A holistic framework for the study of urban traces and the profiling of urban processes and dynamics

Citation	Vaccari, A. et al. “A holistic framework for the study of urban traces and the profiling of urban processes and dynamics.” Intelligent Transportation Systems, 2009. ITSC '09. 12th International IEEE Conference on. 2009. 1-6. © 2009 IEEE
As Published	http://dx.doi.org/10.1109/ITSC.2009.5309518
Publisher	Institute of Electrical and Electronics Engineers
Version	Final published version
Citable link	http://hdl.handle.net/1721.1/58813
Terms of Use	Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Proceedings of the 12th International IEEE Conference
on Intelligent Transportation Systems, St. Louis, MO,
USA, October 3-7, 2009

A holistic framework for the study of urban traces and the profiling of urban processes and dynamics

Andrea Vaccari, Liang Liu,
Assaf Biderman, Carlo Ratti
SENSEable City Lab
Massachusetts Institute of Technology
Cambridge, MA, USA
{avaccari,liuliang,abider,ratti}@mit.edu

Francisco Pereira, João Oliveirinha
Centro de Informatica e Sistemas
Universidade de Coimbra
Coimbra, Portugal
{camara,jmforte}@dei.uc.pt

Alexandre Gerber
Network and Traffic Analysis Research
AT&T Labs Research
Florham Park, NJ, USA
gerber@research.att.com

***Abstract*—Pervasive systems produce massive amounts of data as by-products of their interaction with users. Cell phone calls can inform us on how many people are present in a given area and how many are entering or leaving it. Geotagged photos can tell us where tourists go within the city and how much time they spend in each place. Descriptions of events, products, and services allow us to characterize places based on their most popular activities, products, and services.**

In this paper we illustrate a research agenda that aims at developing a holistic framework for the study of urban traces and the profiling of urban processes and their dynamics which will enable us to better understand how cities function and to develop more efficient urban policies. We also present the results of a preliminary case study in New York City where we analyzed the correlation between cell phone network handovers and traffic volumes and between semantic indexes of public events and local variations in cell phone activity. The results showed that there exist causal relationships between these types of data, and confirmed that there is strong promise in the holistic study of urban traces.

***Index Terms*—Holistic Framework, Urban Traces, Digital Footprints, Geocodable Buzz, Urban Processes, Urban Dynamics**

I. INTRODUCTION

Over the past decade there has been an explosion in the deployment of pervasive systems like cell phone networks and content aggregators on the Internet that produce massive amounts of data as by-products of their interaction with users. This data is related to the actions and opinions of people and thereby to the overall dynamics of cities, how they function and evolve over time. Electronic logs of cell phone calls and geotagged photographs are examples of *digital footprints* [1] that today allow researchers to better understand how people flow through urban space [2], and could ultimately help those who manage and live in urban areas to configure more liveable, sustainable, and efficient cities. Moreover news and descriptions of events, as well as blog posts and online reviews of products and services are forms of *buzz* that can often be geocoded to build semantic indexes of different parts of a city [3], [4].

Among the challenges to face in order to make the best use of this wealth of information is the need to create a methodology that considers the role and value of digital footprints and geocodable buzz in the larger context of urban traces that also include traditional information related to mobility features like traffic volumes, car accidents, transportation schedules, and vehicles' trajectories. In particular, we have to understand what each type of urban trace can and cannot tell us about urban processes, and to understand how we can combine the analysis of different traces to overcome their limitations and to leverage their strengths. Then, we have to formalize these operations into a framework that takes into account different scales and resolutions for each type of trace, and to develop useful tools for the profiling of urban processes and their dynamics.

One critical question in the study of cities is how cities perform in normal conditions or during special events. Until today it has been difficult to provide a precise answer to this question: traditional surveys and people counts through security cameras and satelite imageries are cumbersome in capturing the dynamics of a city and the behaviour of its inhabitants. Today, however, we have the opportunity to answer questions like this in real time by gathering, storing, analyzing, and visualizing urban traces as they are generated. Cell phone calls can inform us on how many people are present in a given area and how many are entering or leaving it. Geotagged photos can tell us where tourists go within the city and how much time they spend in each place. Descriptions of events, products, and services allow us to characterize places based on their most popular activities, products, and services.

Indeed there already exist many projects that focus on the study of digital footprints [1], [5], [6] and the implementation of new mobile services [7], [8], [9] and interfaces [10], [11], [12]. These studies, however, present two major limitations. First, they tend to consider only one data set at a time, and therefore they fall short on the inherent limitations of the particular trace (e.g. spatial resolution or statistical representativeness). Second, they do not corroborate their results with other types of urban traces and with results from more traditional methodologies typical of transportation engineering and urban planning.

