On the Decomposition of Cell Phone Activity Patterns and their Connection with Urban Ecology

Blerim Cici[†◇⋆], Minas Gjoka[◇⋆], Athina Markopoulou[†‡◇⋆], Carter T. Butts[‡○⋆]
Networked Systems[†], CalIT2[⋆], CPCC[◇], EECS Dept.[‡]
Sociology Dept.[○] & Statistics Dept.[○]
UC Irvine, Irvine, USA
{bcici, mgjoka, athina, buttsc}@uci.edu

ABSTRACT

The goal of this paper is to infer features of urban ecology (i.e., social and economic activities, and social interaction) from spatio-temporal cell phone activity data. We present a novel approach that consists of (i) time series decomposition of the aggregate cell phone activity per unit area using spectral methods, (ii) clustering of areal units with similar activity patterns, and (ii) external validation using a ground truth data set we collected from municipal and online sources. The key to our approach is the spectral decomposition of the original cell phone activity series into seasonal communication series (SCS) and residual communication series (RCS). The former captures regular patterns of socio-economic activity within an area and can be used to segment a city into distinct clusters. RCS across areas enables the detection of regions that are subject to mutual social influence and of regions that are in direct communication contact. The RCS and SCS thus provide distinct probes into the structure and dynamics of the urban environment, both of which can be obtained from the same underlying data. We illustrate the effectiveness of our methodology by applying it to aggregate Call Description Records (CDRs) from the city of Milan.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Application— *data mining, spatial databases*; J.4 [**Social and Behavioral Sciences**]: Sociology—*Smart Cities, Urban Planning*

Keywords

Mobile networks; Call-Description Records (CDRs); Time Series Decomposition; Urban Ecology.

1. INTRODUCTION

The modern urban environment is a complex ecosystem, characterized by distinct geographical regions sharing common patterns of socio-economic activity, infrastructure, social cohesion, etc. [5, 24]; further, some of these regions are strongly interacting (via mobility, communication, etc.), while others are relatively isolated from one another [16, 18]. We refer to this intra-urban structure as "urban ecology." Historically, the detection of urban ecology has been difficult as it usually requires extensive local knowledge of the area in question and investigation using time-consuming and expensive techniques (e.g., informant interviews, ethnographic observation, etc.). This poses challenges for urban governance—e.g., urban planning, infrastructure management, administration, and law enforcement—particularly in an era of increasingly rapid urban growth and change (e.g., due to shifting patterns of immigration and economic activity). Likewise for the private sector, effective siting of businesses requires extensive knowledge of the urban landscape; in a global economy, obtaining such knowledge via years of experience "on the ground" in a given location may be expensive or impractical.

It is estimated that within the next forty years, two-thirds of the world's population will be living in expanding urban centers, and the level of urbanization is expected to increase in all major areas of the developing world [2]. Given this, there is clearly a need for methods that will cheaply yield up-to-date information on urban environments, without extensive on-the-ground investigation. A promising tool in that regard is the use of aggregate geo-located data on communication activity, a resource that is increasingly available given the near-universal penetration of mobile devices within urban populations. Because communication behavior is a fundamental aspect of virtually all social and economic interaction, even aggregated communication data can provide rich information on the types and volumes of activities occurring within a given geographical area, and on the degree of interaction (or lack thereof) between occupants of different areas. Such data is inexpensive, can be collected and distributed in a privacy-preserving

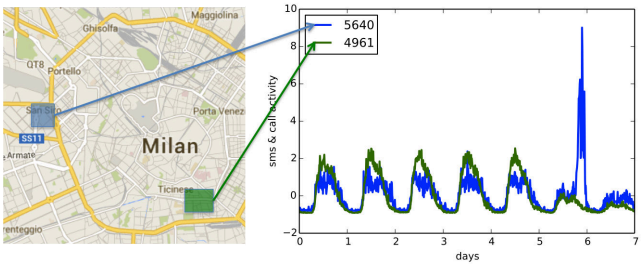


Figure 1: Cell phone activity series (normalized) for two grid squares of Milan, for the week of 11/4/2013–11/10/2013. Square 5640 is located close to the San Siro stadium, while square 4961 is located in a university region. Differences in seasonal patterns (weekday/weekend) reflect stable differences between the university and stadium environments.

This work has been supported by AFOSR MURI award FA9550-09-0643, and NSF CDI award 1028394. This work is based on data made available by Telecom Italia for the *Big Data Challenge* [3].

MobiHoc'15, June 22–25, 2015, Hangzhou, China.

http://dx.doi.org/10.1145/2746285.2746292.

manner (e.g., in the case of aggregate activity), and can be monitored to track developments within the urban landscape.

The question, then, is how aggregate mobile communication data can be used to provide information on urban ecology. In this paper, we provide an approach to this problem that draws on the notion of seasonal decomposition from the field of classical time series analysis. By decomposing communication data across time, frequency, and space, we create distinct time series that provide (respectively) information on routine activities and on deviations from those routines. Subsequent analysis of the resulting multivariate time series allows for the identification of socio-economically distinct regions within the urban environment, identification of socially interacting regions, and other goals of interest to the analyst. As we show, these analyses can be performed efficiently and in an unsupervised or semi-supervised manner, facilitating their use in settings for which the analyst has only limited resources for additional data collection.

Key idea: As is well known, a time series can be decomposed into three components (the *classical decomposition* [6]): a *trend*, representing systematic, non-periodic change over the time scale of observation; a *seasonal* component, representing systematic periodic variation in the phenomenon of interest (possibly with multiple characteristic frequencies); and a *residual* component, representing variation due to idiosyncratic factors and exogenous shocks. The value of the decomposition approach arises from the fact that many social processes either generate gradual, systematic change (e.g., population growth) or are strongly periodic (e.g., on hourly, daily, weekly, or annual time scales); the trend and seasonal components of the time series "selects out" such processes, allowing them to be either measured or removed from analysis (as desired). By turns, many other social processes are characterized by short-term responses to exogenous perturbations, and (when integrated over all perturbations) have little or no systematic component. Information on these processes is contained in the residual series, which can thus be used to study them without contamination from their systematic counterparts.

Data: The data that we use in this study consists of time series of aggregate cell phone traffic sent or received by persons within small areal units in the city of Milan (see Fig. 1), for approximately one-month period. In particular, we use aggregate call-description-records (CDRs) and aggregate SMS activity per area unit, made available for the *Big Data Challenge* [3] competition. In addition, we collected ground truth data from the municipality of Milan and online sources, containing elements such as universities, residential areas, sport centers, parks, etc. This latter data is used here for external evaluation of our methodology and is made publicly available at [4].

Methodology and Results Overview: We apply our approach to the cell phone activity series, using the decomposed series to characterize distinct aspects of urban ecology. We proceed as follows:

First, we begin by decomposing the original cell phone activity series for each areal unit into seasonal and residual components. The decomposition is done using FFT, with the seasonal component corresponding to high-amplitude frequencies and the residual component corresponding to the deseasonalized series in the time domain. The seasonal communication series (SCS) are due to typical patterns of socio-economic activity within an area; e.g. a university generates higher traffic during weekdays and lower traffic during weekends and holidays than a residential neighborhood. The residual communication series (RCS), on the other hand, can represent irregular traffic (e.g. an area may have higher traffic than usual due to a protest or sporting event, or lower traffic than usual due to a strike) and/or due to the influence of one area on another (e.g., due to mobility and/or social interaction).

