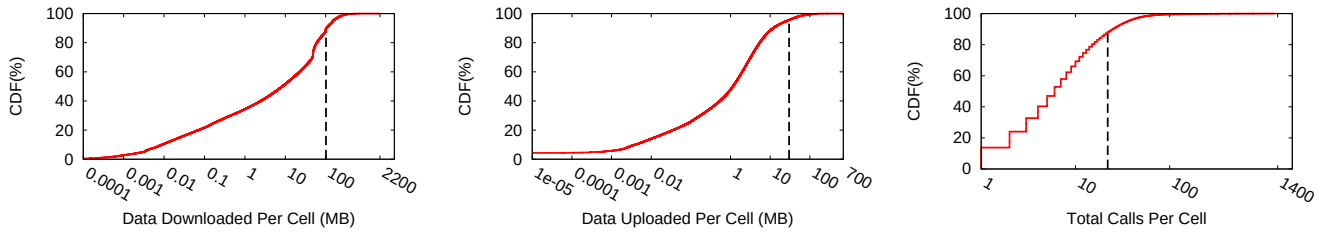
1 Introduction

The proliferation of mobile devices like smartphones and tablets has created an explosion in mobile data traffic. Industry reports predict that mobile data traffic will continue to grow and reach 11.2 exabytes per month by 2017 [1]. Such unprecedented growth poses significant challenges to cellular operators, who must carefully plan network growth and manage network resource usage.

One particular challenge facing cellular networks is how to efficiently handle data hotspots, *i.e.* how to provide reasonable performance to customers when the cell sites they reside in experience significant amount of data traffic. Addressing this challenge requires a concrete understanding of the behavior of data hotspots in today's networks. While previous measurement studies have examined patterns of network-wide traffic [2,3,4,5,6,7,8], none of them has studied the impact of data hotspots in detail.

In this paper, we perform a detailed, empirical study on hotspots in cellular networks using a large-scale dataset from a major European cellular provider. Our dataset is unique because it contains almost 3 months (9 full weeks) of records of user voice and Internet data usage, for 734K+ subscribers and 5327 cell sites in a large metropolitan city. This detailed and large dataset enables us to perform an in-depth study on hotspot behaviors and its evolution dynamics over time.

Specifically, we analyze data usage (and call volume) at both cell and individual user levels, and seek to understand key characteristics, predictability and causes of data hotspots in today's cellular networks, as well as potential correlations between data and call hotspots. We believe understanding these key issues will lead to more efficient and responsive resource management and thus better QoS provi-



(a) Data Download Volume (b) Data Upload Volume (c) # of Calls

Fig. 1 CDF of traffic volume for data download, upload and voice call per 15 minute segment across all the cell sites.

sion in cellular networks. To the best of our knowledge, our work is the first to characterize traffic hotspots in today's cellular networks.

Summary of Key Findings. Our analysis led to the following key findings.

– Hotspots occur randomly across the network. Over time, the large majority (90%) of cell sites have less than 20% of chance to become a hotspot.
– Data hotspots are evenly distributed over hours in a day and days in a week, and do occur during late night and early morning hours. This differs from prior measurement findings where data usage is concentrated during day time hours (8am-8pm) [5]. On the other hand, call hotspots tend to occur during morning and afternoon hours during weekdays, commonly known as the peak voice hours.
– A small portion ($\approx$ 10%) of the 5327 cell sites become hotspots simultaneously. These hotspots are widely spread across the city and do not form large clusters/congested areas. Thus operators can adjust local network resources to efficiently handle these hotspots on the fly.
– Data and call hotspots occur independently. Even within data traffic, download and upload hotspots rarely overlap, indicating that data traffic congestion often happens in one direction rather than in both directions.
– The statistical properties of hotspots remain consistent across the 9 weeks. More importantly, one can leverage observations of hotspots in the current week to reliably predict hotspots in the following week within close vicinity ($\leq$1km) and time segment ($\leq$1 hour) of the current hotspot.
– Deployment of new cell sites brings down traffic volume of neighboring cells, reducing the occurrence of hotspots in local neighborhood.
– While call hotspots are triggered by large user count, data hotspots form when a few "heavy" users start bandwidth-hungry data sessions. By identifying and tracking these heavy users as they move across the city, the network can effectively predict future data hotspots and adapt network resources ahead of time. This can potentially lead to significant improvement in QoS provisioning.

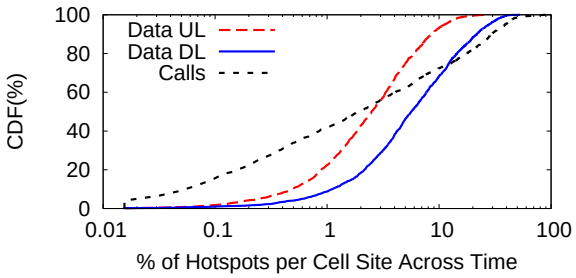


Fig. 2 Frequency of hotspot occurrence across cell sites.

– Using standard machine learning algorithms, one can accurately predict future occurrence of traffic hotspots based on past observations, *e.g.* knowledge of hotspots observed during the previous week.

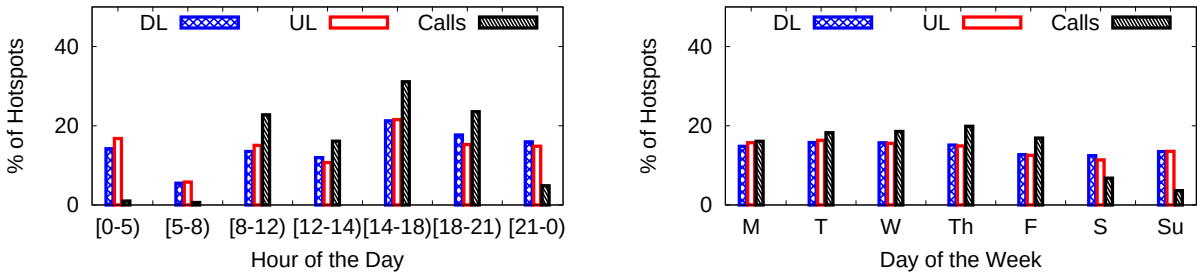
2 Methodology

In this section, we introduce the cellular dataset used in our analysis and our methodology for analyzing traffic hotspots.

Dataset. Our dataset includes records of voice calls and Internet data usage (both download and upload) for 734K+ subscribers in a metropolitan city in Europe. These subscribers belong to a major European cellular operator. The data collection spans 67 days between March and May 2012, and includes data at 5327 cell sites that cover the entire city.

In total, the dataset includes 96M+ voice calls and 61M+ Internet usage sessions. Each voice call record contains the anonymized user identifiers of those initiated and received the call (if the callee belongs to the same operator), the cell IDs where the call started and ended, the time stamp when the call started, and the call duration. Each Internet data session includes the bytes downloaded (or uploaded) by the user, the anonymized user ID, the time stamp, and the cell ID. Finally, each cell ID in our dataset is also associated with the GPS coordinates of the corresponding base station.