We propose a new research agenda that aims at overcoming these problems by developing a holistic framework for the study of urban traces. The proposed methodology has the objective of understanding how the urban processes that define the state of a city as a complex system interact with each other and with the hidden dynamics enacted by events like a public concert or a car accident. Profiling urban processes and their dynamics will enable us to better understand how cities function and to develop more efficient urban policies.

Here we review relevant works aimed at studying mobility in urban spaces and at characterizing the semantics of places. Then we illustrate our research agenda and present the results of a preliminary case study in New York City. We show how aggregate cell phone network handovers correlate to traffic volumes between Manhattan and Brooklyn, and how public events can be classified according to semantic indexes to forecast local variations in cell phone activity. We then conclude with an analysis of the limitation of our results, and with the discussion of future works. In conclusion, we believe that our contributions are the following:

- We propose a framework to support the holistic study of urban traces, which include mobility features, digital footprints, and geocodable buzz. In doing this, we explain how we can profile urban processes and dynamics to assess and develop urban policies, moving from the concept of urban planning to that of urban programming.
- We discuss the results of a preliminary case study in New York City. Our results highlight both the potential of our proposed methodology and the limitations of the data, like spatial resolution and statistical representativeness, that need to be overcome.

978-1-4244-5521-8/09/$26.00 ©2009 IEEE

II. Related Works

A. Sensing Mobility in the Urban Context

Studies of the use of cell phone activity to estimate traffic features have been carried out since the year 2000. An extensive study was conducted by INRET, a French transportation research organization, in the Rhone Corridor, near Lyon, France in 2000. The travel time estimated by the cellular network was compared with the travel time estimation by an inductive loop sensor system. The study concluded that within a location error of 150 meters and 5% of cell phone users on the roads, the average absolute error in travel time estimation would be 10%. It also found a relationship between the volume of calls and the number of incidents on the roads [13], [14].

In 2000, the Virginia Department of Transportation, Maryland State Highway Administration, and US Wireless Corporation participated in a wireless location technology test and concluded that the technology was able to produce speed estimates of moderate quality, due to problems in generating the necessary sample size and map matching. The average mean absolute speed error in miles per hour was between 6.8 and 9.2 [13]. In 2003, Randall Cayford and Tigran Johnson investigated the parameters that could influence the effectiveness of a cell phone-based traffic monitoring system. The study concluded that a network-based location technology could provide measurements on 85% of the roads using approximately 5% of the location capacity of a single carrier, in 5-minute intervals [15].

In 2004, Fontaine and Smith studied the possibility of using a wireless network as a traffic monitoring tool instead of inductive loop sensors. The study provided guidelines for implementation of a wireless location technology-based system, and supported the deployment of this kind of system for field tests [16]. In November 2005, the University of Virginia released its Probe-based Traffic Monitoring report that reviewed 16 planned or completed deployments of wireless location technology. The report stated the lack of performance requirements by the DOTs, insufficient information to determine the quality of the data, and inadequate sample sizes for accurate travel time estimations [17].

The Center of Transportation Studies of the University of Minnesota conducted a study titled *Evaluation of Cell Phone Data Traffic* for the Minnesota Department of Transportation. The study compared data provided by the traffic information provider that uses cell phones as probes with data from inductive loop detectors, instrumented probe vehicles, and ground truth travel time calculated by matching the vehicles license plates through recorded video [18]. Table I summarizes how to get traffic features from cell phone data, and the related advantages and disadvantages.

The existing study has two major shortcomings: 1) invading the personal privacy, because the technique must sniff individual cell phone signal to get the location of individual cell phone carrier to estimate the travel speed and travel time; 2) problems of sample size: there is usually no adequate sample size of cell phones to estimate travel speed since sniffing individual cell phone will cause a lot of bandwidth consumption in the wireless network, which is the major concern of mobile carriers.

On the contrary, our approach has two advantages: 1) It does not penetrate personal privacy and only uses passive traffic log in the cell towers; 2) the 100% sample size: the mobile operators provide all the information in each cell tower, which is very useful to improve the accuracy of prediction.

TABLE I
Estimation of traffic features from cell phone data

Traffic Feature	Cell Phone Data	Advantages	Disadvantages
Demand (O-D matrix)	Handovers, location updates	Sample size, real-time data collection	Limited resolution, does not detect intra-cell trips
Volumes	Handovers, location updates	Ubiquitous coverage	Sample size, does not work in densily populated areas
Speed, Travel Time	Handovers	Ubiquitous coverage, good location accuracy	Multiple roadways within a cell
Congestions	Calls	Real-time data collection and detection	Poor location accuracy
Density	Erlang	Ubiquitous coverage	Sample size, does not work in densily populated areas

B. Perceiving Semantics in the Urban Context

From the user perspective, places are often associated with meaning, and different people relate to places in different ways. For example, a place can be described with geographic, demographic, environmental, historical, and, perhaps also commercial attributes. The meaning of place derives from social conventions, their private or public nature, possibilities for communication, and many more. As argued by [19] on the distinction between the concept of place from space, a place is generally a space with something added - social meaning, conventions, cultural understandings about role, function and nature. Often, it also has temporal properties; the same space can become different places at different times.