Second, we perform hierarchical clustering of different areas based on the different time series and we validate the results using the ground truth data. We show that our SCS clustering scheme successfully segments areas dominated by distinct types of socio-economic activity, and allows for discovery of regions whose activity patterns differ markedly from the rest of the city. The results compare favorably with state-of-the-art approaches such as [26], since SCS can incorporate regular patterns occurring on any time scale; by contrast, state-of-the-art methods estimate regular patterns using average weekday and weekend days (evaluated using binned averages), and cannot therefore exploit patterns of change that occur across multiple days. In addition to using the SCS to identify regular patterns, we show that its counterpart, the RCS, enables the detection of regions that are subject to mutual social influence or in direct communication contact; this was not previously possible using mere activity data. More specifically, we show that the structure of lagged spatial correlations in RCS across areas allows for the detection of regions that are subject to mutual social influence (i.e., disruptions in one area propagate to the other), and of regions that are in direct communicative contact. We validate the latter by showing that RCS correlations between areas are significantly related to the volume of inter-area cell phone traffic, and that this relationship is substantially stronger than for SCS.

In summary, the RCS and SCS provide distinct probes into the structure and dynamics of the urban environment, both of which can be obtained from the same underlying communication data.

The structure of the rest of the paper is as follows. Section 2 reviews related work. Section 3 presents our data sets, and the ground truth we collected. Section 4 presents the decomposition of the original series into SCS and RCS, and some key insights from basic properties of these series. Section 5 presents the clustering methodology. Section 6 presents the results of applying clustering to SCS and its comparison to ground truth, as well as the findings of the RCS analysis. Section 7 concludes the paper.

2. RELATED WORK

This section summarizes the most relevant related work in the intersection between urban dynamics and human activity data. Data generated by human activity can be divided into two broad categories: (1) self-reported data and (2) behavioral traces. In self-reported data, users decide to report their semantically annotated location via check-ins e.g. in Foursquare a user selects a venue from a list of venues that are detected nearby his location. In behavioral traces, users are passively monitored and do not actively select the information that is being revealed. Cell phone activity patterns are an example of behavioral traces, in which some aspects of the users' behavior are *fully* revealed, but without any *semantic information* i.e. the user location is not annotated to indicate venue or category type. This work uses cell phone activity patterns.

Behavioral traces. Cell phone activity patterns, also commonly known as Call Description Records (CDR), capture important aspects of human activity in a city [23], and they have been used to study human mobility, social networks, and urban ecology. Since the focus of this paper is on the latter, we mainly review related work in that area, and we only briefly mention other work in the broad area of cell phone activity analysis.

Toole et al. [27] used aggregated CDRs in order to infer land usage in Boston. They used a supervised learning technique: they built various features from activity series and used them to classify areas of the city into five different categories (residential, commercial, industrial, parks, and other). Their ground truth was a data set of zoning regulations from the municipality of Boston. However, the classifier's accuracy was worse than classifying every area to

belong in the dominant category, due to the high percentage of residential areas. In contrast, we use clustering, a form of unsupervised learning, our ground truth data set contains the facilities in each area, and we study the city of Milan.

Soto et al. [26] also followed an unsupervised learning approach to characterize areas in the city of Madrid. They clustered areas of the city based on their activity signature, *i.e.* the activity pattern for a typical weekday and a typical weekend, where "typical" is defined as average activity in a period of 3 months. They produce five different clusters: industrial and offices, business and commercial, nightlife, leisure and touristic, and residential. We follow a different clustering approach that facilitates selecting solutions for specific purposes (e.g., segmentation versus anomaly detection), and employ a more general time series decomposition that can be used in settings with regular patterns that do not fit into the typical weekday/typical weekend day typology. In Section 6 we compare our spectral approach to that of [26] and show that our approach allows us to detect additional information that is lost under averaging. Additionally, our work differs from [26] in that we make use of residual fluctuations in the time series, which can be used to detect interactions between areas. We note that our clustering and residual analysis techniques could be applied to the averaging scheme used by [26], and many aspects of our approach are hence complementary to this work.

Similar decompositions of activity series have been applied in other settings as well. Calabrese et al. [9] applied eigendecomposition to extract features from Wi-Fi time-series, and then used those features, produced from the top 4 eigenvectors, to cluster access points with similar traffic. We apply time series decomposition on a different type of series and we also take advantage of the residual communication to highlight important aspects of the data.

Finally, CDRs have been used for human mobility analysis, [17], [15], [8] as well as for studying social interactions [14]. These studies have answered various questions, including: what is the potential of ride-sharing for reducing traffic in a city [12]; which are the poorest areas of a city [25], etc.

Self-reported data. Work on urban dynamics and self-reported locations is primarily focused on Foursquare check-ins [13, 20, 21]. The authors of [21] use the categories of Foursquare check-ins in order to cluster similar areas of the city, i.e. two areas are similar if their normalized check-ins are also similar (e.g. both have 70% restaurant and 30% work check-ins). More specifically, each area is represented by a vector of category check-ins aggregated over a time period. Spectral clustering with a cosine similarity measure is used to create clusters of similar areas. However, the paper provides no external validation of the clustering results. In contrast, our evaluation results make use of a separately obtained ground truth data from the official city records and our method does not require semantically annotated location information.

The authors of [13] use the same clustering approach as [21], with some modifications in the distance function, and they validate they results through interviews with city residents. However, the distance function in [13] requires per-user and per-venue check-in information whereas our method is suitable with spatially aggregated data over all users in the same area.

[28] considers that the function of an area is a combination of two aspects: places of interest (POI) categories and human mobility data (e.g. number of people arriving or leaving) in a region. They cluster similar areas using a topic model-based method that combines both aspects. They evaluate their method by using well-known locations, and interviews with residents of the city. However [28] aggregates the human mobility data in typical weekday/weekend fashion, an approach that we show looses information. Additionally, their method requires as input POI categories which, unlike ours, assumes existing knowledge about the city.

3. DATA SETS

Table 1 summarizes the data sets used in our analysis.

Name	Period	Source
Milan Activity	Nov.4-Dec.1 2013	Telecom Italia [3]
Milan Square-to-Square	Nov.4-Dec.1 2013	Telecom Italia [3]
Universities, Businesses, Parks, Population per area, Sport Centers, Bus stops	Jan.1-Dec.31 2013	City of Milan [1]

Table 1: Data Sets

3.1 Cell phone data

The first two data sets were provided by Telecom Italia Mobile as a part of the *Big Data Challenge* [3] competition. They consist of telecommunications activity records in the city of Milan. In this paper, we focused on a 4-week period of November 2013. The decisions about spatial and temporal granularity in the dataset, were already made by the dataset provider (Telecom Italia) when they made the data sets available for the *Big Data Challenge*, and were out of our control. In particular, the city of Milan, an area of 550 km^2, was divided into a 100×100 square grid. Each grid square has the same dimensions: a side length of 0.235 km and an area of 0.055 km^2. This is the areal unit we use throughout the paper, and we refer to it as a "square". The temporal unit is the 10-minute interval.

