Author(s): Jo, Hang-Hyun & Karsai, Márton & Karikoski, Juuso & Kaski, Kimmo

Title: Spatiotemporal correlations of handset-based service usages

Year: 2012

Version: Final published version

Please cite the original version:

Jo, Hang-Hyun & Karsai, Márton & Karikoski, Juuso & Kaski, Kimmo. 2012. Spatiotemporal correlations of handset-based service usages. EPJ Data Science. Volume 1, Issue 1. ISSN 2193-1127 (printed). DOI: 10.1140/epjds10.

Jo et al. *EPJ Data Science* 2012, **1**:10
http://www.epjdatascience.com/content/1/1/10

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Spatiotemporal correlations of handset-based service usages

Hang-Hyun Jo[1*], Márton Karsai[1], Juuso Karikoski[2] and Kimmo Kaski[1]

*Correspondence:
hang-hyun.jo@aalto.fi
[1]Department of Biomedical Engineering and Computational Science, School of Science, Aalto University, P.O. Box 12200, Espoo, Finland
Full list of author information is available at the end of the article

Abstract

We study spatiotemporal correlations and temporal diversities of handset-based service usages by analyzing a dataset that includes detailed information about locations and service usages of 124 users over 16 months. By constructing the spatiotemporal trajectories of the users we detect several meaningful places or contexts for each one of them and show how the context affects the service usage patterns. We find that temporal patterns of service usages are bound to the typical weekly cycles of humans, yet they show maximal activities at different times. We first discuss their temporal correlations and then investigate the time-ordering behavior of communication services like calls being followed by the non-communication services like applications. We also find that the behavioral overlap network based on the clustering of temporal patterns is comparable to the communication network of users. Our approach provides a useful framework for handset-based data analysis and helps us to understand the complexities of information and communications technology enabled human behavior.

1 Introduction

Understanding macroscopic socio-economic phenomena of a large number of individuals has been extensively studied by means of social, physical, and computational sciences [1–3]. Recent access to large-scale digital datasets on human dynamics and social interaction has enabled us to quantitatively investigate the structure and dynamics of human communication networks. Indeed, researchers have studied various datasets, ranging from email and mobile phone communications to social network services, *e.g.* Twitter and Facebook [4–11]. Mobile phones or handsets are now actively utilized to accurately measure or sense human behavior because the handsets equipped with a variety of sensors, including GPS and WiFi, are carried around by the users everyday and all day through. Highly resolved location data collected from handsets have been recently used to uncover human mobility patterns [12–20]. The reliability of data collected from handsets, *i.e.* 'behavioral' data, was tested in the serial studies conducted within the frame of MIT's Reality Mining project [17, 18, 21]. It was shown that the behavioral data are at least comparable to self-report survey data in terms of friendship network and even capturing information that self-reports are missing [18].

The handset usage patterns are known to be diverse among users when measured by the number or duration of the phone sessions and by the amount of data received, to name a few [22, 23]. Within the individual handset usage patterns, temporal inhomogeneities due

to circadian and weekly cycles were also reported [10], which are in close relation to the spatial inhomogeneities, such as nighttime at home and daytime in office. Therefore, for conducting a comprehensive study, it is important to identify the context characterizing the situation of handset user, and then to understand how the context affects service usage patterns [23–27]. However, it is only very recently when the effect of context on the handset-based service usages was investigated. But so far the analysis has been conducted mostly at the aggregate level, while the temporal diversities of service usage among users have been ignored [27].

In this paper, we study spatiotemporal correlations of the service usage patterns of individual users by analyzing a handset-based dataset. This dataset was collected from 124 users' handsets for over 16 months as a part of the OtaSizzle project at Aalto University, Finland [28]. A software installed on handsets collected information about the handset's locations and usages of various services, including web domain visits, applications, emails, voice calls, and short message services, with the resolution of seconds in time and mobile network base stations spatially. After constructing spatiotemporal trajectories of the users we identify several contexts that are meaningful to them by using the context detection method [26]. Other methods include, for example, places of interest or meaningful locations [29, 30] and eigenmode analysis [31–33]. Then, we find correlations between the spatiotemporal trajectories and the service usage patterns. We observe the similarity and diversity in temporal patterns of the service usages and discuss their temporal correlations, time-ordering behavior between services, and behavioral overlap network based on the clustering results. Our approach provides a useful framework for handset-based data analysis, and hence it would be important for better design of information and communications technology (ICT) enabled social environments and services.

This paper is organized as follows. In Section 2 we describe the data collection and preparation methods. In Section 3 several contexts for each user are identified by means of the context detection method applied to user's spatiotemporal trajectory. In Section 4 we uncover the spatiotemporal correlations and the similarity and diversity in temporal patterns of the service usages. Finally, we summarize the results with concluding remarks in Section 5.

2 Handset-based dataset

2.1 Data collection method

The handset-based dataset in this study was collected by the MobiTrack software installed on Nokia Symbian smartphones of 183 participants or users from September 2009 to December 2010, *i.e.* for a period spanning about 16 months. All users were students and staff members of Aalto University, Finland and identified as early adopters of mobile phones and services [34]. The dataset was anonymized so that no personal information of the users could be obtained. We consider only 124 users with the overall duration of handset usage longer than 30 days, see Section 3 for details.

The dataset consists of two kinds of information: locations and service usages. The resolution of locations is limited to the physical area covered by each mobile network base station, *i.e.* cell, denoted by c. Whenever the handset is connected to a new cell or otherwise every half an hour, the identifier of the cell connected by the handset was recorded with a timestamp t with one second resolution. Each cell can be located in the geographical space with a unique pair of latitude and longitude. The geographic information for cells

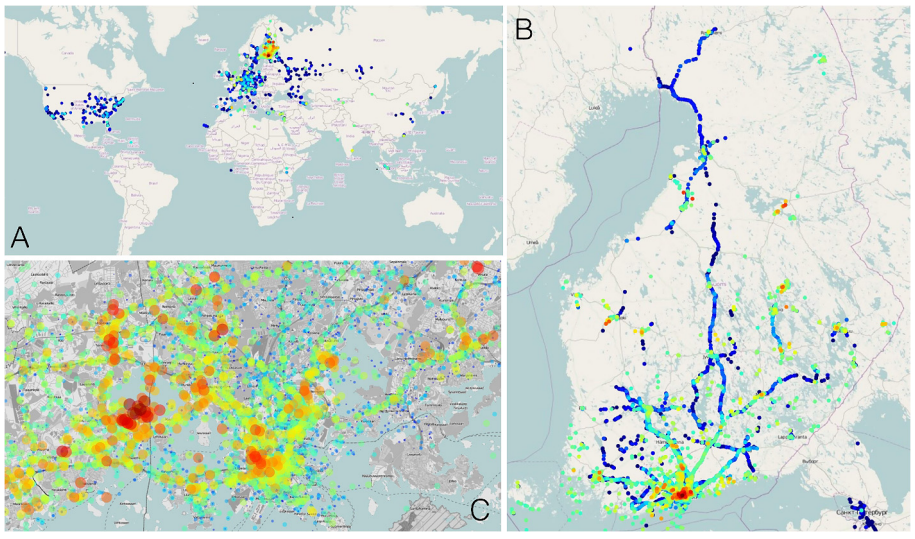

Figure 1 Recording frequencies at mobile network base stations by all users. (A) Over the world, **(B)** in Finland, and **(C)** in the Helsinki municipal area. The higher frequency is denoted by the warmer color. In (C) the size of circle is logarithmic in frequency.

and the maps used in Figures 1 and 6 were collected as a part of the OpenNetMap project and from open databases [35–37]. For all users we have 5,596,041 records at 99,206 different cells. Although only 29.0% of cells could be located in the geographical space, they correspond to 91.3% of records. Figure 1 shows all located cells over the world, in Finland, and in the Helsinki municipal area. In this way, the detailed spatiotemporal trajectory of each user could be constructed in terms of a sequence of cell records $\{(c_k, t_k)\}$, where k denotes the ordered index of record.