Data Analysis. Using real data, we seek to understand the key characteristics, causes, and predictability of data hotspots in cellular networks, as well as their correlation with call hotspots. For this, we analyze the data usage at both cell



(a) Time of the Day (b) Day of the Week

Fig. 3 Temporal distribution of hotspots, for data download, upload and calls.

and user levels. To calculate per-cell data usage, we aggregate downloaded (or uploaded) data traffic (*i.e.* number of bytes) over a period of time for all the users in each single cell. We separate the download and upload scenarios because they have different cell capacities and, thus, hotspot characteristics. We also compute the per-cell call volume as the total number of calls during a time period. For simplicity, we only consider calls during which the user did not change cell, which account for 80% of the total calls.

Our analysis on traffic hotspots consists of four steps. *First*, we identify hotspots and their high-level characteristics, including frequency of a cell being a hotspot, spatial/temporal distribution, and correlation among data and voice call hotspots (see Section 3). *Second*, we explore how hotspots evolve over time and whether they can be reliably predicted from past data (see Section 4). *Third*, we identify key patterns of individual user traffic that lead to hotspots (see Section 5). *Finally*, we apply machine learning approaches to predict the occurrence of traffic hotspots (see Section 6).

3 High-Level Analysis

We begin our analysis by identifying data and voice call hotspots based on their traffic volume. We then present the high-level characteristics of these hotspots.

3.1 Identifying Hotspots

Intuitively, a traffic hotspot occurs when a cell experiences significant traffic, creating concerns on QoS provision. To identify hotspots, a direct solution is to compare the current traffic volume to the cell capacity. Unfortunately, this is infeasible because cellular providers do not reveal any information on capacity.

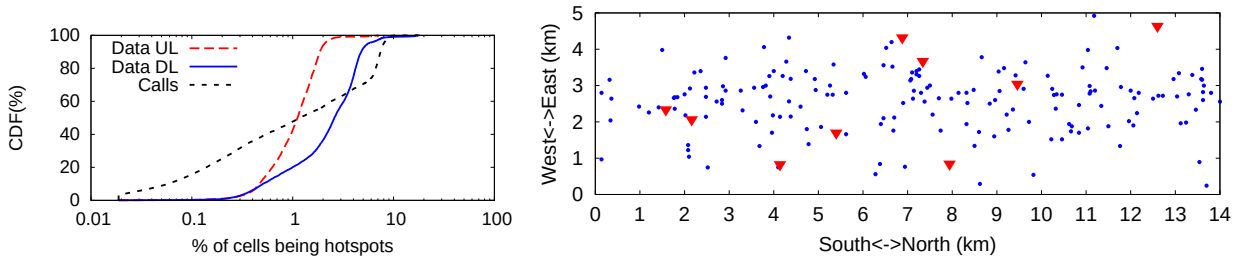
In our work, we seek to identify the set of cell sites and time periods that exhibit very large traffic volume. Naturally, these cells/time periods will contribute to the majority of the traffic across all cell sites during the 67 days. For this reason, we define traffic hotspots as the set of cell sites and time periods where their aggregate traffic volume covers more than 50% of the total traffic in the entire dataset. This definition aligns with a recent work that identifies congested cell sites [9]. Specifically, we divide the 67 days into non-overlapping time slots of X minutes. For each slot we compute the per-cell traffic for all the cell sites, and refer to each cell site/time slot combination as a segment. This produces $5327 \cdot 67 \cdot 24 \cdot 60/X$ segments. We rank these segments by their traffic volume and select the set of heaviest segments whose aggregated traffic exceeds 50% of the total traffic. We have examined X=15, 30, or 60 minutes and they lead to similar results. For brevity, we only show the results of X=15 minutes.

Figure 1 plots the CDF of traffic volume of all the 15-minute segments, while the dotted line marks the start of the heaviest segments which we identified as hotspots. For data download, upload, and voice, we found that 10.5%, 4.4%, and 12.1% of the segments are hotspots, respectively. The download hotspots carry significant traffic, ranging from 102MB to 2154MB per 15 minutes, mapping to a 0.9Mbps-19.1Mbps average download speed. The traffic load in upload hotspots is about one third of download, *i.e.* 30MB-681MB, mapping to an average upload rate of 0.27Mbps-6.1Mbps. Finally, call hotspots have more than 22 calls per 15 minutes but can reach 1301 calls per 15 minutes.

3.2 Key Characteristics

Next we discuss the high-level characteristics of hotspots.

Frequency of Occurrence. Our first question is how frequent a cell becomes a hotspot. Results in Figure 2 show that the frequency varies significantly across cells and is generally small. Consider download that has the highest frequency of occurrence among all three scenarios. We see that 90% of cells have less than 20% of chance to become a hotspot, indicating that hotpots appear randomly across the network. Interestingly, a very small set ($<<$1%) of cells have more than 50% of chance of becoming a hotspot. Thus, the operator can apply special treatment to these cells, *e.g.* allocating extra radio frequency, to improve their QoS provisioning.



(a) Normalized # of hotspots at any time (b) A sample snapshot of hotspots

Fig. 4 Spatial distribution of traffic hotspots. (a) CDF of the number of concurrent hotspots normalized by the total number of cell sites. (b) A sample snapshot of hotspots, where the blue dots represent the cell sites and the red triangles represent the hotspots in the current time segment.

Temporal Distribution. Next we study the temporal distribution of hotspots, *e.g.* the time of the day where hotspots are most likely to appear. For this, we divide the time into 7 periods that are likely to hold different traffic characteristics due to typical user activity pattern: Late Night $[0-5)$, Early Morning $[5-8)$, Morning $[8-12)$, Noon $[12-14)$, Afternoon $[14-18)$, Evening $[18-21)$, and Night $[21-0)$. Figure 3(a) plots the PDF of hotspot temporal statistics for both data (download, upload) and voice. We see that data hotspots are evenly distributed across the 7 time periods, while call hotspots tend to occur during the Morning and Afternoon hours, commonly known as the peak hours for voice calls. This result differs from prior measurement findings [5] that show the users generate most of their traffic between 8am and 8pm. Furthermore, we also study the dependence on the day of the week. Results in Figure 3(b) show that data hotspots are evenly distributed across the entire week, while the number of call hotspots reduces significantly during the weekend.

Spatial Distribution. We categorized the spatial location of hotspots and found that the heaviest ones tend to occur at city downtown, hospitals, universities and soccer stadiums, which is as expected. But more interestingly, a closer look at the hotspot traces shows that multiple hotspots do occur simultaneously in time. Thus, we seek to understand how many hotspots occur simultaneously and whether they are clustered. Intuitively, when a large set of hotspots is clustered into small areas, the network will experience large resource shortage in these congested areas and, thus, degraded QoS performance. But if only a handful of hotspots are widely spread across the city, they can be easily handled by adjusting resources in the local neighborhood.