Taking this perspective, Rattenbury et al [3] identify place and event from tags that are assigned to photos on Flickr. They exploit the regularities on tags in which regards to time and space at several scales, so when "bursts" (sudden high intensities of a given tag in space or time) are found, they become an indicator of event of meaningful place. Then, the reverse process is possible, that of search for the tag clouds that correlate with that specific time and space. Other attempts were also made towards analysing Flickr tags [20], [21], which applied ad-hoc approaches to determine "important" tags within a given region of time [20] or space [21] based on inter-tag frequencies. However, no determination of the properties or semantics of specific tags was provided [3].

In the Web-a-Where project, Amitay et al [4] associate web pages to geographical locations to which they are related, also identifying the main "geographical focus". The "tag enrichment" process thus consists of finding words (normally *named entities* such as "New York" or "Brooklyn Bridge") that show potential for geo-referencing, and then applying a disambiguation taxonomy (e.g. "MA" with "Massachusetts" or "Haifa" with "Haifa/Israel/Asia"). The results are very convincing, however the authors do not explore the idea further than strictly geographical meaning. An extension could be to detect and associate patterns such as those referred above in [3] without the need for explicit location referencing.

In our case, we are not only interested in the semantic aspect of location representation, but also in taking advantage of information available on the Web about public places. With the rapid growth of the World Wide Web, a continuously increasing number of commercial and non-commercial entities acquire presence on-line, whether through the deployment of proper web sites or by referral of related institutions. This presents an opportunity for indentifying the information which describes how different people and communities relate to places and really perceive the *dynamics* of those places, and by that enrich the representation of a Point Of Interest. Notwithstanding the effort of many, the Semantic Web is hardly becoming a reality, and, therefore, information is rarely structured or tagged with semantic meaning. Currently, it is widely accepted that the majority of on-line information contains unrestricted user-written text. Hence, we become dependent primarily on Information Extraction (IE) techniques for collecting and composing information on the Web.

III. Research Agenda

We propose a research agenda for the development of a holistic framework to study urban traces and profile urban processes and their dynamics. In this context, we consider urban traces every type of data that can provide insights about the spatial distribution and temporal evolution of flows of dwellers, tourists, vehicles, events, and opinions in the urban context. Indeed there exist many candidate types of information that fall under this description and many more are emerging as new online services are created. For the purpose of outlining our agenda, we delineated three frames of reference to categorize this data:

Mobility Features

- Traffic volumes and people counts on major roadways;
- Traffic anomalies like congestions and car accidents;
- Spatiotemporal schedules of busses, subways, and trains;
- GPS trajectories of private vehicles (e.g. cars, taxis).

Digital Footprints

- Cell phone network activity (e.g. traffic volumes in Erlang, number of calls, text messages, and location updates, number of handovers from cell to cell);
- Geotagged photos (e.g. where and when photos were taken, descriptive tags, photographer's features like nationality and age).

Geocodable Buzz

- Semantic indexes from content aggregators and gazeteers;
- Sentiment and opinions from blogs and online reviews;
- Metadata related to the entries (e.g. popularity based on the number of linkbacks).

Mobility features, digital footprints, and geocodable buzz are here intended as perspectives from which to look at the data listed above. Mobility features can be explicitly measured to quantify flows of people and vehicles, and to measure the impact of events. Digital footprints are the by-products of people's interactions with pervasive systems and can be analyzed to study specific phenomena like tourism and social activities that characterize different parts of a city. Geocodable buzz refers to textual information that can be referenced in space and time: by mining this information it is possible to build semantic indexes that characterize events, social activities, products and services.

Urban traces, therefore, can be considered as proxies of urban processes and their dynamics, which define the functioning of the city as a complex system. Again, depending on the focus of a research some processes are more relevant than others. Here we are interested in understanding how technology and information are changing the way people move and behave in cities and how they can be leveraged to develop more sustainable cities. For this purpose, we decided to focus on four different processes: dwellers' mobility and tourists' mobility, social activities, and (adoption of) products and services.

These processes vary in space and evolve in time, and are altered by one another and by public happenings and emergencies according to hidden dynamics. Figure 1 illustrates this concept. The state of a city is defined by the urban processes, which change over time due to events that dynamically force the city to a different state. These dynamics are enacted by events like public happenings or emergencies, and by urban policies implemented by city managers. We contend that understanding urban processes and their dynamics will allow researchers to understand how people and information flow in space and time, and practitioners to develop more efficient urban policies that can readjust the state of a city to the desired one.