For service usage data we consider five services: web domain visit (web), application (app), email, voice call (call), and short message service (SMS). Each service usage or event was recorded with a timestamp with one second resolution together with service-specific relevant information. In the case of web domain visits, a URL (Uniform Resource Locator) was extracted and recorded whether it was visited *via* browser or widget. Only the applications visible in the foreground of the handset were recorded so that no process or application running in the background was considered. The records of communication services, such as email, call, and SMS, include the information on whether the user was an initiator or receiver of the communication event, and on the communication partner if available. For more information regarding the data collection method, see [34].

2.2 Data preparation method

The service usage dataset contains events mostly generated by users but it also contains automatic events by the operating system of the handsets. In order to observe the pure human behavior, we systematically filtered out these automatic events. However, some spurious regularities still remain in the web dataset. In the cases of google.com, facebook.com and so on, once a web is connected, the browser might visit the same web automatically for periodic updates and synchronization of accounts until the web is disconnected. To resolve this issue, we obtain the distribution of inter-event time τ, defined as the time interval between consecutive web domain visits by the same user. Several sharp peaks at spe-

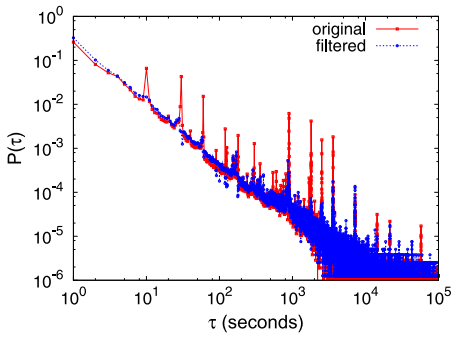

Figure 2 Original and filtered distributions of inter-event time for web domain visits by all users. The inter-event time is defined as the time interval between consecutive web domain visits by the same user. The peaks due to automatic events by the browser have been successfully suppressed after filtering.

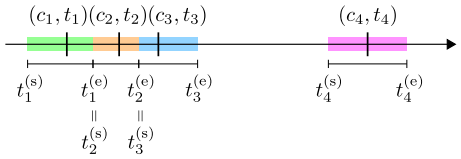

Figure 3 Schematic diagram for deriving temporal boundaries from a sequence of cell records. Temporal boundaries $\{(c_k, t_k^{(s)}, t_k^{(e)})\}$ (colored boxes) are derived from a sequence of cell records $\{(c_k, t_k)\}$ (vertical black lines). It is assumed that the user stays in the cell c_k from the moment of $t_k^{(s)}$ to $t_k^{(e)}$. See the text for details.

cific inter-event times are found, where each peak is mostly related to the single webpage. We remove all the events leading to those inter-event times, except for the event trains consisting of only two events with $\tau = 10$ seconds. It is because some trains with only two events separated by 10 seconds can also be generated by users. As new regularities become visible after filtering, we apply this method recursively until the peaks are suppressed considerably, leading to an approximately 25% of entire events removed. Figure 2 shows that this filtering method for web dataset does not change the overall characteristics of the inter-event time distribution.

We also ignore some user-generated application events associated with other service usages, corresponding to 17% of entire events. For example, the user opens the messaging application when sending or receiving SMSs. These application events might lead to artificial correlations between different service usages. In addition, corrupted events, less than 0.1% of the whole dataset, have been ignored or manually corrected. Finally, we have 792,971 web domain visits, 433,726 application events, 17,976 emails, 79,779 calls, and 79,283 SMSs in the service usage dataset.

3 Context detection from spatiotemporal pattern

In order to detect the contexts for each user, we construct the user's spatiotemporal trajectory from a sequence of cell records $\{(c_k, t_k)\}$. It is necessary to infer the user's location between consecutive timestamps of cell records. From a sequence of cell records, we derive the temporal boundaries $\{(c_k, t_k^{(s)}, t_k^{(e)})\}$ for the user's trajectory, implying that the user stays within the area covered by cell c_k from the moment of $t_k^{(s)}$ to $t_k^{(e)}$, see Figure 3. It is assumed that the user stays in the cell c_k till $t_k^{(e)} = \frac{1}{2}(t_k + t_{k+1})$ and then in the cell c_{k+1} from

$t^{(s)}_{k+1} = t^{(e)}_k$ when $t_{k+1} - t_k \leq 2t_c$. Here we set t_c as half an hour, *i.e.* the time interval for regular cell recording. The time interval between consecutive timestamps longer than $2t_c$ implies that the handset may be turned off, used in offline or airplane mode, or not able to detect any cell nearby. If $t_{k+1} - t_k > 2t_c$, the user is considered to stay in the cell c_k till $t^{(e)}_k = t_k + t_c$ and in the cell c_{k+1} from $t^{(s)}_{k+1} = t_{k+1} - t_c$. Hence, the location is unknown between $t^{(e)}_k$ and $t^{(s)}_{k+1}$. Then, the total time spent, *i.e.* duration, in each cell c is obtained as follows:

$$d_c = \sum_{\{k|c_k=c\}} \left(t^{(e)}_k - t^{(s)}_k\right). \tag{1}$$

If the sum of durations in all the recorded cells, $D \equiv \sum_c d_c$, is less than 30 days, that user is not considered for the further analysis, leading to 124 available users. The average and standard deviation of D for available users are 121 ± 63 days.

In addition, we observe back and forth changes in a short time span between two cells covering the neighboring areas. It can occur even without any real movement of the handset if the handset is located at the boundary of two neighboring cells. To filter out this noisy behavior, the involved cells can be clustered by a sandwich clustering method [26]. Here we consider only one type of sandwich with four records involving two cells, *i.e.* $c_k = c_{k+2} \neq c_{k+1} = c_{k+3}$ with $t^{(e)}_l - t^{(s)}_l \leq t_c$ for $l = k, \ldots, k+3$. Whenever this type of sandwich is detected, every c_k in the temporal boundaries is replaced by or merged into c_{k+1} if $d_{c_{k+1}} > d_{c_k}$, and *vice versa*. Consequently, some geographically neighboring cells can be clustered into one representative cell, which from now on will be considered equally with normal cells. For example, the first row in Figure 4 shows the user 81's temporal boundaries during typical Friday and Saturday. Note that clustering cells for one user is independent of other users' records.