In the Milan Activity dataset, the activity is aggregated within each grid square for each 10-minute interval. Each activity record consists of the following entries: square ID, time-stamp of 10-minute time slot, country code, incoming SMS activity, outgoing SMS activity, incoming call activity, and outgoing call activity. According to the data set release information [3], each activity value corresponds to the level of interaction of all the users in the square with the mobile phone network, *e.g.* the higher the number of outgoing calls made by the users, the higher the outgoing Calls activity. Values of incoming/outgoing Call and SMS activity are normalized and have the same scale. Therefore, we sum up the latter four values,[1] over all country codes in order to come up with a single value which describes the total activity volume in a square during each 10-minute time slot; this is the *original time series*, used in Section 4. Fig. 1 shows an example of the original time series for two grid-squares over a one week-period.

The Milan Square-to-Square dataset contains information regarding the "directional interaction strength" (as per terminology in [3]) between two squares, based on the calls exchanged between users in them. Each activity record consists of the following fields: ID of square i where call was initiated from, ID of unit j where the call was made to, time slot, and value of directional strength from i to j. We can obtain the total directional strength from square i to square j, by summing up over all time slots in the 4-week period.

3.2 Ground Truth

We collect additional data sets at the Municipality of Milan's Open Data website [1], in order to use them as ground truth, for external validation of clustering, in Section 6. For each square defined in the main data sets, we gathered the following *category* information: population, %green area, #sport centers, #of universities, #businesses, and #bus stops. The blue shaded area in Fig. 2 marks

[1]This aggregation helps avoid data sparsity.

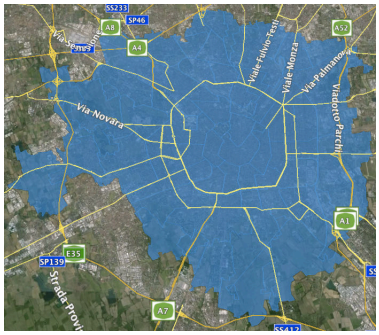


Figure 2: Local identity units of Milan. The blue shaded area shows the part of the city for which we have ground truth information. Ground truth information consist of information regarding facilities in each area, as well as census data.

Figure 3: Squares inside the local identity unit of Pagano. Some squares have full overlap, while others have only partial overlap. The Population of Pagano will be spread uniformly to the overlapping grid-squares proportionally to the overlapping area (between a square and Pagano). Also, the population of a square may be reduced even further if the square contains green areas. A square can receive population from multiple local identity units.

the neighborhoods (or "local identity units") in the city of Milan, for which category information is available.

Gathering the category data and putting them in a ready-to-use format was non-trivial. First, we performed basic data cleaning and post-processing of the collected data. This involved removing duplicate entries in each category, transformation of coordinates from the Italian Gauss-Boaga projection to standard latitude-longitude, and mapping of addresses to latitude-longitude using Google Maps API. Second, we assigned every ground truth element in each category to the corresponding square. Information that appeared as a single latitude-longitude coordinate was easy to assign to a single square; those included businesses, universities, bus stops and sport centers. However, assigning categories such as green areas (which appear as geometric shapes with multiple points) and demographic information (which is reported on top of the local identity units) to the grid squares was more challenging. In the case of the green areas, we calculate the overlapping area between a specific square and a given green area as a percentage of the square.[2] Fig. 3 shows

[2]For example, if the park was large and the whole square was inside it then the square was considered to be 100% green space. For demographic data, we calculate the overlapping area between a square and a local identity unit, we subtract the green-space area, and we assign part of the population of the local identity unit to the square, proportionally to the size of the remaining overlapping area as a percentage of the local identity unit.

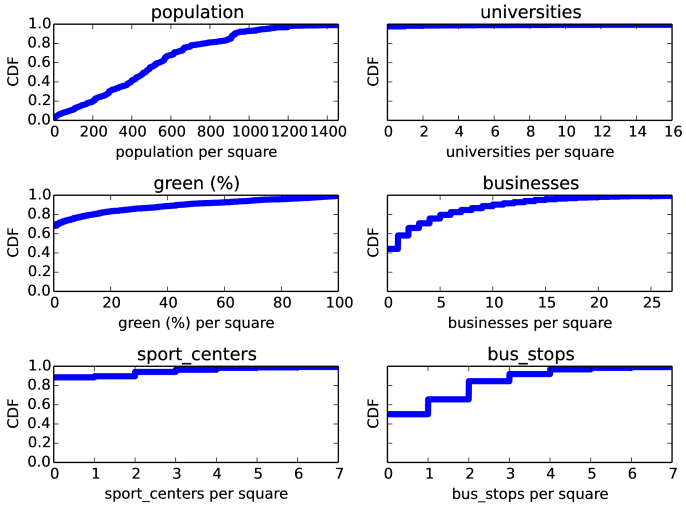


Figure 4: Number of ground truth elements in each square. The above figures indicate how uniformly each ground truth element is distributed in the city. For example we observe that the vast majority of the squares have zero universities, and universities are distributed to only a handful of squares.

the overlap between squares and local identity areas. Fig. 4 shows how ground truth elements are distributed over the grid. Finally, we make the processed ground truth dataset publicly available at [4].

Limitations: Using ground truth to evaluate urban dynamics has some inherent limitations and, in fact, leads to – conservatively – underestimating the success of our method. For example, we were constrained to use only the portion of the data (30% of the squares) for which we were able to obtain ground truth. Furthermore, a category as defined in the ground truth (e.g. "universities") may not capture subtle differences (e.g. universities in the periphery vs. universities in the center), which the time-series approach may be able to correctly capture. These challenges in evaluating clustering based on urban dynamics using ground truth data, have also been observed by previous papers [27]; while other papers choose to use interviews with city residents [26, 28] .

4. ACTIVITY IN TIME AND FREQUENCY DOMAINS

Let $S = [1, n]$ be the set of all grid squares, where $n = 10^4$. Also, let $T = \{t_1, t_2, ..., t_m\}$ be the set of all time units (10-minute time slots as defined by the data provider) in our 4-week period, where $m = 4032$. Finally, for each square $i \in S$, we denote the original activity series as $O_i(T) = \{o_i(t_1), o_i(t_2), ..., o_i(t_m)\}$.