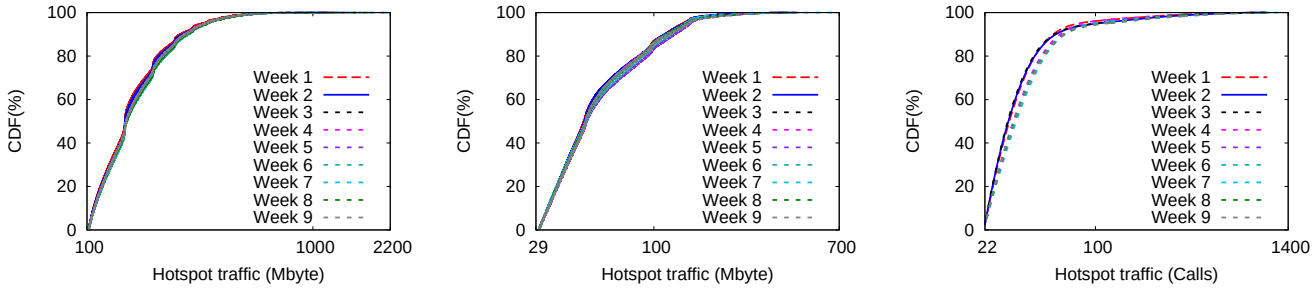
Figure 4(a) plots the CDF of the number of hotspots that occur simultaneously across time, normalized by the total number of cells in the network. We see that only a small portion ($<$10%) of the network becomes congested at any given time. Next, we study the spatial distribution of these concurrent hotspots using the following method. We first calculate the distance between any two cells using the haversine distance formula on their GPS coordinates [10][1]. Given these distances, we apply the complete-linkage hierarchical clustering [11] with a maximum cluster radius of 1km. Since we do not know the radius of each cell site in our dataset, we assume it is between 250m - 1km, the typical values used for cellular cells in urban areas [12]. And as a result we select the cluster radius of 1km (or 2km diameter). The clustering results show that the median cluster size is mostly 1. For downloads and calls the median cluster size can reach 6 and 7 respectively for very few time intervals. On the other hand, the maximum cluster size varies significantly between 1-37 for downloads and calls, and 1-19 for uploads. Overall, we can conclude that at any given time, traffic hotspots are well isolated from each other. This can also be seen from a sample snapshot of the hotspots at a randomly chosen time interval (Figure 4(b)).

Correlation among Data and Voice Hotspots. Finally, we seek to understand how often download and upload hotspots overlap, *i.e.* a single cell becomes a download hotspot and a upload hotspot simultaneously. For this we compute the ratio of the number of overlapping download and upload hotspots and the total count of download and upload hotspots over the 67 days. We also repeat the same for download and voice hotspots. Our results show that the ratio is 14% and 13%, respectively, implying that download, upload, and call hotspots in general occur independently.

4 Hotspot Evolution

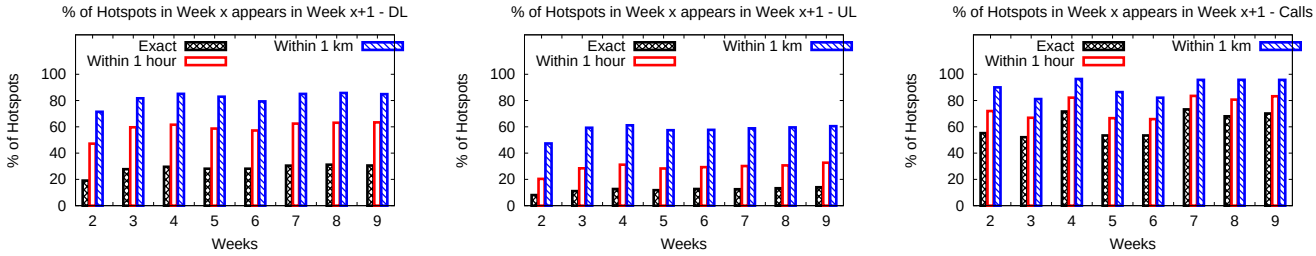
We now turn our attention to the evolution of hotspots in the cellular network. Using our 9-week dataset, we seek to understand three key issues: (1) *hotspot dynamics*, *i.e.* do hotspot characteristics change over time; (2) *predictability*, *i.e.* given knowledge of hotspot occurrence in the past week (or month), can we predict its future appearance? (3) *impact of new cell deployments*, *i.e.* will adding new cell sites trigger significant changes in hotspot behaviors?

[1] The use of haversine distance helps to cluster cells separated by a small physical distance.



(a) Download (b) Upload (c) Calls

Fig. 5 The statistical distribution of hotspot traffic volume remains similar across the 9 weeks.



(a) Download (b) Upload (c) Calls

Fig. 6 Predictability of hotspots using observations from the previous week. We consider three prediction criteria: exact time/exact location, exact time/within 1km, and within 1 hour/exact location of the hotspot observed from the previous week.

4.1 Time Dynamics

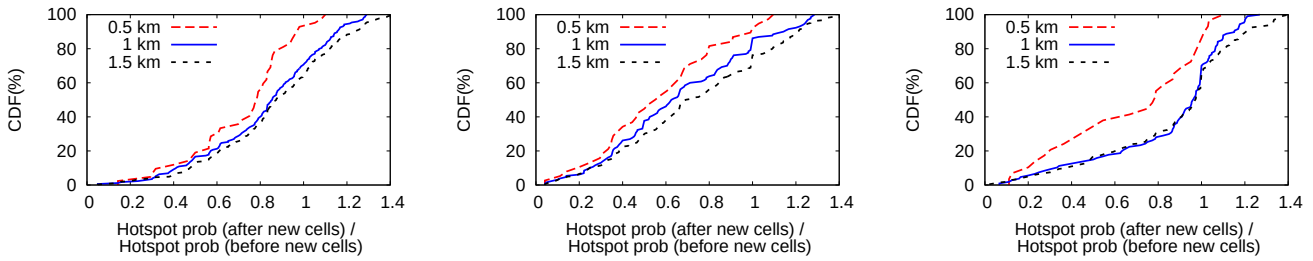
We examined in detail the per-week hotspot characteristics, and found that they stay consistent across the 9 weeks. Specifically, our analysis led to the following findings. *First*, both data and voice hotspots were evenly distributed over time. In terms of data download (upload), 7% (8%) of hotspots appeared during the 1st and 2nd week, respectively, while each of the remaining 7 weeks has 11% of the total hotspots. The call hotspots follow a similar pattern. *Second*, for both data and voice scenarios, the distribution of hotspot traffic volume is consistent across the 9 weeks, as shown in Figure 5. *Third*, the temporal and spatial distributions of hotspots are also consistent over time. This is not surprising and shows that the cellular network is in a stable condition over these 9 weeks.

4.2 Predictability

The above observation motivates us to investigate whether one can predict hotspot occurrence based on past observations. For this, we start from examining how likely a hotspot appeared in week X will again appear at the exact same time and location in week X+1. Figures 6 plots this probability for download, upload and voice hotspots for $X = 2, ..., 8$. We see that the predictability is low for data hotspots, 19%-31% for download and 8%-14% for upload, but moderate for voice hotspots (55%-70%). This aligns with the results in Figure 3(a) where the majority of voice calls take place during "peak" hours.