We find spatiotemporal inhomogeneities of the trajectories of handsets on the individual basis as well as at the aggregate level. As an illustrative example, we obtain the rank curve $d(r)$, defined as the duration in the rth cell c in a descending order according to d_c. The rank curve for all users is highly skewed, such that the first few cells, including one in Otaniemi campus of Aalto University, were visited for more than a few months while 88.9% of cells were visited for less than one hour, as shown in Figure 5. The same inhomogeneities are

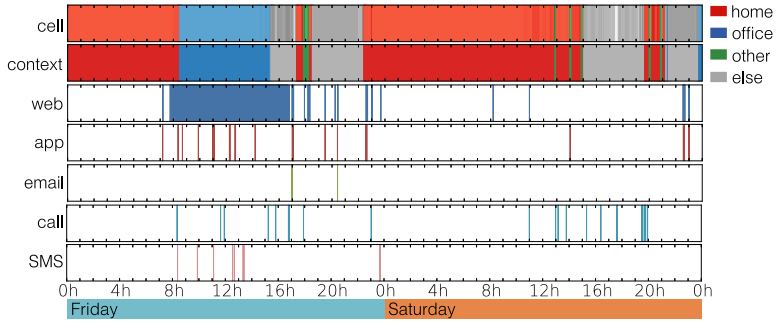

Figure 4 Locations and service usage patterns of a sample user 81 during typical Friday and Saturday. The first and second rows represent cells and contexts assigned to cells. Home, Office, Other meaningful place, and Elsewhere are denoted by red, blue, green, and gray colors, respectively. Different depths of the same color indicate the different cells belonging to the same context. Service usage events are denoted by vertical lines in the rows of web, app, email, call, and SMS (from the third to the bottom).

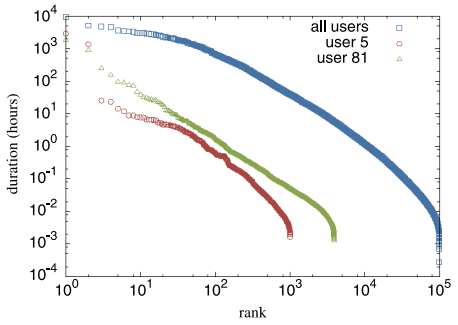

Figure 5 Rank curves of cells. A rank curve is defined as an ordered sequence of durations in cells. We plot rank curves of cells for all users and for sample users 5 and 81.

also observed for individual users. For example, the rank curves for users 5 and 81 are shown in Figure 5, who were selected to show the representative behavior.

The heavily visited cells are supposed to cover meaningful places to the handset user, such as home and office. Since the service usage patterns might be affected by the different characteristics of meaningful places, it is important to identify the context characterizing the situation of user. Here the context is preferred to the meaningful place because the time and place of handset usage are not independent but correlated, *e.g.* nighttime at home and daytime in office [26]. Each cell will be detected as one of five contexts, such as Home, Office, Other meaningful place (Other), Elsewhere (Else), and Abroad. One context can be assigned to several cells. The identifier of a cell contains the mobile country code (MCC), by which Abroad context is assigned to the cells out of Finland. For the cells within Finland, we obtain more detailed durations for each cell c:

1. duration on weekdays ($d_{c,\mathrm{wd}}$),
2. duration on weekdays between 0 AM and 6 AM ($d_{c,0-6}$), and
3. duration on weekdays between 10 AM and 4 PM ($d_{c,10-16}$).

Now we describe criteria for assigning contexts except for Abroad. A cell is detected as Elsewhere (Else) if the duration in that cell is negligible to the total duration as

$$d_c/D < t_{\mathrm{elsewhere}} = 0.02. \tag{2}$$

For example, Else is assigned to the cells along the highways. The threshold value of $t_{\mathrm{elsewhere}}$ has been determined in order to leave only 0.2% of cells, *i.e.* 3.73 cells per user, for other contexts. A cell is detected as Office if the user spends a considerable time in that cell during the working time on weekdays as

$$d_{c,\mathrm{wd}}/d_c > t_{\mathrm{weekday}} = 0.8 \tag{3}$$

and

$$d_{c,10-16}/d_{c,\mathrm{wd}} > t_{\mathrm{worktime}} = 0.5. \tag{4}$$

With above threshold values, at least one Office has been detected for more than half of the users. Note that most users were students so that they might not have any regular

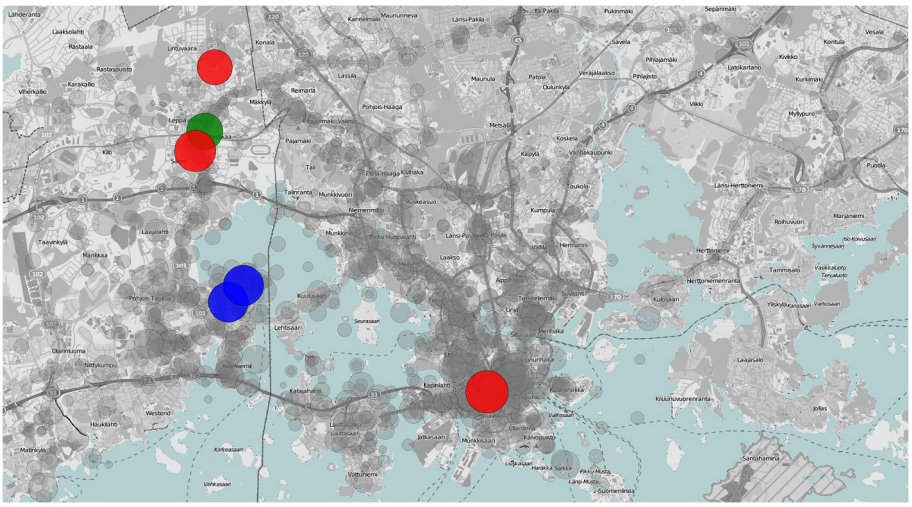

Figure 6 Contexts detected for sample users 5 and 81 in a map of Helsinki municipal area. Each cell is represented by the circle with its radius according to the duration in that cell. The cells are identified as either Home (red), Office (blue), Other meaningful place (green), or Elsewhere (gray).

places to visit during the working time. Next, Home is assigned to a cell if the user spends a considerable time in that cell for nighttime and free time, *i.e.* the remaining time except for the working time, on weekdays as

$$d_{c,0-6}/d_{c,\mathrm{wd}} > t_{\mathrm{nighttime}} = 0.1 \tag{5}$$

and

$$d_{c,10-16}/d_{c,\mathrm{wd}} < t_{\mathrm{freetime}} = 0.3. \tag{6}$$

With above threshold values, at least one Home has been detected for all users except for two of them. Many users turn out to have more than one Home, such as user's own home and his/her parent's home. Finally, the remaining cells are detected as Other meaningful place (Other). Figure 6 shows the locations of detected contexts for sample users in the Helsinki municipal area. We put two sample users' contexts together to avoid privacy issues.

Our context detection method is validated by weekly patterns of duration for different contexts obtained for sample users and at the aggregate level, as depicted in Figure 7. For example, the user 5 without Other detected shows a very regular pattern, especially on weekdays, *i.e.* at Home in nighttime, in Office during the working time, and at Else when moving between Home and Office. Weekly patterns of user 81 are comparable to the temporal boundaries in terms of detected contexts, as depicted in the second row in Figure 4. Weekly patterns of duration aggregated over all users show the overall behavior. Durations at Home, Office, Other, and Else account for 66.8%, 7.0%, 8.5%, and 14.0% of the total duration of all users, respectively.