This research agenda requires a paradigm shift from the concept of urban planning to that of urban programming. Instead of the traditional process which involves developing long-term policies and then validating their effectiveness, iteratively adapting their implementation until the desired effect is reached, we intend to develop tools that allow urban planners to develop short-term policies, validating their effect and adapting their implementation almost in real time. The list below presents some of the questions that we plan to address in our future work:

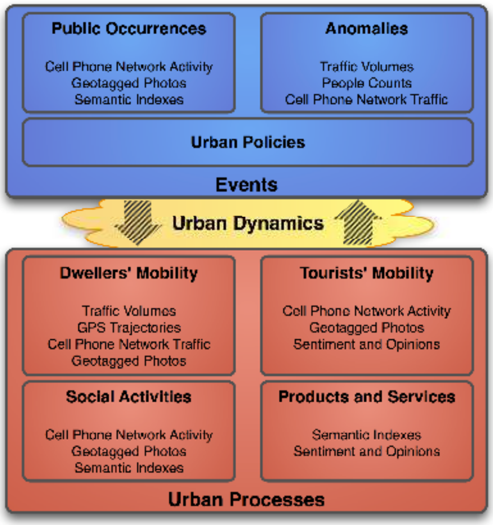

Fig. 1. Cities can be studies as complex systems. Urban processes define the state of a city, while planned and unexpected events change it. By understanding the dynamics of these changes, we can implement urban policies that readjust the state of a city to the desired one.

Dwellers' Mobility. How do dwellers and private vehicles flow in the city? How is transportation demand distributed? Where are traffic conjestions and car accidents?

Tourists' Mobility. How do tourists flow in the city? Where and when do they go? How long do they stay? How do attractiveness and popularity of hotspots evolve in time?

Social Activities. How are places characterized by the people and activities that they welcome? Which places are more attractive or more popular? How do they evolve in time?

Products and Services. How are places characterized by the products and services that people use there? How do buzz related to events, products, and services spread through space?

Planned Occurences. How do planned occurences like concerts or maintenance works impact urban processes? How can we classify occurencies and measure their accomplishment?

Anomalies. How do unexpected anomalies like car accidents impact urban processes, in particular mobility? How can we identify anomalies in real time, quantify their importance, and forecast their evolution?

Urban Policies. How can we anticipate the effects of urban policies? How can we inform urban planners in order to *configure* real time policies that readjust the effects of a occurence or anomaly?

IV. Case Study: New York City

To explore how urban traces relate to each other, we performed a preliminary case study in New York City where we studied show how aggregate cell phone network handovers correlate to traffic volumes between Manhattan and Brooklyn, and how public events can be classified according to semantic indexes to forecast local variations in cell phone activity.

A. Data Sets

Our preliminary case study was focused in the area within Lower Manhattan and West Brooklyn in New York City. We gathered three data sets, related to the three types of urban traces described in Section III:

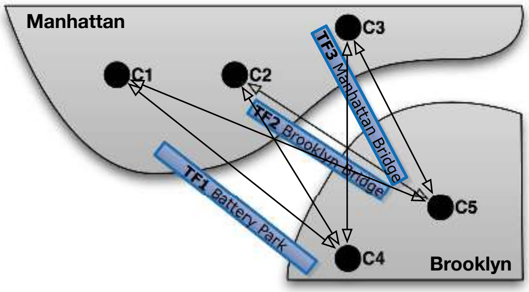

Fig. 2. Schematic representation of the spatial distribution of handovers and traffic volumes. Each cell covers a large area, making it difficult to associate handovers to a specific bridge or tunnel.

Traffic Volumes. Publicly available hourly counts of vehicles crossing the Battery Park Tunnel, Brooklyn Bridge, and Manhattan Bridge during the average working day in March 2008.

Cell phone network activity. Hourly counts of handovers between antennas in Lower Manhattan and antennas in West Brooklyn from August 2007 to August 2008. A handoff is registered when an ongoing call is transferred from one antenna to another because as the caller move out of the coverage of the former and into the coverage of the latter.

Events information. Event descriptions retrieved from *upcoming.org* related to the same area and period of the network activity data set. The data set contains description of 108 events, with the following information: name, venue, geographical location, event text description, date, time, category.

B. Cell Phone Handovers and Traffic Volumes

We studied the causal dynamics between cell phone handovers and traffic volumes with the aim of uncovering predictive relationships between cell phone network activity and traffic features. We averaged hourly handovers to predict traffic volumes and compared our forecasts with ground truth data related to an average working day. Due to the limitations in spatial resolutions of the handovers data set (the coverage of one antenna includes many roads), we limited our analysis to the traffic flows between the two boroughs of Manhattan and Brooklyn.