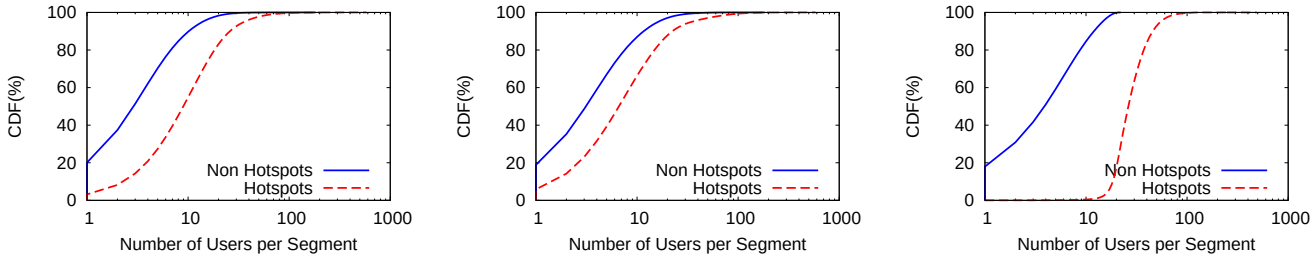
Next, we relax the predictability criterion by examining how likely a hotspot in week X will reappear in week $X+1$ but within 1 hour of original time segment. This improves the predictability largely to 63% for download, 32% for upload and 83% for voice hotspots. As comparison, we then relax the hotspot location criterion to within 1km of the original cell site (*i.e.* the direct neighbor of the original cell) but not the time criterion, the predictability increases to 84% for downloads, 60% for uploads, and 95% for call hotspots. Finally, when we relax both time and location criteria to within 1 hour and 1km, the predictability jumps to 83% for uploads, 96% for downloads, and 98% for calls.

While the above analysis examines predictability using data from the previous week, we also repeated the experiments using data from the previous month. We did not cross-compare the same day of each month, as they may fall in different days of the week. Instead, we compared the first Monday of the first month with the first Monday of the second month and so forth. The resulting predictability is slightly worse than the week-based prediction. Specifically, the predictability is 28%, 58%, and 80% for download, 11%, 27%, 55% for upload, and 61%, 71%, 88% for voice hotspots when we use the aforementioned three different criteria (exact, within 1 hour, within 1km). This shows that hotspots do consistently reappear over time but display some short-term fluctuations. So a week-by-week based prediction is likely to be more accurate than a month-by-month based prediction.



(a) Download (b) Upload (c) Calls

Fig. 7 Impact of new cell deployment on neighboring cells, in terms of CDF of the ratio of the probability of being a hotspot in the week before the deployment and in the week after the deployment. The neighboring cells are those within 0.5km, 1km or 1.5km from the new cells.



(a) Download (b) Upload (c) Calls

Fig. 8 Number of users per hotspot and non-hotspot segment for downloads, uploads, and calls.

Later in Section 6 we show that standard machine learning algorithms can lead to accurate prediction of future traffic hotspots.

4.3 Impact of Adding New Cells

One unique feature of our dataset is that it also captures events of adding new cell sites to the network. Over the 9 weeks, we observed a total of 20 new cell sites. This led us to an interesting question: *does the hotspot (or network) behavior change in the area where new cells are introduced*? Intuitively, adding new cells (and thus extra capacity) will reduce traffic loads among neighboring cells, potentially dissolving existing hotspots.

To answer this question, we examine the neighboring cells around each newly added cell i over the course of one week before and one week after i's deployment. Here we define cell i's "neighboring cells" as those located within distance x, and we vary the value of x between 0.5km and 1.5km to understand the impact range. For a given x, we compute for each neighboring cell of i, the ratio of its frequency of being a hotspot in the week after i's deployment and the frequency in the week before i's deployment. Figure 7 plots the CDF of this ratio for download, upload and call scenarios. We see that among neighboring cells in close proximity (x=0.5km), 90%+ of them had less frequency of becoming hotspots after new cell deployments. The impact reduces at cells further away (based on the results of x=1km and 1.5km). This confirms that adding new cells does relieve network traffic congestion, leading to better user performance.

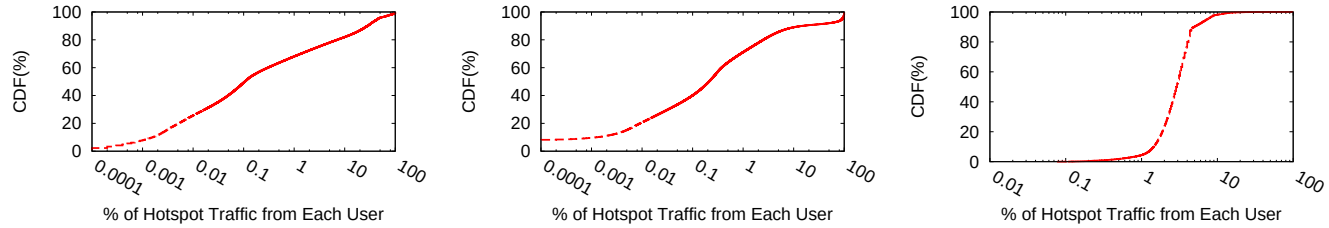
5 Causes of Traffic Hotspots

Finally, we study in detail the user activities to understand the causes of traffic hotspots. Intuitively, a data hotspot appears when either many users download/upload data, or some "heavy users" start bandwidth-hungry applications like video streaming. Thus we seek to understand the dominating cause for the hotspots in our dataset.

# of Users per Hotspot. Figure 8 plots the CDF of the user count per hotspot as well as those of non-hotspot segments. For both download and upload hotspots, it varies largely between 1 and 582. Yet more than 80% of hotspots have less than 20 users, 50% of them have less than 10 users, and only 5% of them have 500+ users. This indicates that data hotspots were mostly caused by a small set of "heavy" users. Furthermore, the user count distribution/range is similar across hotspots and non-hotspots, which means that we cannot simply use user count to predict future hotspots.

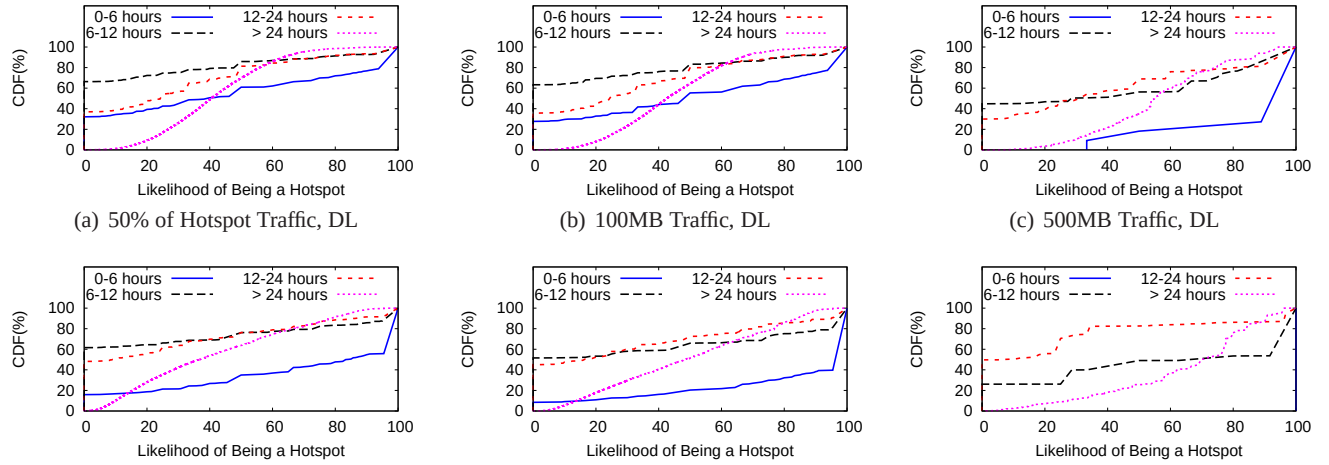
In contrary, call hotspots display a different pattern compared to call non-hotspots – the user count in call hotspots is always above 10 while the mass majority of non-hotspots has less than 10 users. This implies that the user count can be a reliable indicator of call hotspots.