The original time series is expected to have strong periodicity, since it depends on human activity. Therefore, we first map our time series into the frequency domain to identify the dominant seasonal components. We apply the fast Fourier transform (FFT) to convert $O_i(T)$, for all $i \in S$, from the time domain to the frequency domain. FFT is suitable for our purpose because it is a non-parametric method that extracts periodicity, it is useful for series with no obvious trend and provides a spectrogram that is easily interpretable [11]. Fig. 5 shows the power spectrum of all series in the frequency domain. We observe several frequencies with high power (e.g., weekly, daily, and 12-hour cycles). These high-power frequencies dominate the seasonal component of the com-

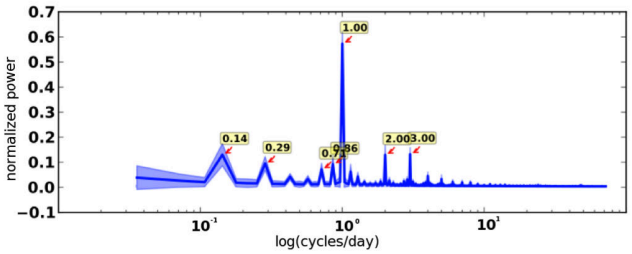
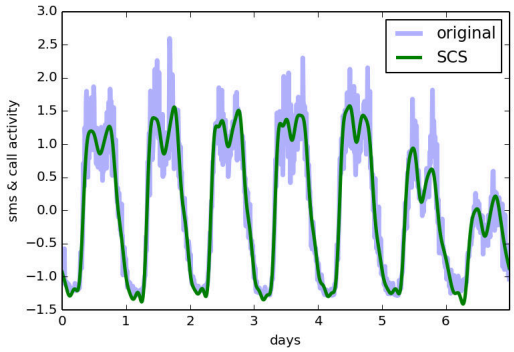
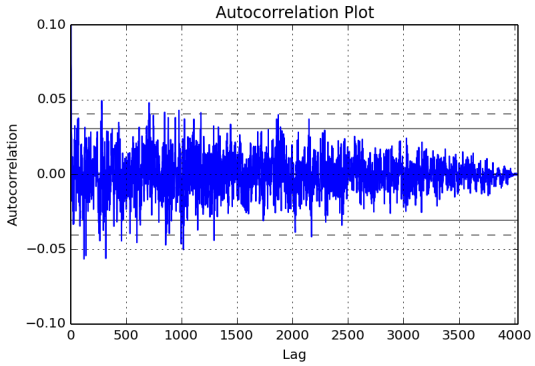


Figure 5: Power spectrum of activity aggregated over all grid squares (blue line indicates mean, shaded area ±1 std dev); marks indicate high-amplitude frequencies, e.g. daily (1 $cycles/day$) and weekly (0.14 $cycles/day$.

munication series, and indicate endemic social processes taking place within the city.



(a) Original series and derived SCS; SCS contains only systematic information, and is hence smoother.



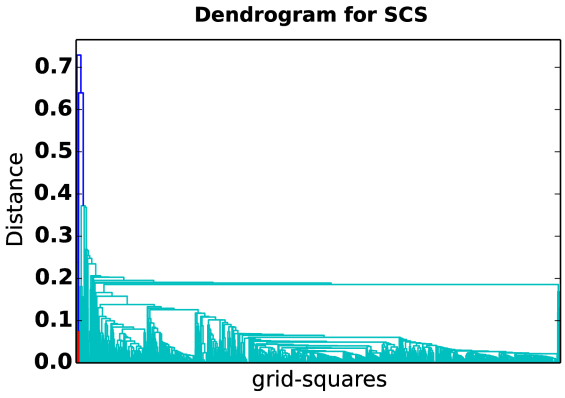
(b) RCS autocorrelation. Low autocorrelation values verify that systematic components have been removed. Note that the y-axis is zoomed in the interval (-0.1, 0.1).

Figure 6: Decomposition of original activity series for grid square 5071.

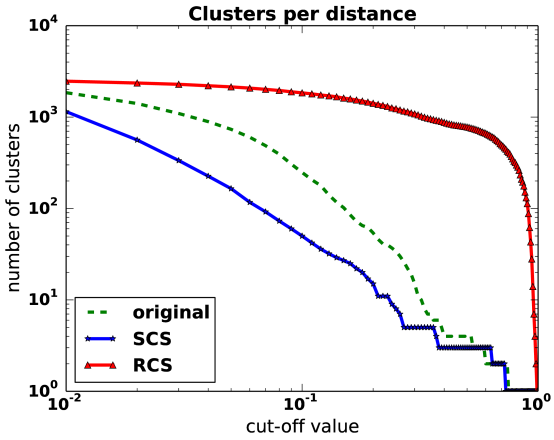
For each grid square $i \in S$, we decompose the original time-series $O_i(T)$, into *seasonal* and *residual* components through the following steps.

1. We select its k highest-power frequencies.
2. We regenerate the seasonal communication series $SCS_i(T)$ using only the top k frequencies of the grid-square, where $SCS_i(T) = \{scs_i(t_1), ..., scs_i(t_m)\}$.
3. We obtain the residual communication series $RCS_i(T)$ by subtracting the basic series from the original series: $RCS_i(T) = \{o_i(t_1) - scs_i(t_1), ..., o_i(t_m) - scs_i(t_m)\}$.

The data analysis in the remainder of the paper uses $k = 30$; this was selected by finding the smallest k such that the RCS autocorrelation function does not differ significantly from that of a white noise sequence, as shown in Fig 6(b). Please note that SCS captures regular activity (which we use in Section 6.1 for clustering of areas in the city), while RCS carries information about similarity of residual activity between two squares (which we use in Sec. 6.2.1 to study cross-square interactions.)



(a) Dendrogram

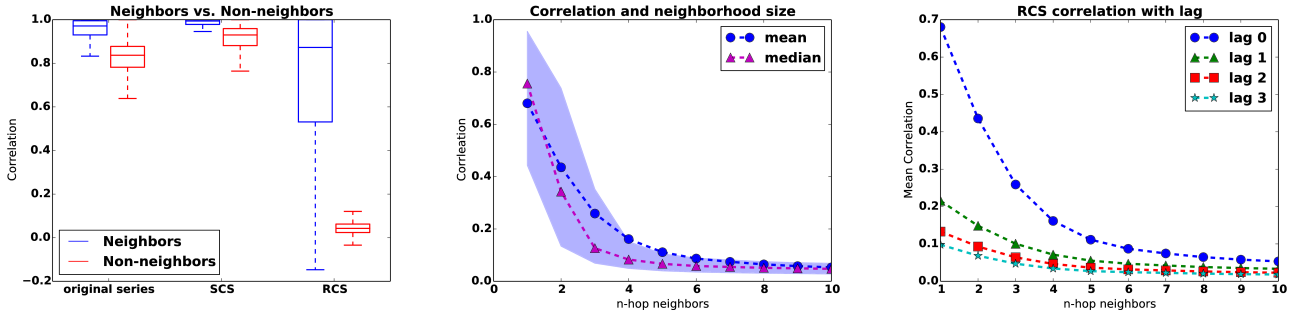


(b) Number of clusters per cut-off distance.

Figure 7: Cut-off distances and clusters. (a) shows the dendrogram for clustering the SCS. At distance 0.74 all squares have been merged into one cluster. By looking at the dendrogram we see that most squares in Milan have quite similar traffic, *i.e.* highly correlated activity series, with a few special squares that are different from the rest. Also, we summarize clustering with SCS, original, and RCS in (b), which shows the number of clusters per cut-off distance.