4 Spatiotemporal correlations of service usages

We investigate correlations between users' spatiotemporal trajectories and their service usage patterns. Here five services, such as web domain visit (web), application (app), email,

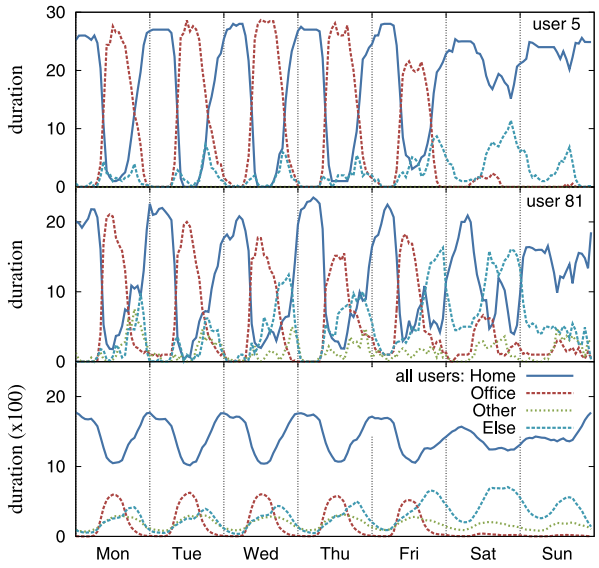

Figure 7 Weekly patterns of duration for the different contexts. Weekly patterns of duration in hours for the different contexts are consistent with the context detection results for users 5 and 81 and for all users (from top to bottom). The typical weekly cycles of humans are observed.

voice call (call), and short message service (SMS), are considered and each service is denoted by s. The spatiotemporal correlation of service usages for user i is fully characterized by the number of events corresponding to the service s in the cell c and at time t, denoted by $n_{is}(c,t)$. For gaining contextual understanding of correlations we consider the contexts instead of cells, *i.e.* $n_{is}(C,t) = \sum_c n_{is}(c,t)$, where the summation is over c detected as context C.

4.1 Contextual correlations of service usages

We first focus on the contextual correlations of service usages with $n_{is}(C) = \sum_t n_{is}(C,t)$. Since services have qualitatively different characteristics, the numbers of events of different services cannot be directly compared to each other but only in terms of fractions and intensities of usages. The fraction of service usage is defined as follows

$$f_{is}(C) = \frac{n_{is}(C)}{\sum_C n_{is}(C)}. \tag{7}$$

Figure 8 (left) shows the fractions for sample users 5 and 81 as well as their means over all users with standard errors, measured by the bootstrap method. The handset of user 5 has never been abroad and no Other context is detected. For this user all service usages are more active at Home and Office than at Else, which is very different from the service usage patterns of user 81. Due to the diversity of the service usage patterns among users, any general conclusion cannot be made on the individual basis. However, by looking at the means with standard errors, it is found that all service usages are the most active at Home, while they are relatively inactive for other contexts. Given the aggregate durations for different contexts obtained in the Section 3, this finding can be explained such that the longer duration for some context means the higher chance for service usage.

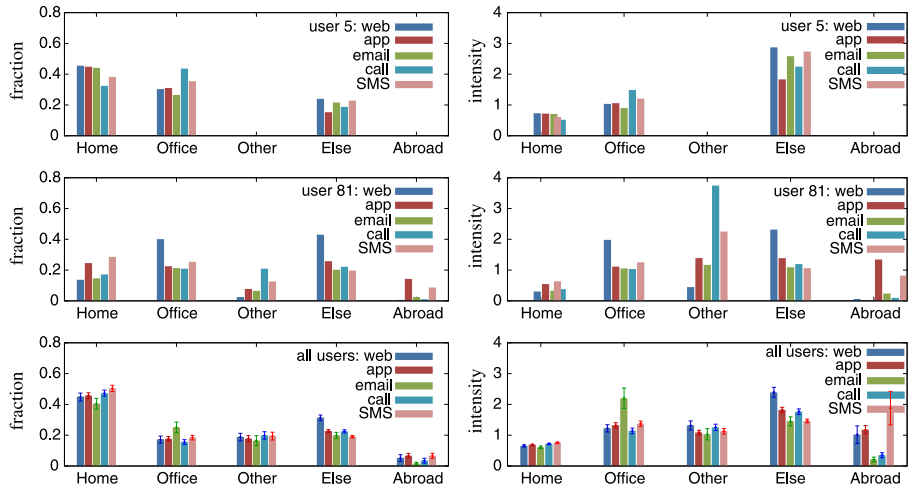

Figure 8 Contextual correlations of service usages. Contextual correlations of service usages are measured for users 5 and 81 and for all users (from top to bottom). Fractions (left) and intensities (right) of service usages are defined in Eqs. (7) and (8), respectively. Standard errors are also provided for the user-averaged statistics.

Accordingly, instead of the fractions of service usages we consider those divided by the corresponding durations as follows:

$$I_{is}(C) = \frac{n_{is}(C)}{\sum_C n_{is}(C)} \cdot \frac{\sum_C d_{iC}}{d_{iC}}, \tag{8}$$

where d_{iC} denotes the duration of user i for context C. The results are shown in Figure 8 (right). Despite of the diversity among users, the means of intensities of different services for the same context have to some extent similar values. The large mean of intensity of email usage in Office might be due to the fact that users prefer emails to calls or SMSs in classes or laboratories during the working time. The large mean of intensity of web usage at Else could be the result of users killing time by surfing the webpages while on the move. One could also say that users while abroad tend to use SMSs more than other communication services. Finally, for all services, only the means of intensity at Home turn out to be less than 1 and most inactive, which could be partly because users have many other activities to do at Home.

4.2 Temporal correlations and time-ordering of service usages

We turn to analyze the temporal correlations of service usages in terms of $n_{is}(t) = \sum_C n_{is}(C,t)$, where the summation is over all contexts with one exception, Abroad. It is because the service usage abroad cannot be considered as normal, as shown in Figure 8. We first obtain weekday and weekend patterns of service usages as

$$n_{is}^{\mathrm{wd}}(t) = \sum_k n_{is}(t + kT_d), \tag{9}$$

$$n_{is}^{\mathrm{we}}(t) = \sum_{k'} n_{is}\left(t + k'T_d\right) \tag{10}$$

for $0 \leq t < T_d$ with $T_d = 1$ day. Here k and k' denote the indexes of weekdays and weekends, respectively. The weekday and weekend event rates of service s for user i are defined as

$$\rho_{is}^{\mathrm{wd}}(t) = \frac{a n_{is}^{\mathrm{wd}}(t)}{\sum_t [a n_{is}^{\mathrm{wd}}(t) + a' n_{is}^{\mathrm{we}}(t)]}, \tag{11}$$

$$\rho_{is}^{\mathrm{we}}(t) = \frac{a' n_{is}^{\mathrm{we}}(t)}{\sum_t [a n_{is}^{\mathrm{wd}}(t) + a' n_{is}^{\mathrm{we}}(t)]}, \tag{12}$$

where $a = 1/5$ and $a' = 1/2$ are weights for normalization. In addition we obtain the weekday and weekend event rates averaged over all users.

In Figure 9 we show the individual event rates for sample users 5 and 81 as well as the event rates averaged over all users. The overall behavior of the individual and user-averaged event rates reflects typical weekly cycles of humans by being more active in the daytime and on weekdays and less active in the nighttime and on weekends. From the user-averaged event rates, we find that email (call) is more used around noon (late afternoon) on weekdays, while email (call) is less (more) used than other services in the weekend daytime. Since most users in our dataset were students and staff members of the university, they might not be making or receiving calls in classes or laboratories in the weekday daytime. Instead they might be using other communication services, such as email and SMS. On the other hand, users might be using call more than email outside class or laboratory on weekends.

