Inferring Content-Centric Traffic for Opportunistic Networking from Geo-location Social Networks

Pavlos Sermpezis
Mobile Communications Dept.
EURECOM,France
pavlos.sermpezis@eurecom.fr

Thrasyvoulos Spyropoulos
Mobile Communications Dept.
EURECOM,France
thrasyvoulos.spyropoulos@eurecom.fr

***Abstract*—Opportunistic networking has been proposed to support a number of novel applications, like content sharing or mobile data offloading, that follow a content-centric communication model, i.e., many users are interested in the same content. Users' traffic demand patterns can crucially affect the performance of such applications, but our knowledge about the characteristics of content demand is limited. Nevertheless, opportunistic networking is known to exhibit strong locality and social characteristics. For this reason, in this paper we argue that some initial insights about opportunistic traffic patterns could be inferred from geo-social network data. In particular, we study the check-in patterns of users in datasets of two real Location-Based Social Networks, towards understanding potential traffic characteristics and implications for opportunistic networking.**

I. Introduction

The recent growth in the number of mobile devices with rich communication and storage resources enables new ways of mobile networking. Mobile devices (e.g., smartphones, laptops) can directly exchange data using short-range communication technologies (e.g., Bluetooth, WiFi Direct) when they are in proximity. Many novel applications, based on such *opportunistic* communications, have been recently proposed, e.g., content sharing [1], [2], [3], collaborative sensing and computing [4], offloading of cellular networks [5], [6], [7], etc.

A common trend among these applications is that they follow a content-centric communication model, where many users are interested in the same content, or in contents belonging to the same category. Example cases are the scenario where a user creates a content and distributes it locally to (and through) neighboring devices (*floating content*) [1], [3], or the scenario where a cellular network offloads mobile traffic by distributing popular content to a few relay users which opportunistically forward it to any other interested user (*mobile data offloading*) [5], [6], [7].

A crucial factor for the performance, or even the feasibility, of opportunistic content-centric applications is the number of users that are involved in the data dissemination process and their traffic demand [1], [6], [7]. When the density of users willing to participate in the opportunistic network is low, a content might not be able to disseminate [1], [7]. Moreover, when a content is not *popular*, i.e., only a few users are interested in it, its delivery through opportunistic communications might be inefficient or unnecessary [6], [7]. As a result, the knowledge of the (i) *user density*, (ii) *content popularity* and (iii) *data demand* patterns (or a correct estimation of them) under different settings is of high importance.

978-1-4799-8461-9/15/$31.00

However, up to now, there is a lack of data about such traffic patterns, since large-scale opportunistic networks have not been deployed and the only experience we have is of small, experimental settings (e.g., see [8]), which are not representative examples of real traffic. Hence, previous studies have mainly inferred traffic patterns based on statistics from other communication paradigms, like the Web [9], peer-to-peer or cellular networks [10], [11]. For instance, a common assumption is that content popularities follow a Zipf-law distribution, similarly to the webpage requests [9].

Despite the resemblance between the mechanisms of novel (opportunistic) and traditional (P2P, Web, etc.) content-centric applications, the respective networking paradigms differ in a number of dimensions, which might lead to different traffic patterns as well. In particular, the inherent *locality* of opportunistic networks (data dissemination in long distances usually comes with large delays) [1], [3] and their *content-centric* applications (e.g., offloading mobile data from overloaded base stations) [5], [6], [7], is a determinant/key factor for the traffic to be exchanged between users.

To this end, in this paper, we infer traffic statistics for opportunistic applications from Location-Based Social Networks (LBSNs). Check-in data from LBSNs might be useful sources of information because they connect the *location* of a user and her *context* (hence, what type of content the potential opportunistic node might be interested in). Therefore, we look at user check-ins in two large LBSN datasets, and their relative popularity as an indirect measure of potential demand for content, and we consider two main categories:
(a) Individual *venue popularity* as a fine-grain indication of popularity related to *local content*.
(b) Venue *category popularity* as an indication of popularity of *types of contents*.