Insights from decomposition and neighborhoods: Here we discuss observations on the correlation of a square and its neighbors w.r.t. the different time series, namely, original, SCS and RCS. These are important to guide the clustering based on these time series, discussed in the next section.

Fig 8(a) evaluates the correlation between neighboring and non-neighboring squares for the original, SCS, and RCS series. In RCS,



(a) Correlation between neighbors vs. non-neighbors for original, SCS, and RCS.

(b) Correlation for n-hop neighbors for RCS.

(c) Correlation and lag.

Figure 8: Correlation and neighborhood. In (a) we observe that in all three cases, the correlation between neighbors is stronger than non-neighbors. However, the difference in the correlation between neighbors and non-neighbors varies. In (b) we observe that as the size of the neighborhood increases the mean correlation decreases – the filled area of the graph represent the region between 25^{th}-percentile and the 75^{th}-percentile. And, in (c) we observe that even when we look at lagged correlation for the RCS, we observe higher correlation, on average, between neighboring squares and a decline in the mean correlation as distance increases. Correlation also decreases when time lag increases.

the correlation between neighbors is much stronger than the correlation between non-neighbors; this is a clear indication that when something occurs in a square, it often spreads to its neighbors. On the other hand, the correlation between non-neighbors is close to zero, which indicates that perturbations of activity are spatially restricted; this is even more clear when we look at Fig. 8(b) that shows how the correlation among neighbors decreases as the size of the neighborhood increases. Also, in Fig. 8(c), we see that what happens in a square at time slot t_0 will affect its neighbors at a later time slot, with the effect dampening over time. This is compatible with an underlying diffusion process.

Fig 8(a) also shows that in SCS the correlation between both neighbors and non-neighbors increased when compared to the original series. Moreover, the difference of the correlation between neighbors and non-neighbors decreased. This shows that SCS is dominated by the day-to-day activity patterns, and is stripped from perturbations that are spatially constrained.

5. CLUSTERING

Our goal is to use the result of the decomposition to segment the city into distinct areas, where the members of same area would have similar activity patterns. We hypothesize that if two squares have similar communication patterns then they have similar local ecologies, whereas if their communications patterns differ then they have different local socio-economic activities; *e.g.* one would expect a university and a stadium to have different activity patterns, since people visit them during different hours.

To achieve our goal, we cluster squares via agglomerative hierarchical clustering. We employ hierarchical clustering because of its generality (requiring no particular assumptions regarding the underlying distance measure) and because it yields a family of solutions (generally expressed as a dendrogram) that contains more information than a single clustering solution. As we show, this information can be exploited to perform both segmentation and anomaly/outlier detection from a single dendrogram. The distance function used in the rest of the paper is based on the Pearson correlation between activity series. More specifically, for two grid-squares $i, j \in S$, their distance in a given activity series A is:

$$dist(A_i, A_j) = 1 - correlation(A_i, A_j)$$

This distance function takes values in the range [0,2]; when two grid-squares are fully correlated they will have a distance of 0, when they are completely uncorrelated a distance of 1, and when they are inversely correlated a distance of 2. Also, we used the average linkage criterion to build the dendrogram since it had the highest cophenetic coefficient in comparison with other linkage types – the cophenetic coefficient is a measure of how faithfully a dendrogram preserves the pairwise distances.

Fig. 7 shows the high-level clustering results. Fig. 7(a) shows the dendrogram for the clustering via SCS; by looking at it we see that for small cut-offs, *e.g.* values between 0.02 and 0.08, we get a segmentation of the city into multiple areas, while for large cut-offs, *e.g.* values higher than 0.28, we get a very large cluster and a few small ones – hence, high cut-offs can be used for segmentation and low cut-offs can be used to detect areas with anomalous activity, without requiring additional computation.

Fig. 7(b) evaluates the number of clusters generated in the original, SCS, and RCS series for the full range of cut-off distances from 0-1. We observe that the results of clustering with RCS differ significantly from that with SCS and original. More specifically, clustering with RCS yields a large number of clusters (1000s) until cut-off distance ~ 0.90. On the other hand, clustering with SCS yields a low number of clusters for cut-off distance as small as 0.10. We speculate that the reason for this result is due to the degree of correlation between neighbors and non-neighbors (also see figures 8(a) and 8(b)). Intuitively, SCS provides information on routine activities in the city. Thus, where the analyst's goal is to cluster urban areas based on regular patterns of activity, the SCS should be employed.

6. RESULTS

The results presented in this section are limited to the grid squares that overlap with the ground truth; this corresponds to the approximately 3K squares represented in the colored area of Fig. 2.

6.1 Clustering based on the SCS

SCS hierarchical clustering requires a choice of where to cut the dendrogram. Examination of the number of clusters by distance, Fig. 7(b), shows that this alone does not produce a natural cutting point for the SCS; thus, we also factor in skewness of cluster sizes,

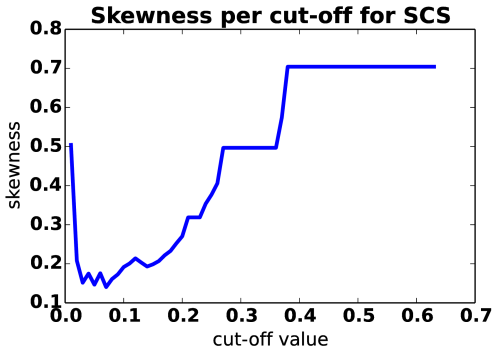


Figure 9: Skewness plot for SCS clustering. We use Pearson's second skewness coefficient: (mean-median)/(standard deviation). The minimum skewness is at cut-off $= 0.07$.

defined as (mean-median)/(standard deviation). Fig. 9 shows the calculated skewness for the full range of cut-off values in the SCS clustering. The rationale for employing skewness to guide distance selection is as follows. One expects that a segmentation of the urban environment into distinct functional areas, *e.g.* university areas, residential areas, *etc.*, will produce clusters of relatively comparable size, with correspondingly low skewness. On the other hand, if the distribution is very skewed, then there is a small number of large clusters which contain the vast majority of squares, and various small clusters with unique activity patterns that are quite different from the rest of the city, *e.g.* stadiums. By choosing low versus high-skewness cut-points, one can then choose to break the city into a few large areas of similar activities, or (respectively) detect small, anomalous areas in the city against an "average" background pattern. Both options provide distinct information, and it is not necessary to employ only one; here, we show both cases.

Low-skewness segmentation: In this case we seek to divide the city into comparably sized clusters, and hence choose a cut-point that yields a size distribution with low skewness. Per Fig. 9, we obtain this via a distance threshold of 0.07, which is the minimum skewness value in the SCS clustering. As Fig. 11 shows, the clusters of Fig. 10(a) successfully segment the city by features of the urban environment. For instance, clusters c1 and c5 have a high density of universities, while cluster c3 has a high density of green space; density is the ratio of the number of ground truth elements over cluster size. Thus, we conclude that the clusters indeed reflect regions with distinct socio-economic and environmental characteristics as reflected by their differences in SCS.