To investigate the temporal correlations between service usages for each user, we calculate the Pearson correlation coefficient (PCC) by using the event rates of services s and s' for user i:

$$r_{i,ss'} = \frac{\sum_t [\rho_{is}(t) - \bar{\rho}_{is}][\rho_{is'}(t) - \bar{\rho}_{is'}]}{\sqrt{\sum_t [\rho_{is}(t) - \bar{\rho}_{is}]^2} \sqrt{\sum_t [\rho_{is'}(t) - \bar{\rho}_{is'}]^2}}, \tag{13}$$

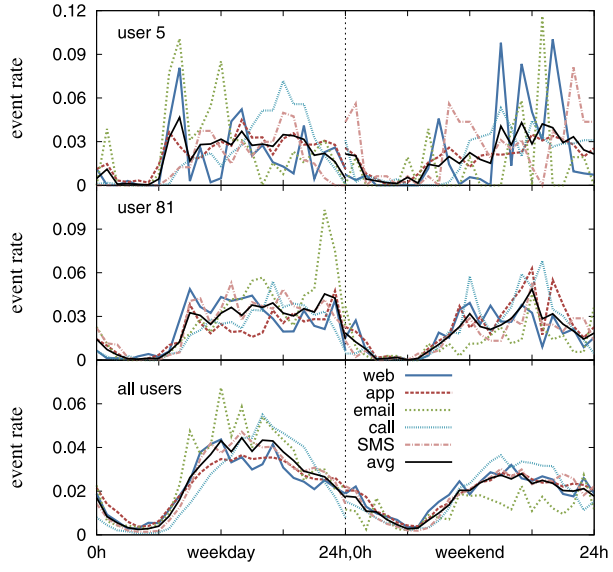

Figure 9 Weekday and weekend service usage patterns. Weekday and weekend service usage patterns are measured for users 5 and 81 and for all users (from top to bottom), showing the similarity and diversity among users. The bin size was set to one hour.

where $\bar{\rho}_{is} = T_d^{-1} \sum_t \rho_{is}(t)$. For the PCC on weekdays and on weekends, $\rho_{is}^{\mathrm{wd}}(t)$ and $\rho_{is}^{\mathrm{we}}(t)$ are used, respectively. The values of PCC turn out in most cases to be positive (not shown here). This is mainly due to the typical weekly cycles of humans as mentioned before. To correct such cycles, for each case of weekdays and weekends we consider de-seasoned event rates defined as

$$\Delta\rho_{is}(t) = \rho_{is}(t) - \frac{1}{S_i} \sum_s \rho_{is}(t), \tag{14}$$

where S_i denotes the number of services the user i have used.

As shown in Figure 10, the values of PCC obtained for the de-seasoned event rates show similar and distinct behavior among users as well as between weekdays and weekends. For example, in the case of user 5, the strongly positive correlation between call and SMS usages on weekdays turns to be slightly negative on weekends. This result is consistent with

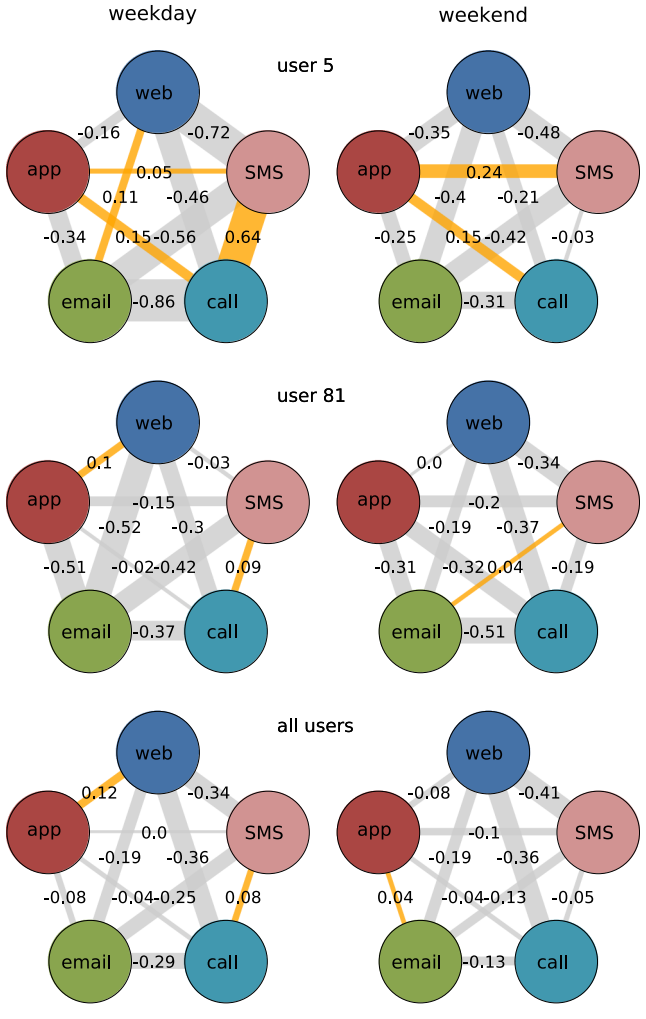

Figure 10 Pearson correlation coefficients among service usages. Pearson correlation coefficients among service usages for users 5 and 81 and for all users (from top to bottom) are obtained from weekday (left) and weekend (right) event rates. Positive and negative correlations are represented by orange and gray lines, respectively.

the temporal patterns depicted in Figure 9. The positive (negative) correlation between services by being used at the same time (at different times) of the week can be interpreted such that those services are complementary (substitutive) with each other [38]. Then, we obtain and compare distributions of PCC over all users for each pair of services. The mean values for web-app and call-SMS pairs (app-email pair) are slightly positive (negative) on weekdays and become slightly negative (positive) on weekends. All other pairs have the negative mean values. The result for positive correlations is inconclusive due to the large standard errors of PCC up to 0.05. However, for the pairs of services with large negative correlations, such as web-call and web-SMS pairs, we can argue that those services might be used in a substitutive way. In order to compare the correlations for weekdays and for weekends, we have conducted the Kolmogorov-Smirnov test. It is found that the distributions of PCC for weekdays and for weekends are significantly different for the pairs of web-app (p-value less than 0.005), app-email (0.03), email-call (0.03), email-SMS (0.03), and call-SMS (0.02). This list of pairs contains all the pairs whose sign of the mean has changed from weekdays to weekends.

For more detailed, *i.e.* event-based analysis of correlations among service usages, we obtain the distribution of time interval between two consecutive or simultaneous events but of different services of the same user. Precisely, the time interval for a pair of services s and s' is defined by $\Delta t_{ss'} = t_{s'} - t_s$ with event timings t_s and t'_s. As shown in the upper panels of Figure 11(a), distributions for some service pairs have a peak at the negative value of $\Delta t_{ss'}$ both for weekdays and for weekends. This indicates that the event of service s follows that of service s'. On the other hand, distributions for other pairs of services do not show any distinct peaks, implying no temporal correlation. This time-ordering behavior could mean that one service usage might effectively induce another service usage. However, we cannot investigate such a process by our dataset. We summarize the results such that communication services, such as email, call and SMS, are followed by non-communication services, *i.e.* web and app, as depicted in Figure 11(b). We also obtain the distributions of time interval for different contexts. We find the overall similar time-ordering behavior (not shown here), except that email is followed by web at Home and that app does not follow communication services abroad. Note that the event-based analysis cannot be directly compared to the analysis of aggregated weekly patterns.