Although, a check-in does not necessarily imply a content request, correlation between the presence of users in certain locations and their traffic demand (volume and type) has been supported by a number of studies [11], [12], [13]. For instance, [11] shows that the location affects the applications accessed by users, [12] demonstrates variations in the traffic depending upon the geo-location of users, while [13] builds a mobile cloud caching system based on the observation that for many mobile applications the specific data that is accessed depends on the current location of the user. The intuition from these studies, coupled with the insights on the opportunistic

communication mechanisms, suggest that the qualitative characteristics or relative statistics of traffic demand is probable to bear a resemblance to the corresponding characteristics of user check-ins. Thus, by analyzing LBSN check-ins, our goal in this paper is to obtain understanding and try to answer interesting questions related to content-centric traffic, like

– *Which distributions better describe the content popularity?*
– *What are some typical values of their parameters?*
– *How traffic intensity might differ in various locations?*
– *How does it change over time in a certain location?*

The rest of this paper is organized as follows. After describing the datasets (Section II), we extract statistics from which we infer traffic patterns that relate to content popularity (Section III), and time varying characteristics of data demand (Section IV). In each section, we first provide the methodology for the data processing, and we conclude it by providing directions for how our results could be used for the evaluation and design of opportunistic networking applications.

II. Datasets

Location-Based Social Networks (LBSNs) have recently become very popular among mobile phone users and businesses, as they offer a new way of social networking and new advertising possibilities. In a LBSN, users, using a mobile or web application, post their presence at a *venue* ("check-in"), where a venue can be a place (e.g., airport), a business (e.g., restaurant), etc., registered in the venue database. Furthermore, it is common for users to associate their LBSN account to their accounts in other online social networks (OSNs), like Facebook or Twitter, in order to notify / share their check-in activity with their social connections (friends, followers, etc.).

In this study, we analyze large datasets from two LBSNs, namely the Foursquare and the Gowalla networks. The two datasets were collected and published by Yanhua Li *et al.* [14] and Theus Hossmann *et al.* [15], respectively. In the remainder, we briefly describe the main characteristics of the datasets and the information we use in our analysis. We refer the interested reader to the initial publications [14], [15] for a complete presentation and a detailed description of the datasets.

Foursquare dataset [14]
Foursquare[1] is a web and mobile application that allows registered users to post their location at a venue ("check-in") and connect with friends. The dataset we analyze (Yanhua Li *et al.* [14]), contains information about 2.4 *million* venues, from 14 geographic regions all over the world, during a period of two months (May-June) in 2012. Data are organized as a list of venues including the following information: (i) *location* of the venues, (ii) *category* they belong to (e.g., bar, gym, theater), and (iii) number of Foursquare users that have visited them (*#users*), number of check-ins (*#checkins*) and the number of "tips" users left during the data collection period.

Gowalla dataset [15]
Gowalla was[2] a location-based social network, where users were able to check-in to close-by venues (e.g., restaurants, office buildings, shops, etc.) through their mobile phones. We analyze a dataset (Hossmann *et al.* [15]) of $350,000$ users that checked-in 27 *million* times in 2.5 *million* different venues all around the world in the period Jan. 2009 - July 2011. Data are organized as a list of "check-ins". A "check-in" logs the (i) *location* of the venue, (ii) the *context* of the venue (i.e., the *category* it belongs), and (iii) the *time* of the check-in.

For ease of reference, in Table I, we present the (subset of the) data attributes used in our analysis for both datasets.

TABLE I: Data attributes in Foursquare and Gowalla datasets.

Foursquare (list of venues)	Venue ID	Location (City)	Category	#users	#checkins
Gowalla (list of checkins)	Venue ID	Location (City)	Category	Time	Check-in ID

III. Content Popularity

In this section, we focus on *content popularity patterns* that might appear in an opportunistic networking application (Section III-A). We present a number of popularity-related statistics, and how they vary along different system dimensions, like the type of content-centric applications (location-based or context-based) and the network size (Sections III-B and III-C). Then, we summarize our findings and discuss some important implications (Section III-D).

A. Inferring Content Popularity from Check-ins

To infer content popularity statistics, we analyze the corresponding statistics of (a) check-ins in different venues and (b) check-ins related to different categories/contexts. As discussed earlier, the correlation between *check-ins* and *content requests* might be significant in opportunistic applications. Thus, we also expect similarities between the relative statistics (e.g., distribution type, coefficient of variation) of these two quantities[3].