For each of the handovers' pairs across the river (see Figure 2) we applied the following process: we get the vector $HO_{i,j}$ (hourly handover timeseries from antenna i to antenna j) and compare it with the traffic statistics count TF, such as A↑ (upflow traffic volume timeseries), A↓ (downflow traffic volume timeseries), then we compute the correlation coefficient between the handover vector and traffic volume vector, $Coef(HO_{i,j}, A\uparrow)$ defined as:

$$Coef(HO_{i,j}, A\uparrow)= \frac{\mathrm{cov}(HO_{i,j}, A\uparrow)}{\sigma_{HO_{i,j}}\sigma_{A\uparrow}} = \frac{E((HO_{i,j}-\mu_{HO_{i,j}})(A\uparrow-\mu_{A\uparrow}))}{\sigma_{HO_{i,j}}\sigma_{A\uparrow}},$$

where σ is the standard deviation, μ the mean of each time series, E is the expected value operator and cov the covariance. We select the top N coefficient handover pairs (the coefficient greater than the predefined threshold of 0.6) and combine them together. After that, we compare the result vector with the traffic count vector to determine the best combination of handover pairs. For example, first we compare $A\uparrow$ with $HO_{13}, HO_{23}, HO_{14}, ...$ and calculate the coefficient of them. From the initial computation result, we select the top N handover pairs whose coefficient is greater than the threshold. Then we do the multi combination test to find the best matching handover pairs and scaling parameter.

We used the traffic volume in Brooklyn bridge as the test case. The results are presented in Figure 3. The red line represents traffic count data, the blue one represents handover data. For both directions, we calculated the coefficient factor of the handover pairs and traffic volume. The results show that there is strong correlation between handover and traffic count data: the correlation coefficient for eastbound is 0.8913, the correlation coefficient for westbound is 0.6516, which means this would be very useful to estimate traffic volume. More accurate estimation method should consider the calling activity patterns, which should impact the handover behavior (handover happens only during the calling activity).

TABLE II
CONCEPTS RELATED TO EVENT ID=353171 FROM UPCOMING.ORG

Concept	Score	Concept	Score
Bach	0.637	Thomas Tallis	0.637
dich	0.637	Torelli	0.637
Eli Spindel	0.637	Vaughn Williams	0.637
Gott	0.637	Fantasia	0.546
Kimberly Sogioka	0.637	donation	0.431
Mozart	0.637	Theme	0.407
Serenata Notturna	0.637		

C. Cell Phone Activity and Events

In which regards to events and cell phone activity data, we searched for causality relations between those two kinds of urban traces. On one side, we have event descriptions, and on the other side we have activity variation levels (e.g. how higher or lower than the usual is the call activity during an event). We thus performed associative and classification analysis in which the classes correspond to the activity variation levels of calls and the attributes correspond to the remaining features (e.g. event description, event time and date).

For each event, the first step is to extract the semantic index (set of words associated to an event) using a sequence of Information Extraction techniques: Part-of-Speech (POS) tagging[22], Noun Phrase chunking [23] and Named Entity Recognition (NER)[24]. POS taggers label each word as a noun, verb, adjective, etc. Then, individual noun phrases are inferred with *Noun Phrase chunking*, which concentrates on identifying *base* noun phrases - *head* noun and its *left modifiers* (e.g. Mexican food). Finally, Named Entity Recognition identifies proper names in documents and may also classify these proper names as to whether they designate people, places, companies, organizations, and the like. The output of this pipeline is a list of words that are associated to that event. We rank the relevancy with TF-IDF (Term Frequency × Inverse Document Frequency), a common measure in Information Retrieval (IR). Term Frequency measures the frequency of a word within a text while Inverse Document Frequency measures how discriminant is a word in a collection of texts (e.g. a word that appears in every text has little value since it does not differentiate the texts; a word that is unique to a document is considered to be potentially relevant).

Below, we show an excerpt of the description text for event 353171 ("The String Orchestra of Brooklyn: Winter Concert", 2007-12-15, 20:00:00, at "St. Ann and the Holy Trinity Church", Category: Music). From this text, using the meaning extraction pipeline described above, we could directly extract the concepts found in Table II.

> Bach: Erbarme dich, mein Gott from the St. Matthew's Passion Torelli: Christmas Concerto Vaughn Williams: Fantasia on a Theme by Thomas Tallis Mozart: Serenata Notturna, k239 Eli Spindel, conductor Kimberly Sogioka, Suggested donation: $10

This initial semantic index is however often very limited since event descriptions in upcoming.org are normally small. To overcome this limitation, we turn to the encyclopedic knowledge of Wikipedia. We select the top 5 words in that index and perform individual search in Wikipedia (which returns the respective page if it exists). From each retrieved Wikipedia page, we extract the abstract. The text with all abstracts together is then subject to the exact same process as described above, thus elaborating a new list of words (from Wikipedia and original text), ordered by TF-IDF. For the example above, the first words extracted include: music, suites, johann sebastian bach, style, forms, partitas, Fugue, composer, infobox, article, English, use, variants, statements, lead, changes, page, july and German.