4.3 Clustering and overlaps in temporal patterns of service usage

As it turns out, the temporal patterns of service usage are diverse from one user to another, while some of them still show similar behavior. To investigate the similarity and diversity of weekly patterns for each service we apply the k-means clustering method [39] to the weekly event rates as $\rho_{is}(t) \equiv \{\rho_{is}^{\mathrm{wd}}(t), \rho_{is}^{\mathrm{we}}(t)\}$. To correct the typical weekly cycles of each service (not of each user), we use the de-seasoned event rates as follows

$$\Delta\rho_{is}(t) = \rho_{is}(t) - \frac{1}{N_s}\sum_i \rho_{is}(t), \tag{15}$$

where N_s denotes the number of users showing any activity in service s. We similarly define the service-averaged event rates for each user for the clustering, to be denoted by avg. In each case we set the number of clusters as $k = 10$ and the cluster index is denoted by $q = 0, \ldots, 9$. Clustering has been conducted 2,000 times with different initial conditions and

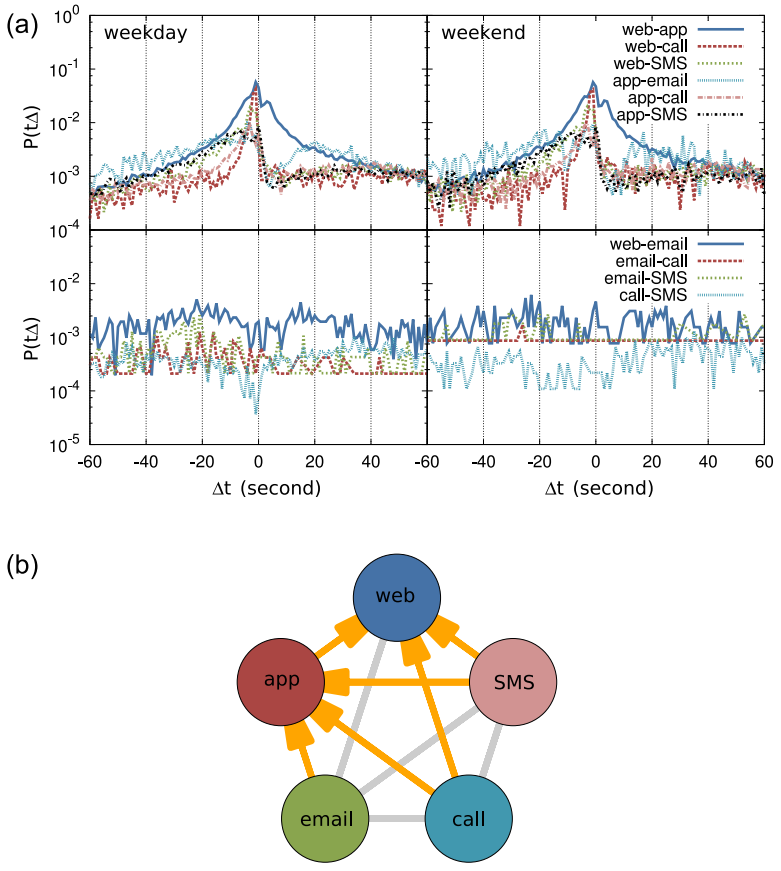

Figure 11 Time-ordering behavior between services. (a) Distributions of time interval $\Delta t_{ss'}$ between consecutive events of different services s and s'. **(b)** Diagram for time-ordering behavior between services based on the distributions of time interval.

Table 1 *k*-means clustering results for weekly patterns of service usages

Service	$q=0$	1	2	3	4	5	6	7	8	9	N_s
web	74	9	7	6	5	3	3	2	1	1	111
app	50	32	10	7	6	6	5	4	3	1	124
email	55	3	3	2	1	1	1	1	1	1	69
call	54	40	14	5	4	1	1	1	1	1	122
SMS	74	14	11	9	5	4	3	1	1	1	123
avg	64	21	16	6	5	5	4	1	1	1	124

We summarize k-means clustering results for weekly patterns of service usages with $k = 10$. q and N_S denote the cluster index and the number of available users for service s, respectively.

here we present the result maximizing the quality of clustering or validity index, defined as the minimum inter-cluster distance divided by the sum of intra-cluster distances [39].

The clustering results are summarized in Table 1 and only a few weekly patterns of dominant clusters are shown in Figure 12. Only one dominant cluster is found in each case of web and email usages, implying similar patterns among users. Weekly patterns of app, call, and SMS usages are clustered into more than one dominant cluster. Compared to the largest cluster ($q = 0$) of call usage, the second largest cluster ($q = 1$) can be characterized by larger activities in the weekday daytime and in the weekend morning. The behavioral difference between dominant clusters in SMS usage is also obvious. The largest cluster

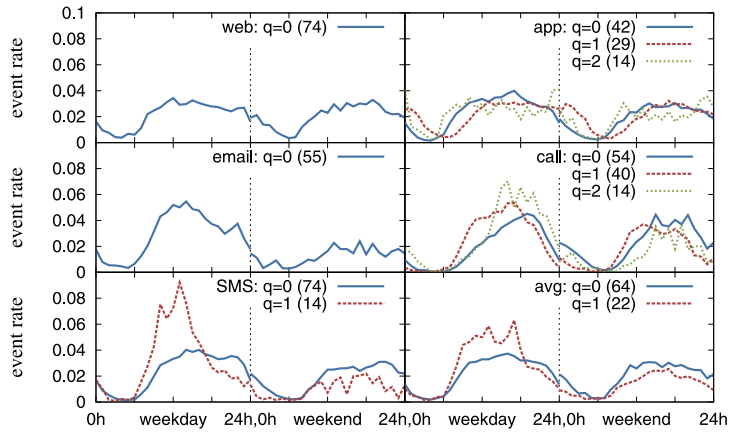

Figure 12 ***k*-means clustering results of users' weekly patterns.** We have used $k = 10$ and plotted only a few dominant clusters with cluster size in the parenthesis, see Table 1 for details. The bin size was set to one hour.

($q = 0$) represents the evening-type users, while the second largest cluster ($q = 1$) does the morning-type users on weekdays. In the case of service-averaged usage patterns, the second largest cluster ($q = 1$) shows the larger (smaller) activity in the daytime on weekdays (on weekends) than the largest cluster ($q = 0$). To check the validity of clustering results, we obtain the Pearson correlation matrices using the de-seasoned event rates, $\Delta\rho_{is}(t)$. All the matrices support the k-means clustering results, see Figure 13. We also tested the effect of the number of means, k, on the clustering and found that the results are qualitatively similar apart from the number of small or outlying clusters.

Finally, in order to get insight into the overall structure of temporal correlations among users and services, we construct an overlap network based on the clustering results. This leads to the network of overlapping communities [40], where nodes and link weights of the network represent users and their overlaps, respectively. Precisely, the behavioral overlap is defined as the number of services in which two users, say i and j, belong to the same cluster as

$$O^{\mathrm{B}}_{ij} = \sum_{s} \delta(q_{is}, q_{js}). \tag{16}$$

Here q_{is} denotes a cluster index for user i's service s, and the Kronecker delta function $\delta(q, q')$ gives 1 if $q = q'$ and 0 otherwise. Figure 14 shows the overlap network with 436 links of $O^{\mathrm{B}} = 4$ and 5. The behavioral overlap $O^{\mathrm{B}} = 5$ of a link, denoted by thick black line, implies that the neighboring users belong to the same clusters for all services, *i.e.* they are fully synchronized. We find cliques consisting of only the fully synchronized users, which we call synchronized cores. The largest synchronized core with 9 users is closely related to the second largest synchronized core except for belonging to different clusters of call usage. These cores are also connected to many other users but not as a synchronized core. This agglomerate structure can be induced by the relatively homogeneous demographics of users in our dataset. However, we like to note that the clustering was applied to the de-seasoned event rates, which have been subtracted by the user-averaged temporal behavior.

