An Empirical Study of Combining Participatory and Physical Sensing to Better Understand and Improve Urban Mobility Networks

Xiao-Feng Xie
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
WIOMAX LLC
PO Box 19184
Pittsburgh, PA, 15213
xie@wiomax.com

Zun-Jing Wang
Department of Physics
Carnegie Mellon University
Pittsburgh, PA 15213
WIOMAX LLC
PO Box 19184
Pittsburgh, PA, 15213
wang@wiomax.com

4997 words + 10 figures + 0 tables

November 14, 2014
ABSTRACT
The rapid rise of location-based services provides us an opportunity to achieve the information of human mobility, in the form of participatory sensing, where users can share their digital footprints (i.e., checkins) at different geo-locations (i.e., venues) with timestamps. These checkins provide a broad citywide coverage, but the instant number of checkins in urban areas is still limited. Smart traffic control systems can provide abundant traffic flow data by physical sensing, but each controlled region only covers a small area, and there is no user information in the data. Here we present a study combining participatory and physical sensing data, based on 3.4 million checkins collected in the Pittsburgh metropolitan area, and 125 million vehicle records collected in a sub-area controlled by an adaptive urban traffic control system. Our aim is to disclose how we could utilize the combined data for a better understanding on urban mobility networks and activity patterns in urban environments, and how we may take advantage of such combined data to improve urban mobility applications such as anomaly traffic detection and reasoning, topic-based nontrivial traffic information extraction, and traffic demand analysis.
INTRODUCTION
Understanding human mobility (1, 2, 3, 4, 5) and activity patterns in urban environments is a significant and fundamental issue from various perspectives, e.g., understanding regional socio-economics, improving traffic planning, providing local-based services, and promoting sustainable urban mobility. Traditionally, relevant information is however rarely obtained due to difficulties and costs in tracking the time-resolved locations of individuals over time.

In recent years, an increasing attention has been placed on introducing smart traffic control systems (6, 7) into urban road networks. The primary objectives of such systems are to reduce travel time, resolve traffic congestion, and reduce vehicle emissions. Recent work in real-time, decentralized, schedule-driven control of traffic signals (7, 8) has demonstrated the strong potential of real-time adaptive signal control in urban environments. The smart and scalable urban traffic control system (9) achieved improvements of over 26% reductions in travel times, over 40% reductions in idle time, and a projected reduction in emissions of over 21%, in an initial urban deployment. To facilitate effective real-time control, vehicle flows can be monitored by different physical sensors, e.g., induction loops and video detectors, and pedestrian flows can be detected and inferred using push-buttons or other devices. Although traffic flow data in fine granularity can be logged in real time, usage of these physical sensing data is often limited to the region that is being controlled. For the broad uncontrolled regions instead, no information is available.

The rise of location-based services provides us another way to achieve the information of human mobility, in the form of participatory sensing (10, 11, 12). With mobile devices, users can share their digital footprints at various geo-locations (i.e., venues) with timestamps through check-ins, e.g., geo-enabled tweets and geo-tagged photos and videos. Using of these services grows fast worldwide, although the instant sampling rate of trajectories is still very limited, and some web-based services, e.g., Waze and Facebook, do not open their location-based data to public. Research work has been conducted to understand temporal, spatial, social patterns, and some combined patterns of human mobility (13, 14, 15, 16, 17, 18, 19, 20, 21, 22).

In this study, we focus on exploring the potentials of combining participatory and physical sensing data in urban mobility networks. The participatory sensing data are collected in the Pittsburgh metropolitan area with the APIs provided by the location-based services including Twitter, Foursquare, Flicker, Picasa, and Panoramio. The physical traffic flow data are collected in an area controlled by the scalable urban traffic control system (9). Basic spatial and temporal characteristics are displayed for both the physical and participatory sensing data.