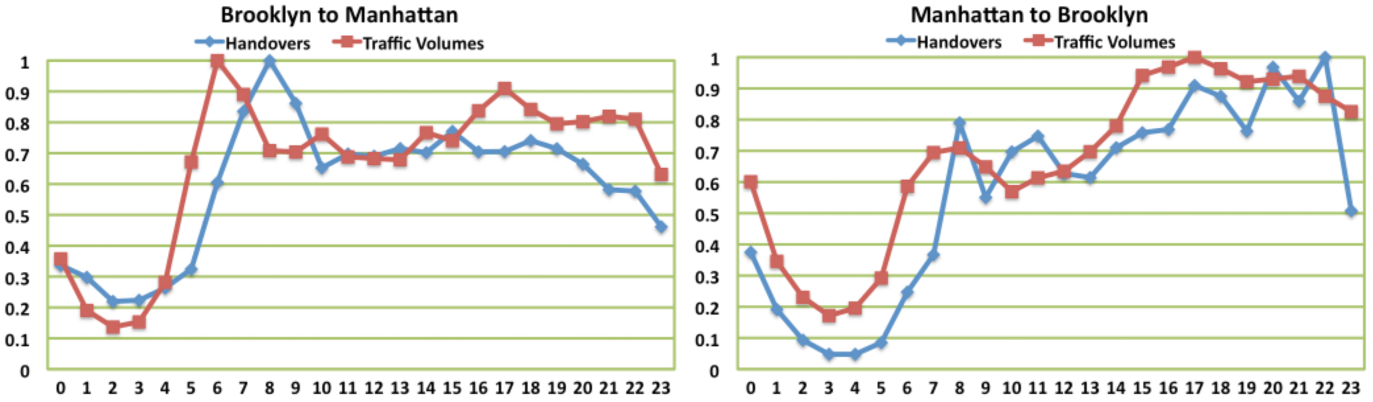

Fig. 3. Normalized hourly handovers and traffic volumes between Manhattan and Brooklyn during the average working day in March 2008.

TABLE III
WIKIPEDIA ENRICHMENT OF 15 UPCOMING.ORG EVENTS (TOP 5 WORDS)

Id	Category	Name	Top-5 Words
353171	Music	The String Orchestra of Brooklyn: Winter Concert	music, suites, johann sebastian bach, style, forms
462350	Media	Aleksey Budovskiy: Russian cartoons recent and classic	country, animators, soviet, language, arms
449040	Family	Easter is ’Egg’cellent in Lower Manhattan	families, children, symbol, nist, units
250921	Education	The Apartment (1960): Movie Nights on the Elevated Acre	star, comedy, part, core, title
396331	Performing/V.A.	Renascence: International New Media Exhibition	premiere, artists, performances, exhibition, term
447037	Other	NY Giants’ Justin Tuck Autograph Signing at J&R Music World	team, bowl, world, game, eastern
282198	Festivals	Stone Street Oysterfest	oysters, pub, term, shell, fogelson
350299	Other	Trinity Church Choir Live at J&R on 12/6	spirit, performance, people, performers, example
692856	Other	Regina Belle Performance & Autograph Signing	cd, autographs, disc, media, minutes
323193	Commercial	IBM and ACORD eForms+ Development Tour	forms, industry, acord, workshop, area
386182	Other	New York Social Network Group Dinner at the Seaport	food, cultures, ships, carmine, methods
765065	Other	Thought @ Rebar	dance, night, physics, body, form
495775	Other	JD Allen Free Live Performance	detroit, population, saxophone, census, michigan
341172	Music	From the Ocean to the Gulf	musicians, students, music, tarab, repertoire

For the 108 events from the study period and area, we extracted 724 words, of which 418 were different from each other. The word “music” was the one that appeared more times (23), followed by the words “internet” and “artists”. Overall, the full word spectrum had this distribution: 41.3% of the words show up only once, 12.9% twice, 14.5% three times, 6% 4 times, then there was a slow decay, i.e. the majority of the words appeared only once or twice, which would be expectable given the small number of events covered. This of course creates serious limitations to the pattern search. In Table III, we show an excerpt of the full list of events and respective word lists.

For the correlation analysis, we selected a number of different classification and association algorithms. For rule and decision tree induction, we chose Apriori [25], JRip [26], ID3[27], and NBTrees[28]. The analysis follows the typical data mining cycle (data preparation, feature selection, training, testing, validation), which was repeated for a number of attribute/class selections, starting with individual correlation analysis of both classes and attributes in order to reach early conclusions on the trends of the data. Particularly, it was observed that calls, sms and erlang discretized values correlated with each other in 54.91% of the instances, calls and sms in 18.41% and calls and erlang in 11.23% of the instances. So, the number of calls was correlated with at least one other class in 84.55% of the cases, which allowed us to focus on call statistics as a representative of a single class of “cell activity”. Another fact observed in this initial phase was that during the “dawn” (between 1 and 7 am), the rate of calls was considerably low, so this period becoming very sensitive to variations. The most steady call behaviour happened mostly during the afternoon.