***Location-based statistics**:* In a number of opportunistic applications users request contents related to the area they reside [1], [2], [3], [13], resembling thus communication in LBSNs. For instance, contents might be a piece of data corresponding to a map, road traffic or local event notification, etc. With respect to our datasets, a user check-in at a venue, indicates some interest of the user in the certain venue/location. Hence, users checking in the same venues are probable to be also interested in a content related to this location [2], [3], [13]. This does not of course mean that these nodes only care about content related to this location. It simply suggests that the larger the number of check-ins the higher the potential traffic demand for local content. To this end, we use the popularity of a venue as an indicator for the popularity of a content related to the venue location.

Our datasets contain information about how many nodes have checked-in at a venue and how many times, which, equivalently, denote the popularity of a venue. Therefore, as a first step towards calculating the popularity statistics, we find for each venue the number of total users (*#users*) that checked in it and the total number of check-ins (*#checkins*)[4].

[1] https://foursquare.com/

[2] Gowalla was launched in 2007 and closed in 2012.

[3] In contrast, the quantitative characteristics or absolute statistical values, like mean value or variance, might differ among these two metrics (check-ins and content requests), due to their different nature.

[4] The two metrics (number of users and check-ins) are used as two different indicators of venue popularity [14].

In the Foursquare dataset this information is already available, whereas for the Gowalla dataset, we calculate it by aggregating the individual check-ins in each venue.

Context-based statistics: In content-centric applications, contents corresponding to a certain context/category are disseminated to interested users through opportunistic communication. Some examples could be news or trending videos belonging to a certain category, delivered through publish-subscribe or mobile data offloading mechanisms, etc. [5], [6], [7]. Content popularity plays a crucial role for the performance of such context-based dissemination mechanisms [6], and an a-priori knowledge (or estimation) of the popularity patterns could lead to a better system design [6], [7].

Since each venue in our datasets belongs to a *category*, we can infer statistics for the content popularity by the number of users that have checked in (or the number of check-ins) at a venue belonging to different categories. To this end, we group the venues per category and then sum the number of users/check-ins of all the venues that belong to each category. This, gives the category popularity and we use it as an indication for the content popularity.

B. Location-Based Statistics

In this section, we first consider the aggregate statistics of venue/content popularity over the whole datasets, and then perform a *per city* analysis, where we calculate the statistics separately for venues located at different cities. This analysis allows us to reveal possible similarities and differences appearing in networks with large or small sizes (aggregate and per-city statistics, respectively).

Aggregate Statistics. For each venue in the datasets, we calculate the total number of users that have checked in (*#users*) and the total number of check-ins (*#checkins*). Then we calculate the *experimental Complementary Cumulative Distribution Function* (eCCDF) of *#users* and *#checkins* over all venues, where the CCDF of a random variable x is

$$\overline{F}_x(t) = P\{x > t\} \tag{1}$$

For instance, if the random variable x is the number of users per venue (*#users*), and only one third of the venues have more than u users checked in them, then we denote $\overline{F}_{\text{\#users}}(u) = \frac{1}{3}$.

In Fig. 1 we present the eCCDF of the *#checkins*[5] per venue in the (a) Foursquare and (b) Gowalla datasets. The first observation is that the venue popularity follows a *power-law* distribution (i.e., a straight line in a log-log plot), with a slightly faster decrease of the *tail* (i.e., right part of the plot, corresponding to high popularity values) in the Foursquare dataset.

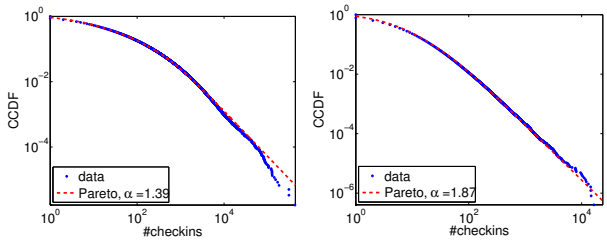
This observation, leads us to fit experimental data (i.e., the eCCDFs) with a well-known power-law distribution, the *generalized Pareto distribution*[6], whose CCDF is given by

$$\overline{F}_x(t) = P\{x > t\} = \left(1 + \tfrac{1}{\alpha} \cdot \tfrac{t-\theta}{\sigma}\right)^{-\alpha} \tag{2}$$

where the exponent α is the *shape* parameter, and σ and θ are the *scale* and *threshold* parameters, respectively.

[5]Using the #users as the popularity metric, gives similar results.

[6]We tried to fit the data with other types of distributions (Gamma, Weibull, etc.) as well. However, the fitting curves deviated more from the eCCDF curve.