High-skewness anomaly detection: In this case we seek to identify one large cluster reflecting the range of "typical" activity patterns and several small areas of anomalous activity; we thus select a cutoff leading to a skewed distribution e.g. Fig. 10(b) with cutoff 0.30 which contains five clusters. In Fig. 12 we see the normal activity in the city in cluster c0, while the other clusters represent strongly anomalous traffic. For instance, cluster c1 corresponds to the San Siro stadium, while cluster c3 contains the Otomercato, a wholesale market for fruit and vegetable.

SCS vs. original series: In Fig. 13 we show a comparison between the clustering with SCS, and the clustering with the original in regard to coverage for the full range of cut-off distances $0-1$. We confirm that clustering with SCS yields better results. This verifies our expectation that removing idiosyncratic variation from the signal clarifies the regular activity within a grid square.

SCS vs. Typical Weekday/Weekend : Prior work [26] handles clustering of time series by aggregating cell phone activity in a typical weekday/weekend over all users in an area. Our method fits data better and requires less assumptions compared to that approach. We compare the performance of our method with that of a typical weekday/weekend using the Entropy of the density distribution for a given category, which is a common external clustering evaluation measure [29]. Table 2 shows the obtained values of Entropy for our method and the clustering via K-means and typical weekday/weekend from [26]. We observe that our method performs better for all ground truth categories. The intuition behind that is that our way of summarizing the cell phone activity is superior in the sense that the seasonal component of FFT (our SCS) holds more information than the typical weekday/weekend approach, as shown in Fig. 14. To quantitatively illustrate that point, we calculate the normalized sum of squared errors for each grid-square i, for both methods as follows:

$$e_i = \frac{\sum_{t=1}^{t=T} (O_i(t) - B_i(t))^2}{T}$$

Category	Entropy for hierarchical SCS	Entropy for [26]
Universities	0.96	0.97
Businesses	0.82	1.33
Green (%)	0.94	1.27
Population	0.97	1.34

Table 2: Segmentation performance via Entropy (lower value is better). Hierarchical SCS clustering produces more functionally distinct clusters for all categories.

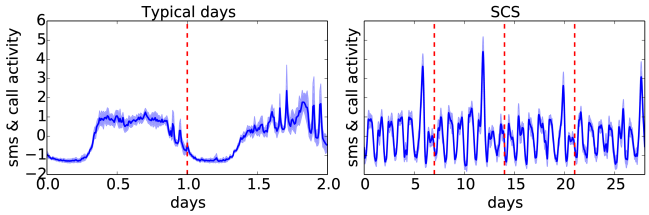


Figure 14: The left figure shows how the typical weekday/weekend approach summarizes the cell phone activity series for the San Siro area. Notice that the high peaks are lost when traffic is aggregated. The right figure shows the SCS traffic in the San Siro area for one month (our method).

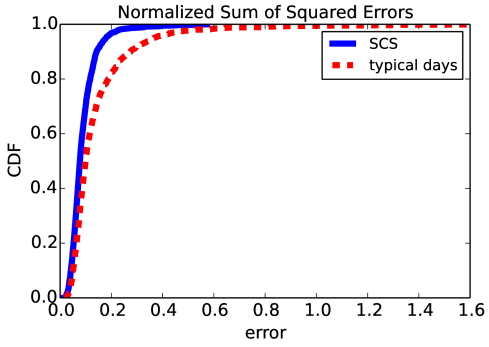
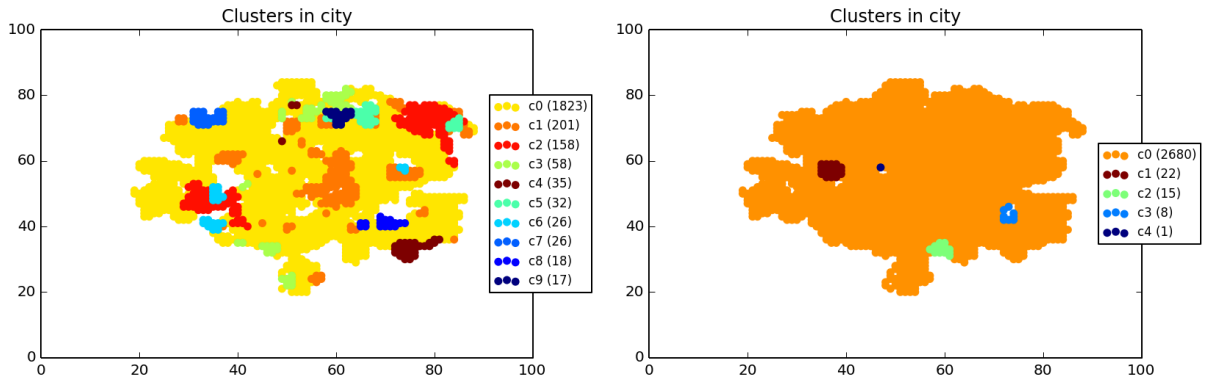


Figure 15: The figure shows how much information is lost from the original series, when we use SCS and when we use typical days. Despite the fact that SCS is created using the top-30 frequencies, while the typical days series is created using a vector of 288 values – 144 values for weekdays, and 144 for weekends – the SCS can reconstruct the original series much better.



(a) Segmenting Milan (cut-off 0.07) (b) Detecting anomalous areas (cut-off 0.30)

Figure 10: Clusters generated by low skewness (left) and high skewness (right) segmentation. When the number of clusters exceeds 10, we show only the top 10 largest clusters.

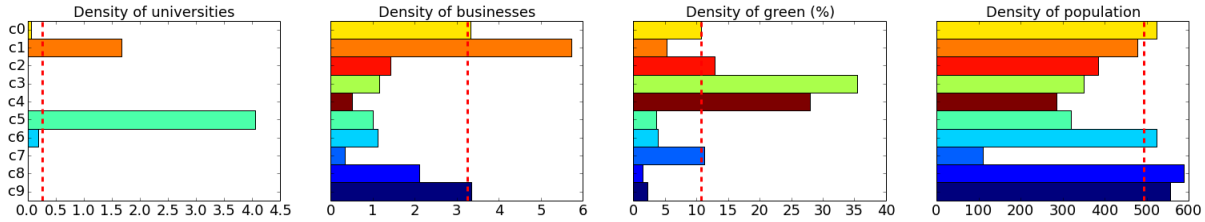
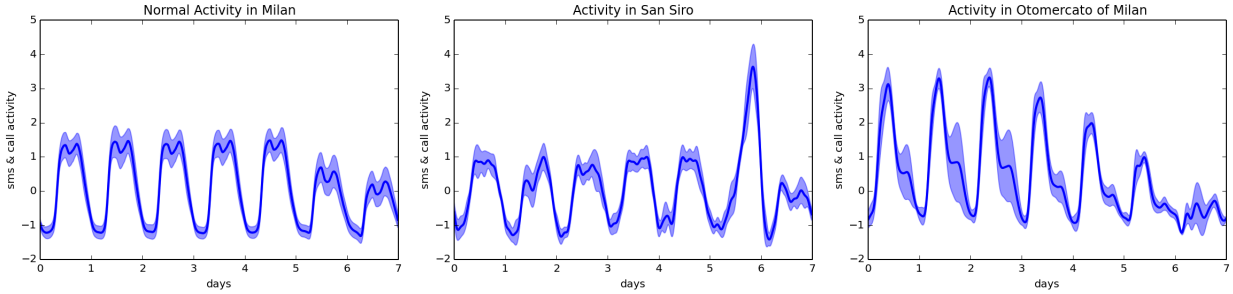
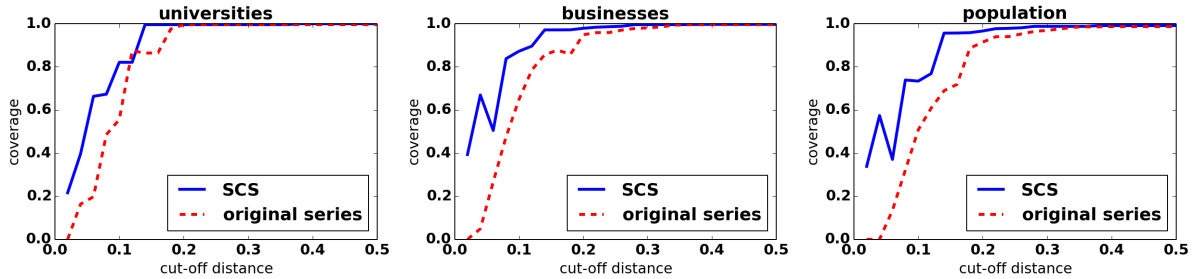