The search for patterns was made in successive approaches: only categories and calls; categories, weekday and/or time of day, calls; categories and/or words, weekday and/or time of day.

The classifier that built the most precise model was JRip, with 72% precision and 76.5% of recall. Then, NBTrees and ID3, with values above 60% in both measures, follow in terms of success. These values reveal that it was difficult to build a classifier model that could reflect well the training set (66% of the whole set) and perform highly in the output set. The association rules algorithms (Apriori) soon became the most effective methodology, providing a number of interesting rules (although, given the small size of the dataset, their support [1] was normally low).

In Table IV, we can see that the majority of the associations could only aim for the “normal” activity class, while the most discriminant factor is category. Furthermore, we can see some surprising results: Educational events at night tend to trigger high activity; Media events (Films) affect negatively the activity; Sports events have no influence at all; Festival events lead to high activity; events during dawn trigger very high activity. While some may reflect reality (indeed, during film playing the number of calls should be low; and any event at dawn affects behaviour in general given its low profile), some others are awkward (Social events during the night do not affect activity; Commercial events during morning affect call profile, differently to the afternoon). It is our conviction that a much higher size of the dataset will unhide some actual regularity, while covering some fragile conclusions.

Another notable issue is that the higher end of the activity patterns (high and very high) allowed for more successful predictions across the models.

We believe that such approach is not only innovative but also useful: a big portion of the city mobility is directly influenced by the events that happen there. And deriving confident associations will allow for a number of important applications, namely prediction and analysis.

[1] Given an association rule $X \Rightarrow Y$, its *support* is the ratio of transactions that contain the itemset X and its *confidence* by the ratio of transactions that contain both X and Y and the ratio of transactions that contain only X

TABLE IV
Top 12 association rules considering all attributes (S/C=Support/Confidence).

Precondition	Activity	S/C
category=Music time=afternoon	normal	35/35
category=Performing/Visual Arts day=weekday	normal	25/25
category=Other day=weekend	normal	25/25
category=Commercial day=weekday	high	15/15
category=Commercial time=morning	high	15/15
category=Festivals	high	10/10
category=Education time=morning	normal	10/10
category=Performing/Visual Arts day=weekend	high	10/10
category=Performing/Visual Arts time=night	normal	10/10
category=Education day=weekend time=afternoon	normal	10/10
word=request	normal	6/6
category=Sports	normal	5/5
time=dawn	very high	5/5
category=Education time=night	high	5/5
category=Media day=weekday	very low	5/5
category=Media time=morning	normal	5/5
category=Commercial day=weekend	normal	5/5
category=Commercial time=afternoon	normal	5/5
category=Social time=night	normal	5/5
day=weekend time=morning	normal	5/5

V. Conclusions

The availability of real time urban traces like mobility features, digital footprints, and geocodable buzz is opening new possibilities for the profiling of urban processes and their dynamics. While existing projects have started to explore the opportunities offered by one type of data or another, we contend there is a need for an holistic framework that can take into considerations all the different types of urban traces and that can provide useful tools to measure and anticipate the functioning of cities, here described in term of processes and dynamics.

In this paper we illustrated a research agenda that aims at developing such framework, and we then showed the results of our preliminary case study in New York City. Our agenda is based on a model of the city as a complex system where urban processes define the state of a city, while planned and unexpected events change it according to hidden dynamics. Urban processes include dwellers' and tourists' mobility, social activities, and adoption of products and services, while events include public occurences like concerts and anomalies like car accidents.

Profiling urban processes and their dynamics will enable us to better understand how cities function and to develop more efficient urban policies. For this reason, we formulated a set of foundamental questions related to urban processes, events, and policies that we intend to address in our future work.

In our preliminary case study we analyzed the correlation between cell phone network handovers and traffic volumes and between semantic indexes of public events and local variations in cell phone activity. These first results show that there exist causal relationships between these types of data, and confirm that there is strong promise in the holistic study of urban traces for the profiling of urban processes and their dynamics. They also highlighted the limitations of these data sets: aggregate network activity has to be complemented with other traces to study mobility or social activity in small geographic areas, and the impact of public events has to be validated on larger data sets as well as against other types of traces like traffic volumes and geotagged photos.

Our future works will include a deeper exploration of both case studies using larger data sets, and the integration of other three types of traces: schedules of public transportation, GPS trajectories, and geotagged photos. With this additional data, we intend to develop new urban indicators (i.e. quantitative masures of specific aspects of a urban process), extending the work presented in [2]. Ultimately, our goal is to identify precise mathematical models that can be extended to the majority of cities, and to demonstrate their validity with more detailed experiments similar to the ones mentioned here.