We compare the behavioral overlap network based on the clustering results to the communication network of users. The communication network can be constructed from the

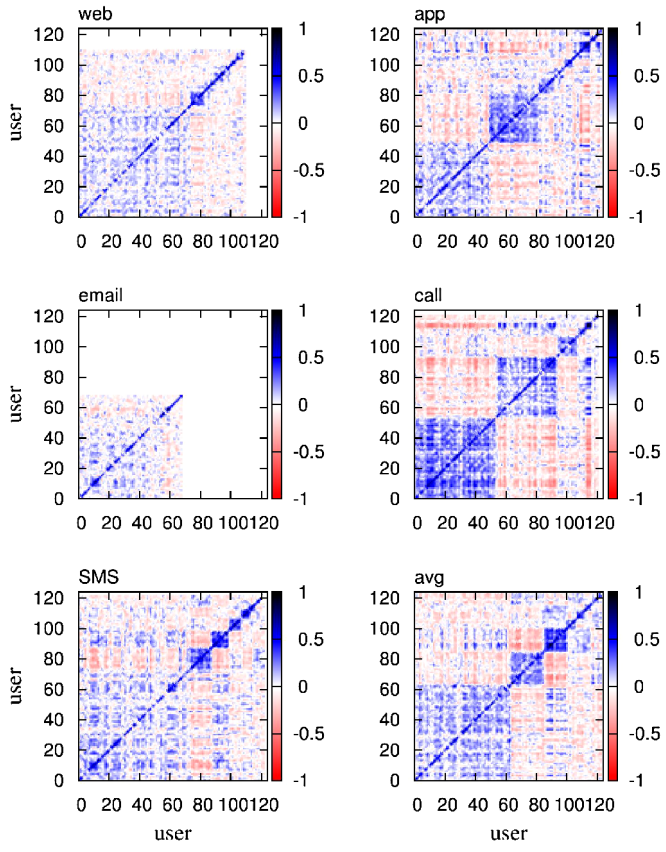

Figure 13 Pearson correlation matrices of users' de-seasoned event rates. These matrices support the validity of the *k*-means clustering results in Table 1. The user index has been sorted according to the corresponding cluster index and blank spaces are due to totally inactive users.

call and SMS datasets containing the information on communication partners. Only 67 out of 124 users and 205 links between users are identified. The topological overlap of a link ij is defined as [6]

$$O_{ij}^{\mathrm{T}} = \frac{|\Lambda_i \cap \Lambda_j|}{|\Lambda_i \cup \Lambda_j| - 2}, \tag{17}$$

where Λ_i denotes the set of neighbors of node i. O_{ij}^{T} has a value of 1 if i and j have exactly the same neighbors except for themselves and it has a value of 0 if they do not have any neighbors in common. Figure 15 shows the overall positive correlation between behavioral and topological overlaps. It implies that connected users sharing more common neighbors show more similar weekly patterns of service usages. Thus, the behavioral overlap network based on the service usages can be used to reveal the communication network structure of users.

5 Summary

We have investigated spatiotemporal correlations and temporal diversities of service usages by analyzing a handset-based dataset collected from 124 users for over 16 months. The dataset consists of locations and service usages. After constructing the precise spatiotemporal trajectory for each user based on the location dataset, we identify several

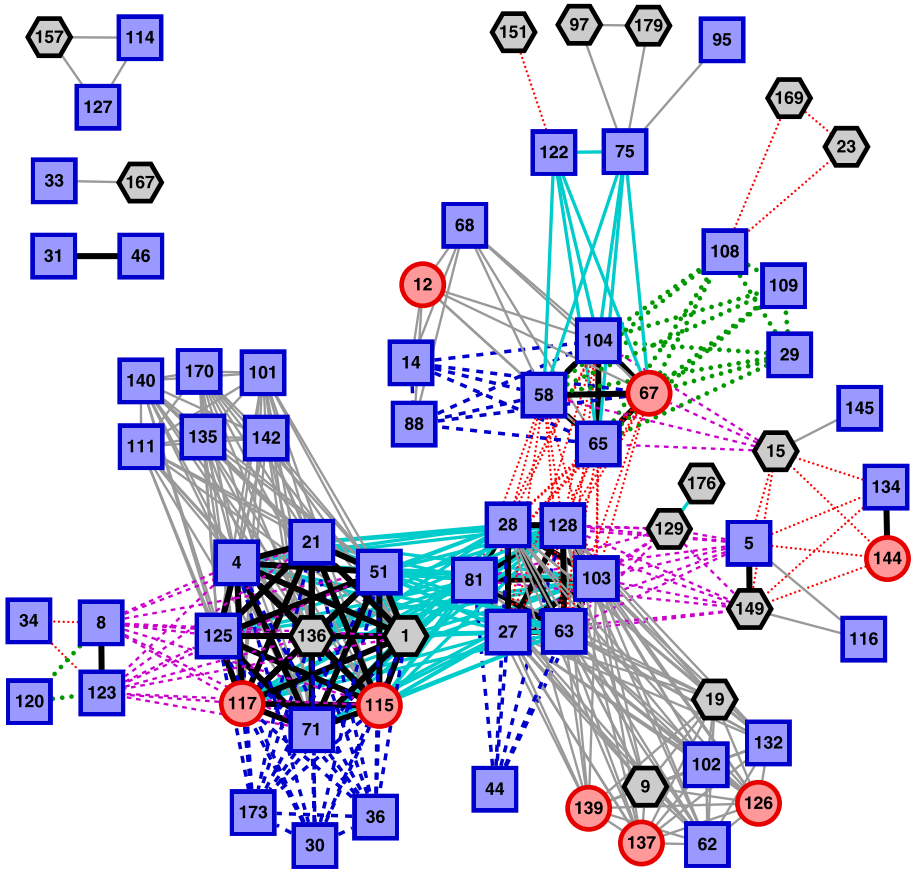

Figure 14 Overlap network constructed based on the clustering results for all services. Circle, square, and hexagonal nodes represent female, male, and unknown gender of users, respectively. Each black solid thick line denotes a link between users who belong to the same clusters for all services. Other colored lines denote the links between users who belong to the same clusters for all but one service: web (dashed thick blue), app (dotted thin red), email (dotted thick green), call (solid thick cyan), SMS (dashed thin violet), or due to the unused service by either user (solid thin gray). This figure was generated using Cytoscape v2.8.1 [41].

meaningful places or contexts by means of context detection method. As contexts, Home, Office, Other meaningful place, Elsewhere, and Abroad are considered. We showed how the context affects the service usage patterns of users, including their web domain visit (web), application (app), email, voice call (call), and short message service (SMS).

In this study we have found the similarity and diversity of weekly patterns among users and services, in terms of temporal correlations, time-ordering behavior between services, and overlap network based on clustering. The services used at the same time (at different times) of the week lead to the positive (negative) correlations between them, which can be interpreted as being complementary (substitutive) to each other. By conducting the event-based analysis instead of weekly patterns we observe the time-ordering behavior between services, such that communication services, *i.e.* email, call, and SMS, are followed by the non-communication services, *i.e.* web and app. Finally, the similarity and diversity of weekly patterns of service usages enable us to classify users into several different clusters, *e.g.* as characterized by the morning-type or evening-type usage patterns, except for the web and email usages. The behavioral overlap network constructed based on the clustering results can be used to reveal the communication or real social network structure of users.