(a) Foursquare (b) Gowalla

Fig. 1: Popularity distribution (#checkins) of venues in the (a) Foursquare and (b) Gowalla datasets.

Since we are interested more in the *qualitative characteristics* of the popularity distributions, in the remainder, we give emphasis to the *shape* parameter α of the fitting distributions, rather than the scale and threshold parameters, which are related more to their quantitative characteristics. Low values of α (e.g., around or less than 1) denote a skewed popularity distribution, whereas large values denote a distribution with its mass being more concentrated. Hence, from Fig. 1, we can see that the popularity distribution in the Gowalla dataset ($\alpha \approx 1.9$) is less skewed than in the Foursquare case ($\alpha \approx 1.4$).

Per City Statistics. We now consider the venues located in each city separately, and perform the same analysis, to investigate if the previous conclusions hold also in smaller (city-scale) networks. To avoid statistical errors due to small samples, we preprocessed the Foursquare dataset and considered only the 113 cities with more than 1000 venues. In the Gowalla dataset, we analysed the data of the 31 cities with the highest number of users.

We calculated the eCCDF of the data (but only for the venues located in a given city). The tails of the data experimental distributions follow a power-law as well; however, they decrease a little faster than the generalized Pareto distributions we fitted to them. Due to the large number of cities, we cannot present the detailed plots for all of them. Instead, we present in boxplots the values of the shape parameter α of the fitting Pareto distributions, in Fig 2.

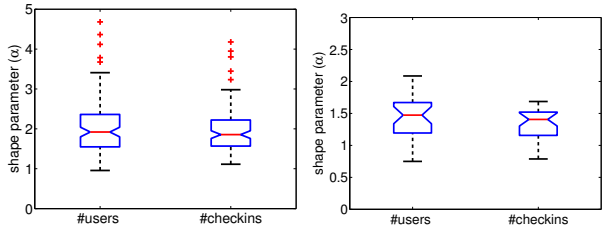
For both datasets, the left boxplots correspond to the case where popularity is inferred from the number of users (*#users*) and the right boxplots to case where it is inferred from the number of check-ins (*#checkins*). In the Foursquare dataset (Fig 2(a)) the values of α lie in the range $\alpha \in [1, 3.5]$, whereas in the Gowalla dataset (Fig 2(b)) they span a smaller range (due to the lower number of cities considered) from values less than 1 up to 2.

C. Context-Based Statistics

We now analyse the data towards calculating context-based statistics. To consider different network sizes, we first group all the venues in the dataset according to the category they belong, and then, we proceed similarly for the venues located in each city separately.

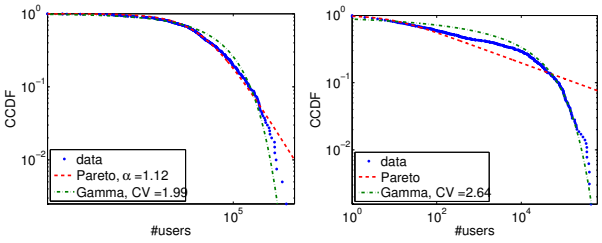
Aggregate Statistics. In Fig. 3 we present the eCCDF of the total number of users[7] per category (i.e., those who checked

[7]Using the #checkins as the popularity metric, gives similar results.



(a) Foursquare (b) Gowalla

Fig. 2: Distribution (presented in boxplots) of the shape parameter α of the generalized Pareto distributions fitted to the venue popularity in different cities.



(a) Foursquare (b) Gowalla

Fig. 3: Popularity distribution of contexts (i.e., set of venues belonging to a certain category) in the (a) Foursquare and (b) Gowalla datasets.

in venues with the same category attribute value) in the (a) Foursquare and (b) Gowalla datasets, along with two fitting distributions: a generalized Pareto distribution (see Eq. (2)), and a *gamma distribution*, whose CCDF is given by

$$\overline{F}_x(t) = P\{x > t\} = 1 - \frac{1}{\Gamma(\alpha)} \cdot \gamma(\alpha, \beta \cdot t) \qquad (3)$$

where α and β, are the *shape* and *rate* parameters, and $\Gamma(\cdot)$ and $\gamma(\cdot)$ are the *gamma function* and the *lower incomplete gamma function*, respectively. A gamma distribution can also be expressed via its mean value μ and coefficient of variance $CV = \frac{\sigma}{\mu}$ (σ is the variance), which relate to α and β as:

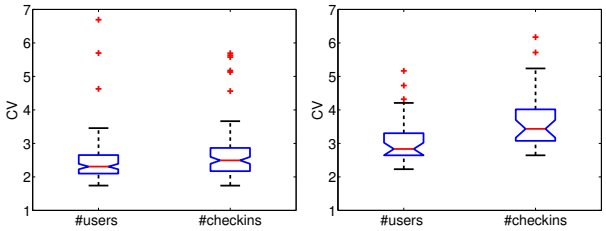
$$\mu = \frac{\alpha}{\beta} \qquad \text{and} \qquad CV = \frac{1}{\sqrt{\alpha}} \qquad (4)$$

Since we are interested in the variability (skewness) of content popularity, we focus only on the *shape* parameter α, or, equivalently, on the coefficient of variation[8] CV.

What can be observed in Fig. 3, is that the tails of the data distributions (eCCDF) decrease faster than these of the corresponding power-law distributions. In this case a *gamma distribution*, which has an exponential decrease of the tail, could better capture the tail of the eCCDF. Nevertheless, in the case of Foursquare dataset, a generalized Pareto distribution could be also used as an approximation for the main body of the real data distribution. Moreover, comparing the results for the context-based popularity (Fig. 3) with the corresponding location-based statistics (Section III-B), it is evident that, in general, the former are less skewed (mainly, due to the aggregation/grouping of venues in categories).

Per City Statistics. As previously, we analysed the statistics in each city separately by grouping venues belonging to the same

[8]The CV serves also as an indication for the heterogeneity of a popularity distribution.



(a) Foursquare (b) Gowalla

Fig. 4: Distribution (presented in boxplots) of the coefficient of variation CV of the gamma distributions fitted to the context popularity in different cities.

category. For all cities in both datasets, we observed that the tails of the data distributions (eCCDF) decrease exponentially, and, therefore a gamma distribution could better capture the context-based popularity patterns. In Fig. 4 we present the distribution of the values (in boxplots) for the coefficients of variation (CV) in each scenario. It is evident that the heterogeneity (indicated by the value of CV) of the context popularity varies, but, in general, it is large (i.e., larger than 1) for the majority of the scenarios.

D. Conclusions and Implications

Our findings show that popularity patterns may vary when considering different types of applications and network sizes. In Table II, we summarize the observations, show what kind of fitting distributions better approximate the content popularity in each scenario, and present the ranges of values of the shape parameters (α and CV for the Pareto and gamma distributions, respectively) we observed in the data. These statistics could be used as realistic example cases in simulation/analytic studies for, e.g., performance prediction of content-centric schemes [6], design of caching policies [5], or network dimensioning [7]. Depending on the scenario considered, e.g., location-related content, context-based content, large networks, small networks, large areas, small areas, etc., one might need to select different statistics, as Table II suggests.

TABLE II: Content Popularity Statistics in Various Scenarios

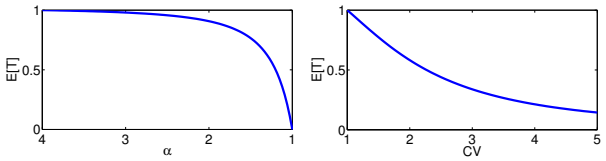
Content Type	Network Size	Pareto	α	Gamma	CV
Location	Large	✓	[1.2 , 1.8]	✗	-
Context	Large	✓	[1.1 , 1.2]	✓	[2 , 2.8]
Location	City	✓	[0.75 , 3.5]	✗	-
Context	City	✗	-	✓	[1.75 , 5]

To further demonstrate how our findings can be used, we consider the following example: the expected delivery delay of a content $E[T]$, under a single-hop content dissemination scheme and under the optimal content placement policy (see [6, Result 5]), is given by [6]

$$E[T] = c \cdot \frac{\left(E_p[\sqrt{x}]\right)^2}{E_p[x]} \qquad (5)$$

where c is a constant depending on nodes' mobility and storing capacity, and $E_p[\cdot]$ denotes an expectation taken over the content popularity distribution (x denotes the random variable).