We first study human mobility patterns for a better understand of the urban mobility network. User checkins are examined to disclose the distribution of user behaviors, a fundamental statistical property of mobility pattern. Geo-location based cluster analysis is performed to identify personal favorite places of users in the studied regions. User entropy is measured to reveal the degree of predictability of user activities. Time-dependent mobility patterns are analyzed to show the regularity of user behaviors, based on primary and secondary most visited places of users.

We then further inform how we may benefit from combining physical and participatory sensing data to improve urban mobility applications with three examples. First, we evaluate the attraction and limit of using sensing data for anomaly detection and reasoning of traffic flow. Second, we examine if nontrivial information could be extracted from participatory sensing data to effectively recognize traffic congestion in temporal and spatial dimensions. Third, by choosing two zones in the controlled region, we perform a close check on the correlation between physical and participatory sensing patterns. For the two zones, we also illustrate how we may investigate
the origin and destination (O-D) patterns that are valuable for urban mobility from the transitions between user checkins.

sensing data

![Image](image_url)

(a) Participatory Sensing Region  
(b) Physical Sensing Region  
(c) Checkin Locations  
(d) Heat Map of Checkins

FIGURE 1: Participatory and Physical Sensing Regions in Pittsburgh, PA.

DATA DESCRIPTION

Data Collection

We implement our study in the Pittsburgh metropolitan area. The participatory sensing data contains a list of checkins. Each checkin can be represented as a tuple \(<userID, venueID, time, [comment]\)>, where userID is associated with a unique user, venueID is associated with a venue at the geo-location of \((latitude, longitude)\) with the precision of six decimal places. Our checkin data were collected (between March and July of 2014) from the geo-APIs of some location-sharing services, including geo-enabled tweets from Twitter and geo-tagged photos from Flickr, Picasa
Xie and Wang

and Panoramio. We also included existing checkin data directly crawled from Foursquare (23). For studying urban mobility patterns, we only consider checkins at venues in the spatial latitude/longitude bounding box of (40.309640, -80.135014, 40.608740, -79.676678), as shown in Figure 1a. For studying up-to-date patterns, we only consider the recent data within the range of dates [1/1/2012, 7/1/2014]. The collected data contains 3,399,376 checkins of 74,658 users at 2,198,572 venues.

The physical traffic flow data are collected in a road network which is currently controlled by the smart and scalable urban traffic control system (7, 8, 9) (see Figure 1a, and for more details see Figure 1b). The system was first installed on nine intersections (A to I) in the East Liberty neighborhood since June 2012 (see the pink region in Figures 1a and 1b), and then expanded to nine more intersections (J to R) in the Bakery Square neighborhood since October 2013 (see the green region in Figures 1a and 1b). The total area includes five major streets, Penn Ave, Centre Ave, Highland Ave, East Liberty Blvd, and Fifth Ave, with dynamic traffic flows throughout the day. For collecting real-time flow data, detectors were deployed on each entry/exit lane at the near end of each intersection and on entry lanes at the far end of boundary intersections. For each detector, a vehicle record is generated at the time when a vehicle is detected to pass the detector. Compared to that in a checkin, there is no userID and comment information available in a vehicle record. In total, the traffic flow data contains 125,369,318 vehicle records generated at 126 stop-bar detectors by the end of 7/1/2014.

**Basic Spatial and Temporal Characteristics**

**Spatial Distribution of Checkins**

We first display the spatial distribution of all checkins in the whole region. Figure 1c shows the geo-distribution of checkins. It reflects the highly non-uniform dynamics of human mobility in the urban area. Figure 1d shows the heat map of checkins, where the high-density regions are clearly colored in red. Notice, one of the red regions in Figure 1d overlaps with the controlled region in Figure 1b.

**Temporal Patterns**

We are interested in the recurrent nature of human mobility over time. To investigate temporal mobility patterns, each week is segmented into $24 \times 7 = 168$ hourly bins (starting from Monday). For obtaining seasonal results, each year is divided into four seasons (A to D), and the binned results are averaged over 13 weeks in each season. We considered all four seasons in 2013 and the first two seasons in 2014.