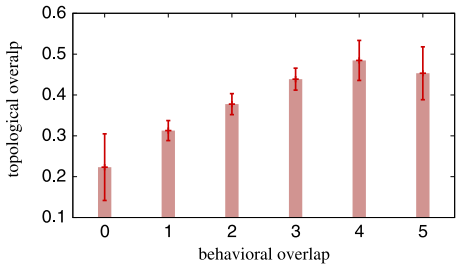

Figure 15 Topological overlap as a function of behavioral overlap. We observe the overall positive correlation between topological overlap from communication network of users and behavioral overlap based on the clustering results.

Our findings on the spatiotemporal correlations of service usage patterns for different contexts enable us to better understand the behavior of humans and what that implies. This is also important for better design of information and communications technology (ICT) enabled social environments and services. However, more detailed analysis with higher resolution is required to reveal the underlying mechanism or the origin of spatiotemporal correlations.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
HJ and MK designed and performed the data analysis, and wrote the manuscript. JK, MK, and KK initiated the study. JK was in charge of setting up the data collection panel and collecting the data. All authors participated in the revision of the final manuscript.

Author details
[1]Department of Biomedical Engineering and Computational Science, School of Science, Aalto University, P.O. Box 12200, Espoo, Finland. [2]Department of Communications and Networking, School of Electrical Engineering, Aalto University, P.O. Box 13000, Espoo, Finland.

Acknowledgements
The research data were collected in the OtaSizzle project that is funded by Aalto University's MIDE program and Helsinki University of Technology TKK's 'Technology for Life' campaign donations from private companies and communities. The authors thank MobiTrack Innovations Ltd. for providing the mobile audience measurement platform. The sponsoring from Nokia and Elisa to this work is also acknowledged. Financial support by Aalto University postdoctoral program (HJ), from EU's 7th Framework Program's FET-Open to ICTeCollective project no. 238597, by the Academy of Finland, the Finnish Center of Excellence program 2006-2011, project no. 129670 (MK, KK), and by Future Internet Graduate School and MoMIE project (JK) are gratefully acknowledged.

Received: 7 April 2012 Accepted: 9 October 2012 Published: 6 November 2012

References

1. Goyal S (2009) Connections: an introduction to the network economy. Princeton University Press, Princeton
2. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. Rev Mod Phys 81(2):591-646
3. Lazer D, Pentland A, Adamic L, Aral S, Barabási AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Computational social science. Science 323(5915):721-723
4. Eckmann JP, Moses E, Sergi D (2004) Entropy of dialogues creates coherent structures in e-mail traffic. Proc Natl Acad Sci USA 101(40):14333-14337
5. Barabási AL (2005) The origin of bursts and heavy tails in human dynamics. Nature 435:207-211
6. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási AL (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 104(18):7332-7336
7. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media. In: Proceedings of the 19th international conference on World Wide Web, WWW '10. ACM, New York, pp 591-600
8. Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N (2008) Tastes, ties, and time: a new social network dataset using Facebook.com. Soc Netw 30(4):330-342
9. Kovanen L, Karsai M, Kaski K, Kertész J, Saramäki J (2011) Temporal motifs in time-dependent networks. J Stat Mech Theory Exp 2011(11):P11005
10. Jo HH, Karsai M, Kertész J, Kaski K (2012) Circadian pattern and burstiness in mobile phone communication. New J Phys 14:013055

11. Karsai M, Kaski K, Barabási AL, Kertész J (2012) Universal features of correlated bursty behaviour. Sci Rep 2:397
12. González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779-782
13. Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási AL (2008) Uncovering individual and collective human dynamics from mobile phone records. J Phys A, Math Theor 41(22):224015
14. Wang P, González MC, Hidalgo CA, Barabási AL (2009) Understanding the spreading patterns of mobile phone viruses. Science 324(5930):1071-1076
15. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. Science 327(5968):1018-1021
16. Song C, Koren T, Wang P, Barabasi AL (2010) Modelling the scaling properties of human mobility. Nat Phys 6(10):818-823
17. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. Pers Ubiquitous Comput 10(4):255-268
18. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proc Natl Acad Sci USA 106(36):15274-15278
19. Krings G, Calabrese F, Ratti C, Blondel VD (2009) Urban gravity: a model for inter-city telecommunication flows. J Stat Mech Theory Exp 2009(7):L07003
20. Bagrow JP, Lin YR (2012) Mesoscopic structure and social aspects of human mobility. PLoS ONE 7(5):e37676
21. Aharony N, Pan W, Ip C, Khayal I, Pentland A (2011) Social fMRI: investigating and shaping social mechanisms in the real world. Pervasive Mob Comput 7(6):643-659
22. Falaki H, Mahajan R, Kandula S, Lymberopoulos D, Govindan R, Estrin D (2010) Diversity in smartphone usage. In: Proceedings of the 8th international conference on mobile systems, applications, and services, MobiSys '10. ACM, New York, pp 179-194
23. Soikkeli T, Karikoski J, Hammainen H (2011) Diversity and end user context in smartphone usage sessions. In: Next generation mobile applications, services and technologies (NGMAST), 2011 5th international conference on. IEEE Press, New York, pp 7-12
24. Dey AK (2001) Understanding and using context. Pers Ubiquitous Comput 5:4-7
25. Verkasalo H (2009) Handset-based analysis of mobile service usage. PhD thesis, Helsinki University of Technology, Espoo, Finland
26. Soikkeli T (2011) The effect of context on smartphone usage sessions. Master's thesis, Aalto University, Espoo, Finland. http://aalto-fi.academia.edu/TapioSoikkeli/Papers
27. Karikoski J, Soikkeli T (2011) Contextual usage patterns in smartphone communication services. Pers Ubiquitous Comput. doi:10.1007/s00779-011-0503-0
28. OtaSizzle project. http://sizl.org
29. Montoliu R, Perez DG (2010) Discovering human places of interest from multimodal mobile phone data. In: Proceedings of the 9th international conference on mobile and ubiquitous multimedia, MUM '10. ACM, New York
30. Nurmi P, Koolwaaij J (2006) Identifying meaningful locations. In: Mobile and ubiquitous systems: networking and services, 2006 third annual international conference on, pp 1-8
31. Eagle N, Pentland AS (2009) Eigenbehaviors: identifying structure in routine. Behav Ecol Sociobiol 63(7):1057-1066
32. Reades J, Calabrese F, Ratti C (2009) Eigenplaces: analysing cities using the space-time structure of the mobile phone network. Environ Plan B, Plan Des 36(5):824-836
33. Park J, Lee DS, González MC (2010) The eigenmode analysis of human motion. J Stat Mech Theory Exp 2010(11):P11021
34. Karikoski J (2012) Handset-based data collection process and participant attitudes. Int J Handheld Comput Res (in press)
35. HIIT OpenNetMap project. http://opennetmap.rista.fi/
36. OpenCellID. http://www.opencellid.org
37. Location-API. http://location-api.com
38. Karikoski J, Luukkainen S (2011) Substitution in smartphone communication services. In: Intelligence in next generation networks (ICIN), 2011 15th international conference on. IEEE Press, New York, pp 313-318
39. Gan G, Ma C, Wu J (2007) Data clustering: theory, algorithms, and applications. SIAM, Philadelphia (illustrated edn)
40. Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814-818
41. Smoot ME, Ono K, Ruscheinski J, Wang PLL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27(3):431-432

doi:10.1140/epjds10
Cite this article as: Jo et al.: **Spatiotemporal correlations of handset-based service usages.** *EPJ Data Science* 2012 **1**:10.