Traffic Volume per User. We also look into the traffic volume contributed by individual users inside each hotspot. Figure 9 plots the CDF of traffic volume of each individual



(a) Download (b) Upload (c) Calls

Fig. 9 CDF of individual user's contribution to hotspot traffic.



(a) 50% of Hotspot Traffic, DL (b) 100MB Traffic, DL (c) 500MB Traffic, DL

(d) 50% of Hotspot Traffic, UL (e) 50MB Traffic, UL (f) 200MB Traffic, UL

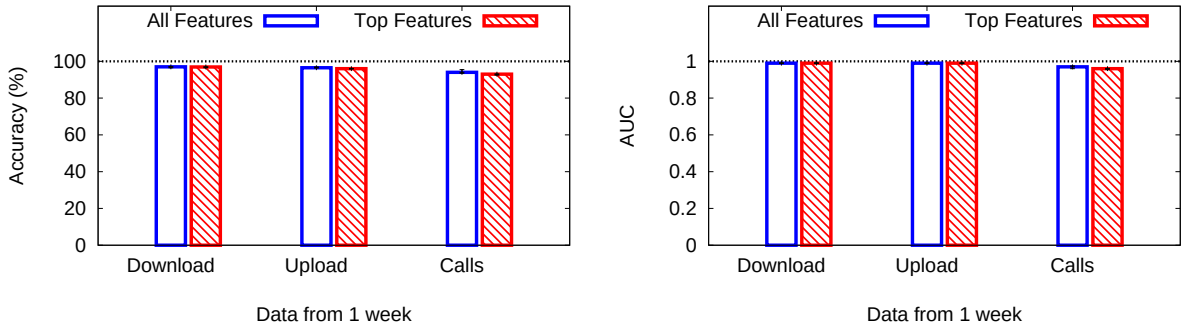
Fig. 10 The likelihood of a cell site becoming a hotspot when being visited by a heavy user within 6 hours, 12 hours, 24 hours after the heavy user's first appearance in a hotspot, or no time limit.

user normalized by the total traffic of her current hotspot. Clearly there is a large skew across the users. For download hotspots, 20% of the users contribute to 90% of the traffic. The contribution of these top 20% users increases to 99% for upload hotspots and 95% for voice hotspots. From the traces, we even observe that certain individual users did access 1GB data or make 180+ calls over 15 minutes.

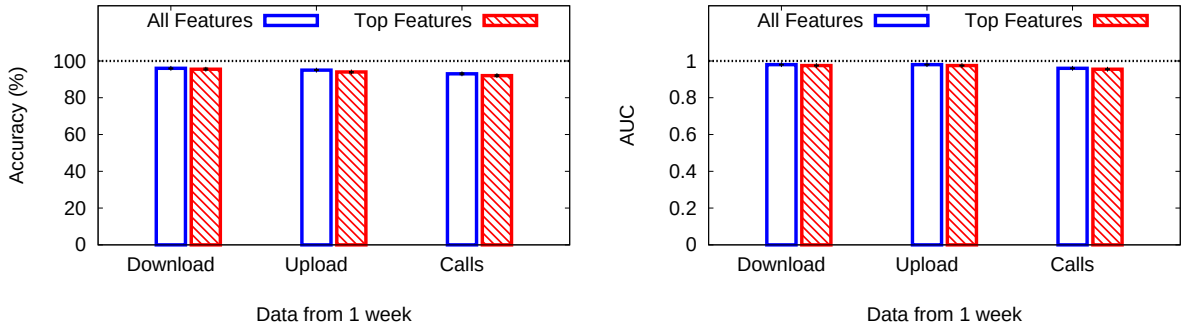
Benefits of Tracking Heavy Users. The above results show that hotspot traffic is dominated by a very small set of "heavy" users. This leads to a natural question: *Can we track these heavy users as they move around the network, and use their trajectory to predict future hotspots*? To validate this hypothesis, we study the chance of a cell visited by a heavy user becoming a hotspot. For this we first define a heavy user as a user producing more than X% of traffic in a current hotspot, or when her raw data traffic goes beyond M bytes. We first locate a user satisfying one of these criteria and then track the cells she visits after leaving the current hotspot. We also segment the tracking period to 0-6 hours, 6-12 hours, 12-24 hours, >24 hours after she leaves the current hotspot. This allows us to identify the benefit of tracking a heavy user in short- and long-term. Note that as the tracking period shifts, the number of cells a heavy user visits also changes.

Figure 10 shows the results for X=50%, Y=100MB or 500MB for download and Y=50MB or 200MB for upload. We did not show any result for X >50% because they lead to similar conclusions. We see that the cells visited by a heavy user during the first 6 hours (after she leaves the initial hotspot) are very likely to become a hotspot. This is particularly true when the user carries massive traffic *i.e.* 500MB for download or 200MB for upload. After that, the probability drops moderately for download and sharply for upload. Furthermore, we also examine the number of heavy users in each hotspot that depends on the above criteria. Interestingly, for all three criteria, each hotspot has only one heavy user if there is any. Specifically, when using X=50%, there are 53K+ (download) or 26K+ (upload) heavy users, which provide a good coverage on the network cells. But for Y= 500MB (download) or Y=200MB (upload), there are only roughly 4K heavy users, which provide less coverage but more accurate prediction results.

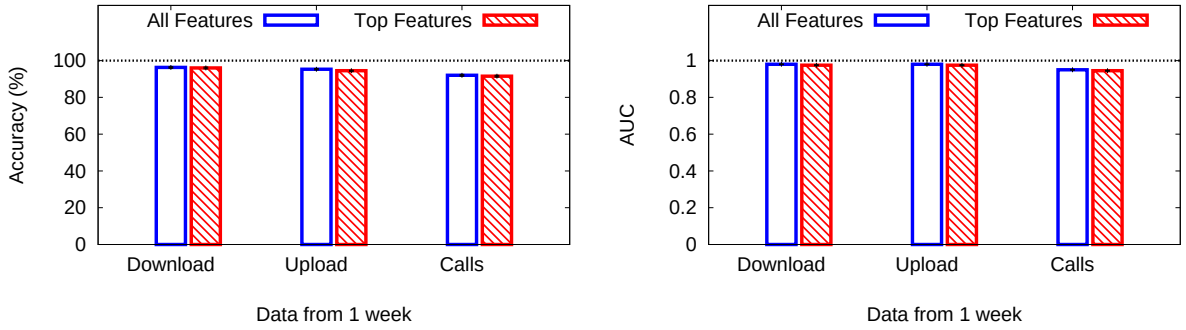
Overall, the results indicate that tracking heavy users for another 6 hours after they appear in a hotspot would help predict future hotspots.