Figure 11: Densities per category for top 10 clusters of Fig. 10(a); red dashed line indicates the Milan average. SCS clustering separates regions with distinct urban environments (e.g., commercial vs. green space).



(a) Normal area (c0) (b) San Siro (c1) (c) Otomercato (c4)

Figure 12: Average activity series for the clusters of Fig. 10(b); different clusters have very distinct seasonal patterns.



(a) Universities (b) Businesses (c) Population

Figure 13: Coverage for SCS clusters vs. original series clustering. Coverage is defined as the percentage of the ground-truth elements "covered" by the clusters with higher than mean concentration of the element type. SCS clustering shows stronger segmentation (higher coverage) for all cut-off values.

In our method $B_i(t) = SCS_i(t)$, whereas in the method of [26] $B_i(t)$ is obtained by using the typical weekday five times and the typical weekend twice to create a typical week, and repeating the typical week four times to create a time series of length T, which corresponds to a month. Fig. 15 shows the cumulative distribution of the errors over all grid-squares. SCS has smaller errors which indicates that it holds more information.

Our method requires less assumptions regarding the nature of the data. It works well in a different culture, e.g. some countries, such as Egypt, designate Friday as a weekend day, while others have Thursday-Friday weekends. To apply the typical weekday and weekend division requires knowledge of the culture, while our approach does not require that. Additionally, the typical weekday approach does not capture mid-week variations that might appear in certain areas.

6.2 What we learn from RCS

RCS reflects the response to perturbations rather than regular behavior. Therefore, we treat them differently: we use cross-correlation between the RCS time series of grid squares, and we investigate how the RCS of one square affects its neighbors and how it correlates with square-to-square communications.

6.2.1 *Using RCS to study cross-square interactions*

A more powerful use of the RCS is to examine interactions between grid squares. Specifically, examination of lagged cross–correlations in the RCS for two squares shows how squares affect each other: i.e., for two cross-correlated squares, a perturbation in one square at time t will be associated with a change in the other square at a later point in time. We denote the cross-correlation "distance" for squares B_i, B_j at a given lag by:

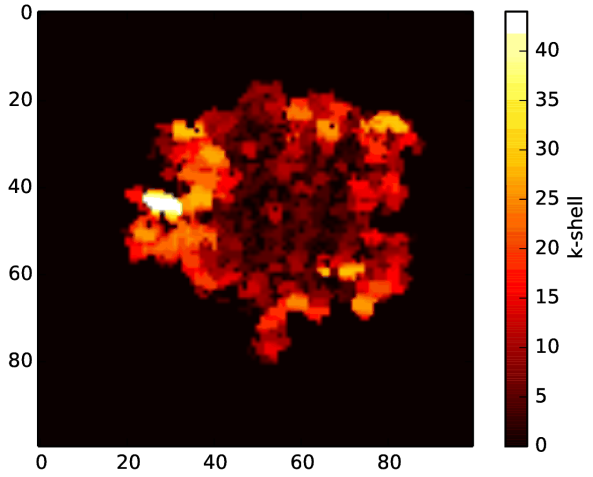
$$dist'(Bi, Bj, lag) = 1 - correlation(Bi, Bj, lag)$$

Note that, unlike a true distance, $dist'$ is not symmetric, i.e. $dist'(Bi, Bj, lag) \neq dist'(Bj, Bi, lag)$. We thus build a directed graph where the nodes are the grid squares and square i is adjacent to square j at a given lag iff

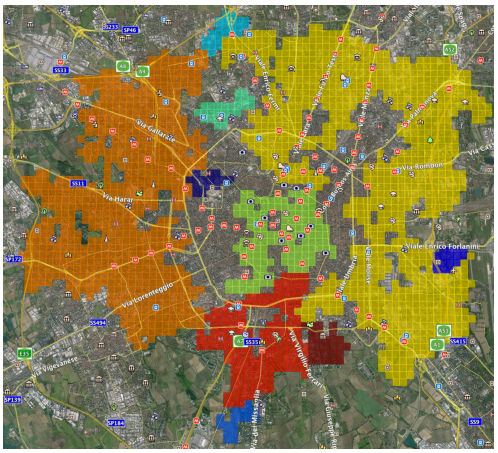
$$dist'(Bi, Bj, lag) < thresh$$

where $thresh$ is an analyst-selected threshold that filters out weak time-lagged correlations. We set $lag = 1$ which corresponds to a 10-minute difference between two squares; it is the most fine-grained unit we can get from our data. We found experimentally that a threshold 5 standard deviations away from the mean yields good results. Given the cross-correlation digraph, we may examine the strongly connected components of the graph to find areas that are subject to mutual social influence (e.g., there are paths by which events in any square can affect communication activity in any other square at a later time).

What we learn from the strong connected components of the graph: In Fig. 16(b) we show the top 10 strongest connected components of the graph, as described in the previous paragraph. We observe a different, but interesting, segmentation of the city from that obtained by SCS, with a clear structure becoming apparent. The center of the city is a connected component (green color), completely separate from the rest. This means that perturbations occurring inside a grid-square on the center, will most likely propagate to other grid-squares in the center. Another example is the dark blue connected component on the center-right side which corresponds to the Milan Linate airport. More generally, the large clusters identified in Fig. 16(b) reflect *socially connected regions* of the city, with events in any given square tending to reverberate within its regional cluster (but not to propagate beyond).



(a) Heat-map showing the k-shell score of each grid square. We observe that squares in the center of the city have lower k-shell scores in comparison with the periphery. This shows that, for this particular directed graph, squares in the periphery are more socially well-connected than squares in the center.



(b) Largest Strongest Connected Components

Figure 16: k-shell heat-map, and 10 largest strongly connected components for the directed graph of RCS cross-correlations with $lag = 1$.