We now consider two cases: (C1) a location-based and (C2) a context-based application. As suggested by Table II, content



(a) case (C1) (b) case (C2)

Fig. 5: Expected delivery delay $E[T]$ (normalized) vs. increasing heterogeneity of content popularity, i.e., (a) decreasing α for case (C1) and (b) increasing CV for case (C2).

popularity is captured better by a Pareto distribution[9] (with parameters $\{\alpha, \theta\}$) for case (C1), and a gamma distribution (with parameters $\{\mu, CV\}$) for (C2). Then, the expected delivery delay for the cases (C1) and (C2), is given by (see Eq. (5))

$$E[T^{(\texttt{C1})}] = c \cdot \frac{\alpha \cdot (\alpha-1)}{(\alpha-0.5)^2} \quad \text{and} \quad E[T^{(\texttt{C2})}] = c \cdot \left(\frac{\Gamma\left(\frac{1}{CV^2}+0.5\right)}{\frac{1}{CV} \cdot \Gamma\left(\frac{1}{CV^2}\right)} \right)^2$$

From the above expressions, one can see how $E[T]$ changes for *increasing heterogeneity* of content popularity (note: heterogeneity increases when (C1) the shape parameter α decreases, or (C2) the CV increases). To demonstrate this more clearly, in Fig. 5 we plot such changes. Two main observations are: (i) in both cases $E[T]$ decreases as heterogeneity increases; (ii) the effect of heterogeneity in the two cases (that correspond to location- and context- based application types) is different, in the first case the curve of $E[T]$ is concave, while in the second case the curve is convex.

IV. Time Variations of Traffic Demand

In this section we study the variations over time (over long and short term time scales) of the content traffic (inferred from check-ins), focusing on characteristics that relate to opportunistic networking. We analyse the data only for the Gowalla network, since the Foursquare dataset does not contain time information for individual check-ins.

We denote the set of the check-ins at venue i as $\mathbf{C}_i$. Then, we define the traffic demand related to venue i at time t as

$$n_i(t) = \sum_{c \in \mathbf{C}_i} 1_{\{c \in [t, t+\tau)\}} \tag{6}$$

where τ is a positive constant (i.e., the *time granularity*) and $1_{c \in [t,t+\tau)}$ is 1 if the check-in c took place during the time interval $[t, t+\tau)$; otherwise it is 0. The above definition says that a check-in during $[t, t+\tau)$ implies an increment in the traffic demand related to the given venue.

To be able to derive useful conclusions for a wide range of settings, including both short-term and long-term variations of traffic demand, a fine time granularity is needed when calculating the values $n_i(t)$[10]. To achieve this, we should select a small value of τ. For opportunistic networking applications (e.g., content sharing [2], mobile data offloading [5]), time granularity choices could be from a few minutes to a few hours. However, due to the sparseness (in time) of the check-ins in our dataset, we need to preprocess the data in order to extract useful information and fit the popularity functions $n_i(t)$:
(1) We consider the 40 most popular venues (in #checkins).
(2) We select a fine time granularity τ, equal to 1 *minute*.
(3) For each venue, and for each minute of an 24 *hour* interval (i.e., a whole day; in total $60 \cdot 24$ minutes), we aggregate the observations over all days, i.e., (for $t = 0, 1, ..., 60 \cdot 24$)

$$\hat{n}_i(t) = \sum_{day} n_i(day + t) = \sum_{day} \sum_{c \in \mathbf{C}_i} 1_{\{c \in [day+t, day+t+\tau)\}}$$

A. Long-term variability

We first, investigate how traffic demand changes throughout a day. Fig. 6 shows the (relative) traffic demand time variations in 4 different venues[11]. To smooth the data ($\hat{n}_i(t)$) we used a *moving average filter* with span $T = 60$ minutes. Specifically, the presented curves correspond to the values

$$n_i^{'}(t) = \frac{1}{T} \sum_{k=0}^{T-1} \hat{n}_i(t-k) \tag{7}$$

where $T = \min\{t, 60\}$ and $t = 0, 1, ..., 60 \cdot 24$.

Some main observations (with respect to opportunistic content-centric communications) from the plots we present here and the whole dataset (40 venues), are the following:

- In all venues, popularity becomes zero only for one period T_0 per day. The values of T_0 vary from a few hours, as in the *SFO San Francisco International* (Fig. 6(a)), to almost half a day, e.g., *Austin Convention Center* (Fig 6(b)).
- For the rest of the day (i.e., $t \notin T_0$), the popularity in many venues remains relatively stable, while other venues experience 1 to 3 high popularity periods.
- The highest variability of popularity is observed in a few venues, where the popularity doubles its value (to its maximum value) or becomes zero (from its max value), in a period of 1 to 2 hours. E.g., in the *Stockholm Centralstation* (Fig 6(c)) and the *Epcot* (Fig 6(d)), between $13h00$ and $15h00$.