Figure 2a gives the seasonal checkin patterns in the participatory sensing region. It shows that the number of checkins has increased significantly over the seasons. Figure 2b shows the checkin frequency normalized by the total checkin size in each season. It shows that the checkin frequency has similar patterns for different seasons. The social day (11, 20) of Pittsburgh starts at around 4AM, and the checkin frequency peaks at around 8-9PM. A high checkin activity during Sunday is disclosed by Figure 2b.

For vehicle flow in the controlled region in Figure 1b, we consider two pivotal intersections which service most vehicles in this road network: the intersection D of Centre Ave and Penn Ave in East Liberty, and the intersection P of Fifth Ave and Penn Ave in Bakery Square. For intersection P, the vehicle flow data is only available for the two seasons in 2014. The seasonal average vehicle flow patterns with physical sensing data of the intersections D and P are respectively shown in Fig-
FIGURE 2: Seasonal Temporal Patterns for the Participatory and Physical Sensing Data.
During weekdays, the traffic flow shows three peak values, which are respectively in the morning (8AM), middle day (12PM), and afternoon (5PM) (i.e. Monday through Friday, corresponding to the first five pattern curves in Figures 2c and 2d). While during weekends, the traffic pattern shows a decreased peak-traffic-flow value and only one peak value at around 1PM (i.e. Saturday and Sunday, corresponding to the last two pattern curves in Figures 2c and 2d). It turns out that the traffic flow in weekdays is heavier than that in weekends.

UNDERSTANDING HUMAN MOBILITY PATTERNS

User Checkin Statistics

Figure 3 presents some basic user checkin statistics in the participatory sensing region. Let $N_i$ be the number of checkins for user $i$, it is shown that the probability distribution of $N_i$ follows a scaling law (see Figure 3a). We now further consider the distribution of the radius of gyration ($r_g$) for each user. For user $i$, let $V_i^j$ be the venue of the $j$th checkin, the radius of gyration is then $r_g^i = \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (\text{dist}(V_i^j, V_c))^2}$, where the $\text{dist}$ function gives the distance between a checkin $V_i^j$ and the center of mass $V_c$ for the checkins of user $i$. As shown in Figure 3b, most user activities are confined to a limited neighborhood within 10 km, which is similar to the finding in ref. (4). Figures 3c and 3d show respectively the distributions of the time and distance intervals between consecutive checkins made by users. The probability of inter-checkin time decreases as the time increases, and interestingly, it shows apparent daily and weekly patterns (see Figure 3c). The probability of inter-checkin distances (Figure 3d) is quite similar to that of $r_g$ in Figure 3b.
Finding Checkin Places
Each user tends to stay in a limited number of places, where each place is defined to accommodate similar checkins/activities of the user in its vicinity (17, 24, 25). This definition of places is able to tolerate some geo-location tracking errors. Notice that some existing tracking techniques (e.g., mobile tower) might have the location accuracy as low as several hundred meters (25, 26).

For our collected data, we identify the checkin places by clustering, where each cluster \( C \) defines the place of a set of checkins. We use an unsupervised clustering method, DBSCAN (27), which is a density-based clustering algorithm that requires only two parameters, i.e., \( \text{eps} \) to define a neighborhood threshold and \( \text{minPts} \) to define a density threshold. By default, \( \text{eps} = 250 \text{ meters} \) and \( \text{minPts} = 2 \) points were used.