What we learn from the k-shell decomposition: k-shell decomposition is a well-known technique in graph theory for identifying regions of high local cohesion [22] and has been used as a visualization tool for studying networks such as the Internet [10]. It involves identifying maximal sets of nodes with at least k neighbors who are also in the set (k-cores) and then identifying the highest-number to which each core belongs. In this article, we apply the k-shell decomposition on the RCS-based digraph, identifying the squares that are more/less cohesively connected to their neighbors; spatial regions with high values of k are strongly interactive (in the sense that perturbations in one location can propagate to other locations in the region through multiple, redundant correlation paths). In Fig. 16 we observe that squares in the center of the city have lower k-shell scores in comparison with the periphery. This shows that, within Milan, there are several spatially peripheral regions with high local connectivity, while squares near the city and along major arterials tend to be either isolated (with respect to propagation of shocks) or connected via locally tree-like structures.

6.2.2 Correlation of SCS and RCS with Milan Square-to-Square Communication

QAP test results	SCS	RCS
Correlation	**0.05**	**0.27**
Min random	-0.018	-0.005
Mean random	0	0
Max random	0.011	0.004

Table 3: Correlation of SCS and RCS inter-square correlations with Milan Square-to-Square data set, with QAP test replications ($n = 100$). Correlations are significant at $p <$0.01.

Finally, we compared the Milan Square-to-Square call volume data set with the SCS and RCS square-to-square cross-correlations, assessing the resulting relationship using the Quadratic Assignment Procedure (QAP) [19, 7]. QAP is a technique for testing an observed bimatrix statistic (here, matrix correlation) against a null hypothesis of no association, while controlling for the underlying structure of the matrices being compared; the technique is a form of matrix permutation test, in which the distribution of bimatrix statistics obtained under row-column permutation of the input matrices is used to form a null distribution.

We transformed the Milan Square-to-Square data set into a NxN matrix, where N is the number of grid squares, and we denote it as MM. An entry $MM_{i,j}$ corresponds to the symmetrized communication strength between squares i, j in the Square-to-Square data set. Similarly, we created two additional NxN matrices: 1) matrix RM, with $RM_{i,j} = RCS\_Correlation(i, j)$, corresponding to the residual correlation for squares i, j, and 2) matrix BM, with $BM_{i,j} = SCS\_Correlation(i, j)$, corresponding to the SCS correlation for squares i, j.

As we can see from Table 3, there is a significant correlation between the Square-to-Square data set both for SCS and RCS ($p <$0.01 in both cases), but for the residual series it is almost six times stronger. Thus, we see that the cross-correlations between squares are associated with direct contact between persons in the respective areas, and this is a substantially stronger effect than the baseline similarity in calling pattern within each square. This further validates our above intuition that the RCS cross-correlations provide information on social interaction across areas within the city.

7. CONCLUSION

In this paper, we studied the decomposition of cell phone activity series, via FFT, into two series: 1) the seasonal communication series (SCS) produced from high-amplitude frequencies, and 2) the residual communication series (RCS) produced after subtracting SCS from the original series. As shown, the SCS can be used to characterize typical patterns of socio-economic activity within an area, while the RCS can be used to capture both irregularities due to novel events and the influence of one area on another. For the first part, we perform an external evaluation of the produced clusters using a ground truth data set that we gathered from the municipality of Milan. Our SCS clustering, produces clusters of areas with similar characteristics as shown in ground truth data. RCS allows to identify regions such that disruptions in one area propagate to the other, and regions that are in direct communicative contact. The RCS and SCS thus provide distinct probes into the structure and dynamics of the urban environment, both of which can be obtained from the same underlying data.

Our techniques are applicable to other geo-social activity data sets,e.g. Twitter and Foursquare, and can be used to reveal patterns of how areas related to each other; in future work we plan to apply our techniques to cell phone activity data from other cities, as well as other type of geo-social activity data. These findings will provide the network operator with information that can improve planning, operations and anomaly detection. In future work, we plan to exploit our findings in order to do cell phone activity prediction.

8. REFERENCES

[1] Milan's public data. `http://dati.comune.milano.it/`, '14.

[2] World urbanization prospects: The 2011 revision. United Nations Department of Economic and Social Affairs/Population Division.

[3] Big data challenge. `http://www.telecomitalia.com/tit/en/bigdatachallenge.html`, 2014.

[4] Source code and data release. `odysseas.calit2.uci.edu/cellphoneactivity`, 2014.

[5] A. S. Berger. *The City: Urban Communities and their Problems, '78*.

[6] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, New York, second edition, 1991.

[7] C. T. Butts. Perform quadratic assignment procedure (qap) hypothesis tests for graph-level statistics., 2014.

[8] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 2011.

[9] F. Calabrese, J. Reades, and C. Ratti. Eigenplaces: Segmenting space through digital signatures. *IEEE Pervasive Computing*, 2010.

[10] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir. A model of internet topology using k-shell decomposition. *PNAS*, 2007.

[11] C. Chatfield. *The Analysis of Time Series: An Introduction*. 1995.

[12] B. Cici, A. Markopoulou, E. Frias-Martinez, and N. Laoutaris. Assessing the potential of ride-sharing using mobile and social data: A tale of four cities. In *Proc. of Ubicomp*, 2014.

[13] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proc. of ICWSM*, 2012.

[14] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *PNAS*, 2009.

[15] M. Ficek and L. Kencl. Inter-Call Mobility Model: A Spatio-temporal Refinement of Call Data Records Using a Gaussian Mixture Model. In *Proc. of Infocom*, 2012.

[16] C. S. Fischer. Toward a subculture theory of urbanism. *AJS*, 1975.

[17] V. Frias-Martinez, C. Soguero, and E. Frias-Martinez. Estimation of urban commuting patterns using cellphone network data. In *Proc. UrbComp*, 2012.

[18] R. Grannis. The importance of trivial streets: Residential streets and residential segregation. *American Journal of Sociology*, 1998.

[19] D. Krackhardt. QAP Partialling as a Test of Spuriousness. *Social Networks*, 9:171–186, 1987.

[20] A. Noulas, C. Mascolo, and E. Frias-Martinez. Exploiting foursquare and cellular data to infer user activity in urban environments. In *Proc. of MDM*, 2013.

[21] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *Proc. of SMW*, 2011.

[22] B. Pittel, J. Spencer, and N. Wormald. Sudden emergence of a giant k-core in a random graph. *Journal of Combinatorial Theory*, 1996.

[23] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning b Planning and Design*, 33(5):727, 2006.

[24] R. J. Sampson, S. W. Raudenbush, and F. Earls. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science 1997*.

[25] C. Smith-Clarke, A. Mashhadi, and L. Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proc. of CHI*, 2014.

[26] V. Soto and E. Frias-Martinez. Automated Land Use Identification using Cell-Phone Records. In *Proc. of HotPlanet*, 2011.

[27] J. L. Toole, M. Ulm, M. C. González, and D. Bauer. Inferring land use from mobile phone activity. In *Proc. of UrbComp*, 2012.

[28] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *KDD*, 2012.

[29] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. Technical report, 2001.