References

[1] F. Girardin, F. Calabrese, F. Dal Fiore, C. Ratti, and J. Blat, "Digital footprinting: Uncovering tourists with user-generated content," *IEEE Pervasive Computing Magazine*, vol. 7, no. 4, pp. 36–43, 2008.

[2] F. Girardin, A. Vaccari, A. Gerber, and C. Ratti, "Quantifying urban attractiveness from the distribution and density of digital footprints," *Journal of Spatial Data Infrastructure Research*, vol. 4, pp. 175–200, 2009.

[3] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2007, pp. 103–110.

[4] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2004, pp. 273–280.

[5] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *IEEE Pervasive Computing Magazine*, vol. 6, no. 3, pp. 30–38, 2007.

[6] M. Gonzàlez, C. Hidalgo, and A. Barabàsi, "Understanding individual human mobility patterns," *Nature Magazine*, 2008.

[7] A. Pawling, T. Schoenharl, P. Yan, and G. Madey, "Wiper: An emergency response system," *6th International Conference on Information Systems for Crisis Response and Management*, 2008.

[8] M. Arikawa, S. Konomi, and K. Ohnishi, "NAVITIME: Supporting pedestrian navigation in the real world," *IEEE Pervasive Computing Magazine*, 2007.

[9] R. Murty, A. Gosain, M. Tierney, A. Brody, A. Fahad, J. Bers, and M. Welsh, "Citysense - an urban-scale wireless networking testbed," *IEEE Conference on Technologies for Homeland Security*, 2007.

[10] J. Liu and F. Zhao, "Towards semantic services for sensor-rich information systems," *2nd IEEE/CreateNet International Workshop on Broadband Advanced Sensor Networks*, 2005.

[11] A. Santanche, S. Nath, J. Liu, B. Priyantha, and F. Zhao, "Senseweb - browsing the physical world in real time," *Microsoft Research*, 2006.

[12] S. Eisenman, N. Lane, E. Miluzzo, R. Peterson, G. Ahn, and A. Campbell, "Metrosense project - people-centric sensing at scale," *ACM Conference on Embedded Networked Sensor Systems*, 2006.

[13] B. Smith, H. Zhang, M. Fontaine, and M. Green, "Cellphone probes as an atms tool: Smart travel lab report no. stl-2003-01," Center of Transportation Studies, University of Virginia, Tech. Rep., 2003.

[14] A. Hendricks, M. Fontaine, and B. Smith, "Probe-based traffic monitoring: Nchrp project 70-01," University of Virginia Center for Transportation Studies, Tech. Rep., 2005.

[15] R. Cayford and T. Johnson, "Operational parameters affecting the use of anonymous cell phone tracking for generating traffic information," in *Proceedings of the Transportation Research Board 82nd Annual Meeting. Transportation Research Board*, 2003.

[16] M. Fontaine and B. Smith, "Improving the effectiveness of traffic monitoring based on wireless location technology: Vtrc 05-r17," Virginia Transportation Research Council, Tech. Rep., 2005.

[17] Y. Yim, "The state of cellular phobes: Report ucb-its-prr-2003-25," California PATH, Tech. Rep., 2003.

[18] H. Liu, "Evaluation of cell phone data traffic: Cts project number 2007022," Center for Transportation Studies, University of Minnesota, Tech. Rep., 2003.

[19] S. Harrison and P. Dourish, "Re-place-ing space: the roles of place and space in collaborative systems," in *CSCW '96: Proceedings of the 1996 ACM conference on Computer supported cooperative work*. New York, NY, USA: ACM Press, 1996, pp. 67–76.

[20] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins, "Visualizing tags over time," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM, 2006, pp. 193–202.

[21] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," in *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. New York, NY, USA: ACM, 2006, pp. 89–98.

[22] K. Toutanova, D. Klein, and C. Manning, "Feature-rich part-of-speech tagging with a cyclic dependency network."

[23] L. Ramshaw and M. Marcus, "Text Chunking using Transformation-Based Learning," in *Proceedings of the 3rd Workshop on Very Large Corpora: WVLC-1995*, Cambridge, USA, 1995.

[24] V. Krishnan and C. D. Manning, "An effective two-stage model for exploiting non-local dependencies in named entity recognition," in *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 1121–1128.

[25] T. Scheffer, "Finding association rules that trade support optimally against confidence," in *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, 2001, pp. 424–435.

[26] W. W. Cohen, "Fast effective rule induction," in *Twelfth International Conference on Machine Learning*, 1995, pp. 115–123.

[27] R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[28] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," in *Second International Conference on Knoledge Discovery and Data Mining*, 1996, pp. 202–207.