(a) Predicting Hotspots vs. non Hotspots (Support Vector Machine)



(b) Predicting Hotspots vs. non Hotspots (Random Forests)



(c) Predicting Hotspots vs. non Hotspots (Bayesian Networks)

Fig. 11 Prediction result using Support Vector Machine, Random Forests, and Bayesian Networks. The model performance is evaluated by accuracy (left) and Area under ROC curve (right) .

6 Predicting Traffic Hotspots

The observations in Section 4 motivate us to investigate whether one can predict future hotspot occurrence. In this section, we experiment with machine learning (ML) classifiers to determine whether we can accurately predict traffic hotspots using knowledge of hotspots during the previous week. More specifically, we seek to answer four key questions:

- Can we achieve accurate weekly forecast on traffic hotspots, *i.e.* predicting traffic hotspots in the current week using knowledge of hotspots in the previous week?
- Can we predict traffic hotspots in multiple weeks using knowledge from a single week?
- Which ML models produce the most accurate prediction?
- What are the most essential features for predicting traffic hotspots?

We take three steps to answer the above questions. First, we collect a set of features based on cell site activities during each week of our dataset. Second, we use these features to build machine learning classifiers for predicting hotspots. Finally, we run feature selection to determine the top features that provide the best prediction on future hotspots.

Rank	Prediction Categories		
	Download	Upload	Calls
1	Basic Hotspots-F4 (0.91)	Basic Hotspots-F4 (0.91)	User Features - F14 (0.88)
2	Traffic Features-F9 (0.91)	Traffic Features-F9 (0.90)	Basic Hotspots-F4 (0.85)
3	Traffic Features-F10 (0.89)	Traffic Features-F10 (0.88)	Traffic Features-F9 (0.80)
4	User Features-F14 (0.66)	User Features-F14 (0.62)	User Features -F13 (0.74)
5	User Features-F15 (0.43)	User Features-F15 (0.49)	User Features-F15 (0.60)
6	Basic Hotspots-F3 (0.36)	Basic Hotspots-F3 (0.20)	Basic Hotspots-F3 (0.30)
7	User Features -F13 (0.32)	User Features -F13 (0.19)	Traffic Features-F8 (0.25)
8	Traffic Features -F8 (0.30)	Traffic Features -F8 (0.17)	Traffic Features-F10 (0.20)

Table 1 The top 8 features and their categories ranked by information gain (values shown in parentheses).

6.1 Background in Machine Learning Prediction & Classifiers

Machine learning constructs efficient and accurate algorithms that learn from and make predictions on data [17]. Machine learning algorithms usually build a model from input data to make data-driven predictions. The input data is associated with a set of features or attributes, often represented as a vector. Our work uses supervised learning algorithms that assign labels to the input data and show in which category each data belongs to (e.g., hotspot or non-hotspot). The data is partitioned into two sets, training and testing, and the first set is used to train a learning algorithm while the second to evaluate the performance of it. Machine learning algorithms that can be used to predict labels (previously unknown) for test data are called classifiers. In our work, we apply multiple classifiers to predict if cell sites (represented as a set of features) will become hotspots or not.

Using the implementation in WEKA [13] with default parameters, we build multiple machine learning classifiers including Support Vector Machine (SVM), Random Forests (RF) and Bayesian Networks (BN). SVM [14] is a supervised learning algorithm that, given a set of instances that belong to one of two categories, builds a model by assigning new instances into one category or the other. A SVM model is a representation of the instances as points in space so that the instances of each category are clearly separated as much as possible. More formally, a SVM constructs a hyperplane or set of hyperplanes that can be used for classification. The best hyperplane is the one that provides the largest margin between the two categories. RF [15] uses multiple decision trees at training time and outputs as the final predicted category for each instance the mode of the categories that are the outputs of individual decision trees. And BN captures the probabilistic relationship between a set of random variables.

To build a training set for our classifiers, we randomly sample the same number of hotspots and non-hotspots at each 15 minute interval for all the days of the week. We create a dataset of 70K hotspots and 70K non-hotspots for each week of our dataset.

6.2 Features

Based on our analysis on traffic hotspots, we explore different classes of features (a total of 18) to profile cell site behaviors during each week.

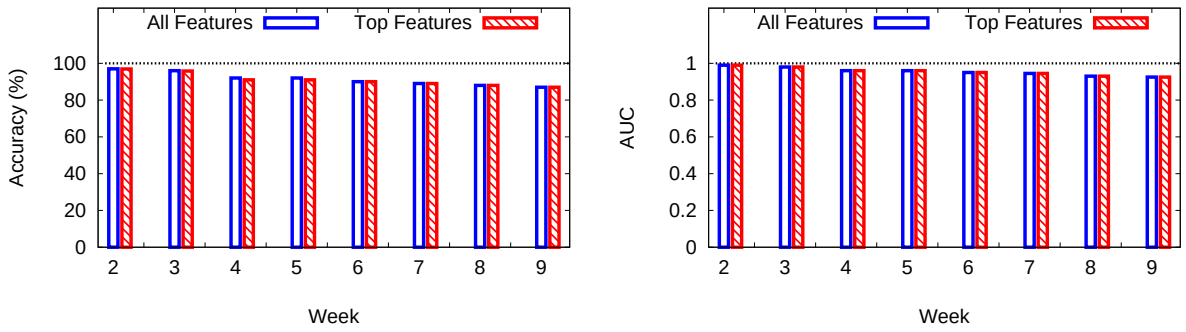
- *Basic Hotspot Features (F1-F5)* which include number of hotspots, number of non-hotspots, number of hotspots/non-hotspots at the same time interval for all the days of previous week, indication if the cell site was hotspots/non-hotspots during the exact time in the previous week.
- *Traffic Features (F6-F10)* which include average number of bytes/calls in hotspots, average number of bytes/calls in non-hotspots, average number of bytes/calls when the cell was hotspot/non-hotspot during the same time interval for all the days in the previous week, number of bytes/calls during the same time interval in the previous week.
- *User Features (F11-F18)* which include average number of users in hotspot, average number of users in non-hotspots, average number of users when the cell was hotspot/non-hotspot at the same time interval for all the days in the previous week, number of users at the same time interval in the previous week, average number of heavy users, average number of heavy users at the same time interval for all the days of the previous week, number of heavy users at the same time interval in the previous week.