The above observations could be useful in a number of content-centric applications. For example, in the information sharing application of [1], a content "floats" if the density of users remains above a threshold. Hence, for locations corresponding to the first observation, we can infer that one needs to re-inject the content *only once per day* (specifically after the end of the period T_0), since there is only one period where user density decreases below a given threshold. Moreover, the third observation indicates that in some locations there would be a fast increase of traffic demand. This could possibly overload (locally) the cellular network, and lead to a need for an enhancing offloading mechanism, e.g., [5], [7], whose design/dimensioning could benefit from the knowledge of the traffic demand curves of Fig. 6, see, e.g., [7].

B. Short-term variability

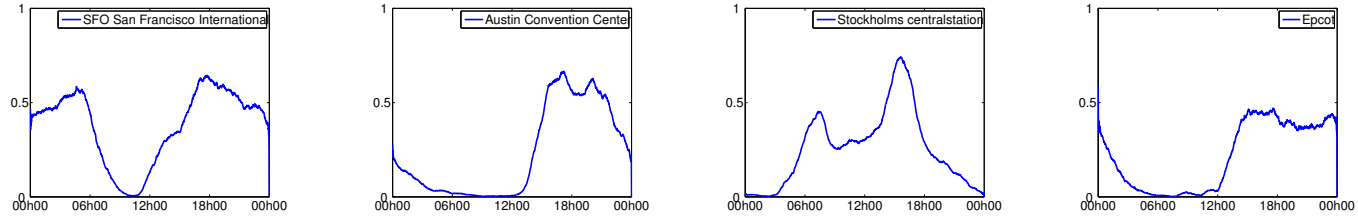
In some applications, contents might have short lifetimes, e.g., a few minutes. In these cases, the analysis of the previous section for long-term variations (span $T = 60$, i.e., 1 hour) might not be adequate to describe the desired characteristics. To this end, in Fig. 7 we present results for short-time traffic demand variability (fluctuations) in the different venues.

We consider different values of span T for the function $n_i^{'}(t)$ (see Eq. (7)) in order to capture different time-scales of

[9] The generalized Pareto distribution for $\sigma = \frac{\theta}{\alpha}$, is equivalent to the Pareto distribution.

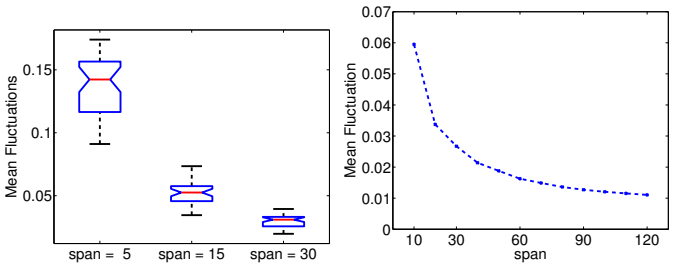
[10] The fine time granularity is used to detect short-term changes, and changes for larger time-scales can then be captured by using a moving average filter.

[11] We present plots for a representative subset of the 40 most popular venues.



(a) (b) (c) (d)

Fig. 6: Time variations of traffic demand throughout a day in 4 popular venues.



(a) (b)

Fig. 7: (a) Boxplot of *mean values of the popularity fluctuations* ($\overline{\Delta n_i'}$) in different venues. (b) Mean values of the *mean values of the popularity fluctuations* ($\overline{\Delta n_i'}$) in different venues, i.e., $E[\overline{\Delta n_i'}]$.

variability. For each value of T, the short-term traffic demand variability (or fluctuation) is calculated as

$$\Delta n_i'(t) = \frac{|n_i'(t+1) - n_i'(t)|}{n_i'(t)} \quad , \; t = 0, 1, ..., 60 \cdot 24 \qquad (8)$$

In Fig. 7(a) we present the mean values of the popularity fluctuations in each venue for $T = 5, 15, 30$ *minutes*, i.e.,

$$\overline{\Delta n_i'} = \frac{1}{60 \cdot 24} \cdot \sum_{t=0}^{60 \cdot 24} \Delta n_i'(t) \qquad (9)$$