User Entropy
User entropy (25) is a fundamental quantity to capture the degree of predictability for a user. For user \( i \), let \( C_i^k \) be the \( k \)th cluster, and \( K \) be the number of clusters. After the clustering, the temporal-uncorrelated user entropy \( S_i \) is defined as
\[
S_i = -\sum_{k=1}^{K} p_i(k) \log_2 p_i(k),
\]
where \( p_i(k) = \frac{|C_i^k|}{\sum_{k=1}^{K} |C_i^k|} \) is the probability that cluster \( k \) was visited by user \( i \). A lower user entropy means a higher degree of predictability for visitation patterns. Figures 4a and 4b respectively show the user entropy versus the checkin numbers \( (N) \) and the cluster numbers \( (K) \). The users in a single cluster are not shown in the log-scale figures since their user entropy are 0. As shown in Figure 4a, a user with a large number of checkins might have very low user entropy. The trend is more reasonable and clearer in Figure 4b, where a larger cluster number often leads to a higher user entropy.
FIGURE 5: Vehicle Flow Pattern in Year 2013 for the Turning Movement CDE (in Fig. 1b) at Intersection D. (The Data in 09/01, 09/02, and 10/02 are Excluded due to the Disk Full Issue.)

Regularity
We study the mobility regularity of users by computing the probability of finding users in their primary and secondary most-visited places (which is ranked using $p_i(k)$) at hourly interval in a week, and we use DBSCAN clustering with $\text{eps} = \{250, 1000\}$ meters. Figure 4c shows the time-dependent mobility patterns for the primary and secondary most visited places of users.

We first check the case for the primary most-visited places. The regularity is high during the night, and it peaks at around 4AM, the start of a social day. In weekdays, there are minima during morning (around 8AM), middle day (1PM), and afternoon (6PM), which are corresponding to the transitions to other places for commuting or having lunch. While in weekends, there is no apparent local minimum during morning, i.e. before 10AM. This pattern indicates that the primary most-visited places contain a large portion of home places.

For the secondary most-visited places, the regularity peaks at around 8AM. In weekdays, the regularity curve reaches minima during night. In weekends, however, the regularity curve shows little difference between daytime and night. Therefore, the secondary most-visited places are very likely associated with work places.

Notice that when we change $\text{eps}$ from 250m to 1000m (i.e., clustering becomes more coarsed), the increase in probability for the primary most-visited places is much more than the decrease in probability for the secondary most-visited places. This means that many users perform their other activities (i.e. except for the primary and secondary activities) near their home places. During night, the probability is nearly 90% for the primary place and 10% for the secondary place.

PROMISES FOR URBAN MOBILITY APPLICATIONS
In this section, by presenting three examples, we inform how we may take advantage of the combined participatory and physical sensing data to improve urban mobility applications.

Anomaly Detection and Reasoning
In 2013, the bridge on South Highland Ave (location shown in 1b) was closed for replacement. This cut off the connection between Shadyside and East Liberty. Therefore a large portion of traffic (including all bus lines) on Highland Avenue was forced to pass through intersection D, which is a pivotal node that services most vehicles in this road network.

Figure 5 shows the vehicle flow pattern in year 2013 for the right-turn movement at intersection D (see Figure 1b, from C to D to E). The flow significantly increased between early March
and late October. This is a typical change point problem for anomaly detection in time series of traffic. Using analysis techniques of statistics (28) on traffic data, we can approximately detect change points, but we are not able to find the exact time and the reason leading to the events.

Analyzing participatory sensing data is helpful at this point. A search of our geo-tagged data gives five checkins with the time of the bridge reopening ceremony. For example, at 5:47PM, October 23, a user mentioned joyfully, Wow the bridge on South Highland is open. Life was rough for a while. We did not find a checkin associated with the bridge closure for our collected data, which might be due to the closure event happened too long time ago and the earlier data was not kept by our location-sharing services. A search of the non-geo-tagged Tweets pinpoints the time both for closure and reopening events (for which nine and fourteen tweets have been found respectively, posted by some of the users or their first-degree friends). In further studies, fusing non-geo-tagged and geo-tagged information will be very useful for anomaly detection and reasoning, especially in case the available geo-tagged checkin information is not sufficient. In fact, many users generate both non-geo-tagged and geo-tagged information in practice. In presence of a high user regularity as shown in Figure 4c (which is a very common case), the locations of non-geo-tagged information of the users can be inferred from their geo-tagged checkins. In addition, locations can also be estimated using content-derived information (29).