6.3 Experimental Results

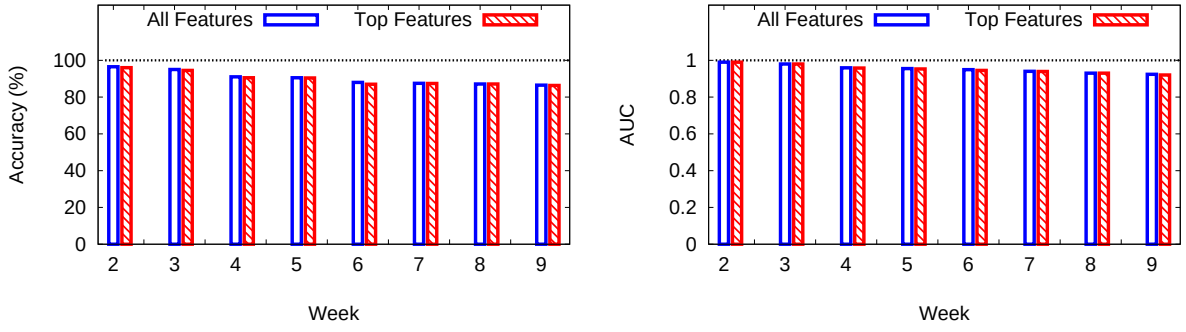
Using these classifiers, we created different experiments to understand the predictability of hotspots. For each experiment, we run 10-fold cross validation and report the results in terms of classification accuracy and area under ROC curve (AUC). Accuracy refers to the ratio of correctly predicted instances over all instances. AUC is another widely used metric and higher AUC indicates stronger prediction power. For example, AUC > 0.5 means the prediction is better than random guessing.

We now present the results of our experiments.

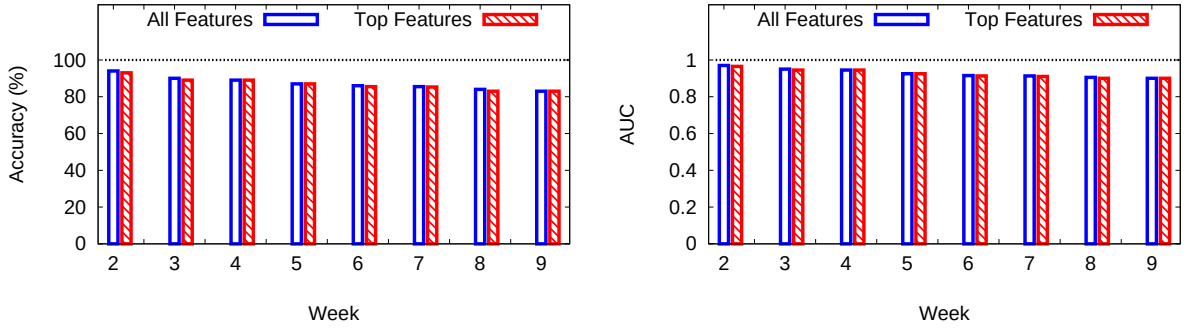
Accuracy of Weekly Forecast. Our first experiment uses data from the previous week to predict hotspots of the cur-



(a) Download



(b) Upload



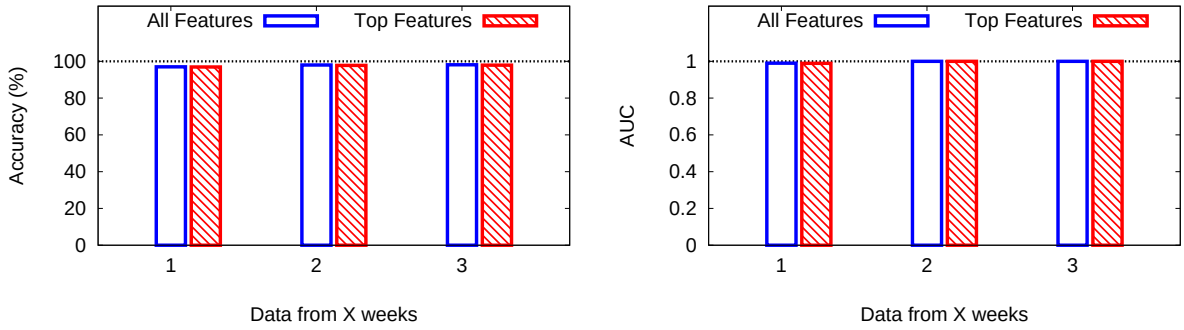
(c) Calls

Fig. 12 Prediction results for Support Vector Machine when using only the first week of data to predict the following weeks.

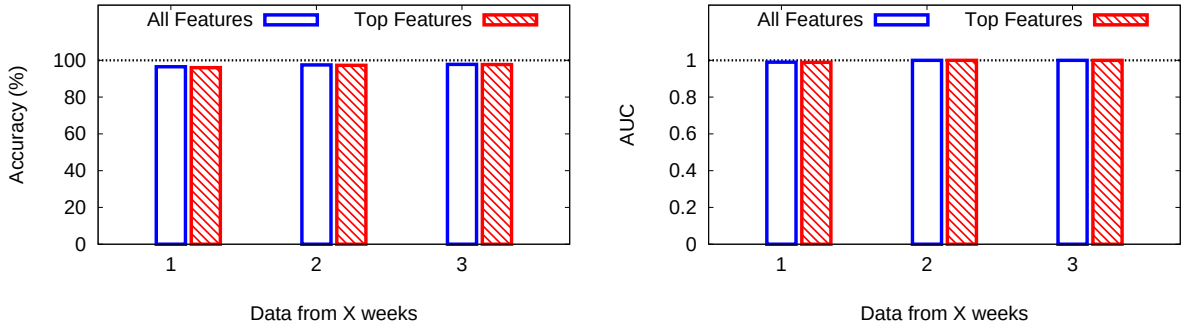
rent week. The results of this experiment for download, upload, and calls are shown in Figure 11. The accuracy and AUC are the average of 8 weeks as the results for each week are similar to each other. We see that the accuracy and AUC are consistently high across the three classifiers and the three traffic scenarios. This confirms the strong weekly predictability observed in Section 4.2.

Top Features. We perform feature selection on the eighteen features described earlier. Specifically, we rank features based on their *Information Gain* [16], which measures each feature's distinguishing power over the two classes of data. We list the top 8 features in Table 1. As expected, prediction power varies significantly, and information gain drops quickly after the top 4 features. To validate their prediction power, we repeat each ML experiment with only the top 4 features. The results in Figure 11 show that using just the top 4 features performs similarly to the full classifiers, but with much less complexity.