The boxplots in Fig. 7(a) correspond to the distribution of the mean values $\overline{\Delta n_i'}$ for $i = 1, ..., 40$ (i.e., over all the 40 venues). In Fig. 7(b) we present similar results for different values of spans ($T \in [10, ..., 120]$ $min.$). Specifically, we plot the mean values of the mean traffic demand fluctuations, i.e.,

$$E[\overline{\Delta n_i'}] = \frac{1}{40} \cdot \sum_{i=1}^{40} \overline{\Delta n_i'} \qquad (10)$$

Some main observations for the short-term traffic demand are:

- As expected, for larger span T, the traffic demand fluctuation decreases. This indicates that the parameters of a content-centric system (e.g., mobile data offloading [5], [7]) would change less frequently when large delays (i.e., T) are tolerated, resulting to a less frequent need for re-tunning, etc.
- For moderate short-time variability (i.e., span $T = 15$ or 30 *min.*), the fluctuations are very small (less than 8%). For very short-time variability (i.e., span $T = 5$ min.), the fluctuations are higher than before, but still less than 20%. These suggest a (relatively) smooth change in the performance of content-centric mechanisms even under short lifetimes.

V. Conclusion

In this work, we analysed two large datasets of check-ins in LBSNs. Motivated by the correlation between user check-ins and content traffic demand in location-based and context-based opportunistic applications, we inferred statistics for content popularity patterns and time variations of content traffic demand. The analysis and presentation of the results were oriented on opportunistic networking, and we believe that the conclusions and implications we provided will be useful for evaluating existing solutions, as well as for future research.

References

[1] J. Ott, E. Hyytiä, P. Lassila, J. Kangasharju, and S. Santra, "Floating content for probabilistic information sharing," *Pervasive Mob. Comput.*, vol. 7, no. 6, pp. 671–689, Dec. 2011.

[2] J. Ott and J. Kangasharju, "Opportunistic content sharing applications," in *Proc. ACM NoM Workshop*, 2012.

[3] S. Gaonkar, J. Li, R. R. Choudhury, L. Cox, and A. Schmidt, "Micro-blog: Sharing and querying content through mobile phones and social participation," in *Proc. ACM MobiSys*, 2008.

[4] M. Conti, S. Giordano, M. May, and A. Passarella, "From opportunistic networks to opportunistic computing," *Communications Magazine, IEEE*, vol. 48, no. 9, pp. 126–139, sept. 2010.

[5] Y. Li, M. Qian, D. Jin, P. Hui, Z. Wang, and S. Chen, "Multiple mobile data offloading through disruption tolerant networks," *IEEE Transactions on Mobile Computing*, no. PrePrints, 2013.

[6] P. Sermpezis and T. Spyropoulos, "Not all content is created equal: Effect of popularity and availability for content-centric opportunistic networking," in *Proc. ACM MobiHoc*, 2014.

[7] P. Sermpezis, L. Vigneri, and T. Spyropoulos, "Offloading on the edge: Analysis and optimization of local data storage and offloading in HetNets," Tech. Rep. http://arxiv.org/abs/1503.00648, 2015.

[8] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Trans. on Mobile Computing*, vol. 6, no. 6, 2007.

[9] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *Proc. IEEE INFOCOM*, 1999.

[10] Z. Li, J. Lin, M.-I. Akodjenou, G. Xie, M. A. Kaafar, Y. Jin, and G. Peng, "Watching videos from everywhere: A study of the PPTV mobile VoD system," in *Proc. ACM IMC*, 2012.

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

[12] M. Shafiq, L. Ji, A. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3G cellular data network," in *Proc. IEEE INFOCOM*, 2012.

[13] P. Stuedi, I. Mohomed, and D. Terry, "Wherestore: Location-based data storage for mobile devices interacting with the cloud," in *Proc. ACM MCS Workshop*, 2010.

[14] Y. Li, M. Steiner, L. Wang, Z.-L. Zhang, and J. Bao, "Exploring venue popularity in Foursquare," in *Proc. IEEE INFOCOM Workshops (NetSciCom)*, 2013.

[15] T. Hossmann, G. Nomikos, T. Spyropoulos, and F. Legendre, "Collection and analysis of multi-dimensional network data for opportunistic networking research," *Comput. Communications*, vol. 35, no. 13, 2012.