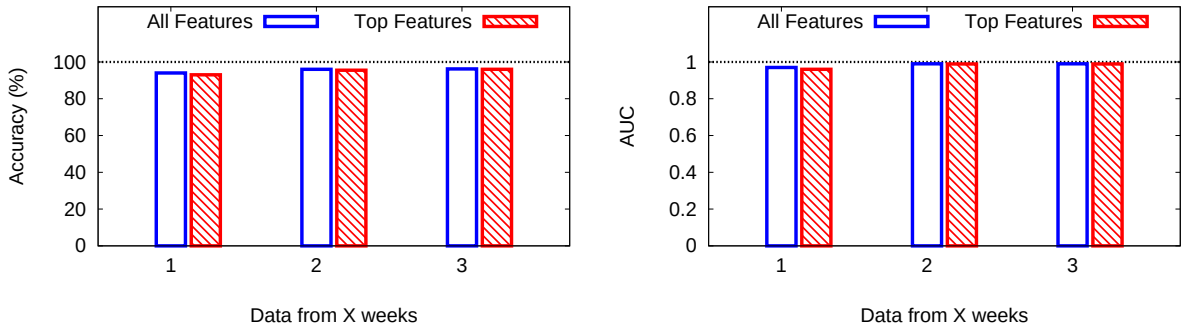
Next we take a closer look at the top features. We see that the download and upload classifiers rely on different order of features compared to the call classifier. The call classifier relies heavily on the user features, while the upload and download classifiers are driven by the basic hotspot and traffic features. This aligns with our earlier analysis, where we show that call hotspots occur mostly when a large number of users make calls. On the other hand download and upload hotspots occur due to sudden increase in data transfer by heavy users.



(a) Download



(b) Upload



(c) Calls

Fig. 13 Prediction results for Support Vector Machine when using X weeks of data (where X = 1, 2 or 3) to predict the following week.

Predicting Multiple Weeks Ahead. We now explore whether one can predict data hotspots in the following 8 weeks using observations from the first week. Figure 12 shows the accuracy and AUC results for these 8 weeks using the SVM classifier for download, upload, and calls. The results for the other two classifiers are similar and omitted for brevity. As expected, accuracy and AUC decrease slightly with the week index, from 97%, 96%, and 94% down to 87%, 86%, and 83% for downloads, uploads, and calls respectively. This means that hotspots remain predictable even when using staled data. On the other hand, since hotspot characteristics can change week by week, to achieve high predictability hotspots must be estimated using data in close range.

Predictability Limit. Finally, we seek to understand the amount of data (in terms of the number of weeks) required to achieve the highest accuracy and AUC when predicting hotspots for the following week. Figure 13 shows the accuracy and AUC results when using 1-3 weeks of data to build the features for the SVM classifier. We omit the results for the other two classifiers because they are similar. Interestingly, the results indicate that using more than one week, *e.g.* two weeks, of data achieves higher accuracy, *i.e.* 98% compared to 97% for downloads. But the gain quickly saturates at three weeks. Overall, these results confirm that traffic hotspots can be accurately predicted using weekly forecast.

7 Related Work

Recently, there have been various measurement campaigns and studies on cellular network traffic, characterizing network performance, capacity, usage patterns, and subscriber behaviors (*e.g.* [2,4,5,3,6,7,8]). Most of them use measurement data with limited scale in space or time, *e.g.* one week of cellular usage data.

Specifically, the work in [5] analyzes subscriber behavior and the impact on data traffic, by monitoring subscribers' mobility and temporal activity patterns and correlating them with the data consumption. A follow-up work extends the analysis to spatial correlation of radio resource usage by clustering base stations based on their RF resource usage and producing spatially connected clusters. Another work [6] applies such spatial correlation to a fine-grained application-level analysis, characterizing geospatial dynamics of different applications. The work of [7] performs a similar fine-grained analysis on applications, adding the dimension of device type as well as proposing both Zipf-like and Markov models to capture geospatial dynamics.

Researchers have developed usage prediction models. The work in [8] proposes a usage prediction model for text, calls and data based on entropy theory. The model applies to general traffic across the cellular network rather than hotspots. A more recent work applies spatial-temporal throughput analysis on cellular traffic, proposing various traffic models for both weekdays and weekends. Finally, the work of [4] develops a traffic model driven by user behaviors.

Our work differs from these existing studies by analyzing traffic hotspots and their evolution in time. Our dataset is unique in that it contains 9-weeks of cellular usage in terms of both data download/upload and voice calls. This allows us to examine hotspot characteristics in long-term (over multiple weeks rather than a single week), and to cross-compare data and voice hotspots.

8 Conclusion

We study data hotspots in today's cellular networks using a 9-week city-wide dataset from a major cellular operator in Europe. Our analysis led to several key findings on data hotspots, including random distribution across space and time, independency between data and voice hotspots as well as independency between data download and upload hotspots, predictability using observations from prior week and by tracking heavy users over time. From these observations we also identify potential solutions to reliably identify/predict traffic hotspots prior to their arrival, thus enabling network operators to take efficient and responsive actions, *e.g.* adjusting local network resources in real-time, to effectively address hotspots and maintain quality of service to their customers.

References

1. Cisco Visual Networking Index, `http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html`
2. Nan, E., Chu, X., Guo, W., Zhang, J.: User data traffic analysis for 3G cellular networks. In: Proc. of CHINACOM, pp. 468–472 (2013)
3. Paul, U., Subramanian, A., Buddhikot, M., Das, S.: Understanding spatial relationships in resource usage in cellular data networks. In: Proc. of INFOCOM Workshops, pp. 244–249 (2012)
4. Laner, M., Svoboda, P., Schwarz, S., Rupp, M., Users in cells: A data traffic analysis. In: Proc. of WCNC, pp. 3063–3068 (2012)
5. Paul, U., Subramanian, A., Buddhikot, M., Das, S.: Understanding traffic dynamics in cellular data networks. In: Proc. of INFOCOM, pp. 882–890 (2011)
6. Shafiq, M. Z., Ji, L., Liu, A. Z., Pang, J., Wang, J.,: Characterizing geospatial dynamics of application usage in a 3G cellular data network. In: Proc. of INFOCOM, pp. 1341–1349 (2012)
7. Shafiq, M. Z., Ji, L., Liu, A. X., Wang, J.: Characterizing and modeling Internet traffic dynamics of cellular devices. In Proc. of SIGMETRICS, pp. 305–316 (2011)
8. Zhou, X., Zhao, Z., Li, R., Zhou, Y., Zhang, H.: The predictability of cellular networks traffic. In: Proc. of ISCIT, pp. 973–978 (2012)
9. Mobile broadband with HSPA and LTE - capacity and cost aspects, `http://www.developingtelecoms.com/business/opinion/123-white-papers-case-studies/3211-mobile-broadband-with-hspa-and-lte-capacity-and-cost-aspects.html`
10. Sinnott, R. W.: Virtues of the haversine. In: Sky and Telescope, vol. 68(2), pp. 159 (1984)
11. Defays, D.: An efficient algorithm for a complete link method. In: The Computer Journal (British Computer Society), vol.20(4), pp. 364–366 (1977)
12. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A.: Identifying Important Places in People's Lives from Cellular Network Data. In: Proc. of 9th International Conference on Pervasive Computing, pp. 133–151 (2011)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, R., Witten, I. H.: The WEKA data mining software: an update. In: SIGKDD Explor. Newsl., vol.11(1), pp. 10–18 (2009)
14. Cortes, C., Vapnik, V.: Support-Vector Networks. In: Machine Learning, vol.20(3), pp. 273–297 (1995)
15. Breiman, L.: Random Forests. In: Machine Learning, vol. 45(1), pp. 5–32 (2001)
16. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. In: Journal of Machine Learning Research, vol.3, pp. 1157–1182 (2003)
17. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. In: The MIT Press (2012)