**Topic-Based Traffic Information Extraction**

Many users generate checkins with the information of their current traffic conditions in travel. Tweet semantics has been used to build indicators for long-term traffic prediction (30). By taking
the search topic “traffic” as an example, we examine more potential usages of the participatory sensing checkins for obtaining important traffic information.

Figure 6a shows the temporal patterns of the checkins for the main “traffic” topic. We can clearly see that checkins mainly happened during the morning (around 8AM) and afternoon (around 5PM) of weekdays, especially the latter, reflecting the rush hours by commuters. Figure 6b further shows the temporal patterns of the checkins for both “traffic” (as the main topic) and “accident” (as the sub-topic). It discloses that people should pay more attention to avoid accidents during the morning rush hours, especially for the peaks in Figure 6b.

![Figure 6a](image1.jpg)
![Figure 6b](image2.jpg)

**FIGURE 6**: Temporal Checkin Patterns for Sub-Topics of the “Traffic” Topic.

Figures 7a and 7b show the spatial patterns of the checkins on the main “traffic” topic, where each red circle point represents a checkin. A global view (see Figure 7a) indicates most checkins are located on major highways and congested neighborhoods, while a local view (see Figure 7b) shows most checkins are located on intersections in the urban road network. These par-

![Figure 7a](image3.jpg)
![Figure 7b](image4.jpg)

**FIGURE 7**: Spatial Checkin Patterns for Sub-Topics of the “Traffic” Topic.
ticipatory sensing information is nontrivial for marking popular route choices and for identifying congested intersections and road segments where traffic conditions should be improved significantly.

Figure 7c gives the checkin results using sub-topics on major roads, including “I-376”, “I-279”, “Rt-28”, “Rt-51”, and “Penn Ave”, where the checkin points are colored differently for each road. It turns out that the segments with heavy traffic on these roads can be quite precisely pinpointed with the checkin results based on these sub-topics. Figure 7d shows the checkin results using sub-topics on “accident”. The spatial checkin distribution is surprisingly broad for the “accident” sub-topic. This implies the importance of further improving the urban road safety (which might be addressed in the emerging autonomous and connected vehicle technology).

As shown in Figure 7b, some traffic-related checkins are located in the controlled region (the pink and green colored regions), therefore we can combine physical sensing traffic data with these participatory points for a deeper analysis. For each checkin, we know its time $t$ and location. Hence we can further investigate the traffic flow in the time window $[t-4h, t+4h]$ and at its closest intersection in the controlled region.

Figure 8 shows the vehicle flow patterns at Intersections A and Q that are associated with two checkin examples using the “traffic” topic (T1 and T2 in Figure 7b which are respectively pertinent to a road closure and an accident event), where we set $t = 0$. In Figures 8a and 8c, both the traffic flow in the invested time window $[t-4h, t+4h]$ (denoted as “Real-Time Data”) and a 10-week average traffic flow (denoted as “10-Week Average”) are shown, where the 10-week average traffic flow is calculated for capturing the weekly periodic patterns of human mobility. In Figures 8b and 8d, the detrended flows are obtained by subtracting the “10-Week Average” flow from the “Real-Time” flow in Figures 8a and 8c, respectively. In each detrended flow, a short-term flow disruption can be clearly identified by the deep deviation for sufficient long time. This provides us a plain information on how the non-recurrent incident impacts local traffic. The two events impacted the traffic for about 4 hours and 1 hour respectively (see Figures 8b and 8d). Interestingly, checkins pinpointed the events during the early stage for both the events. It indicates the potential of using checkins for incident detection, by narrowing down the searches to specific times and locations.

A Tale of Two Zones
Our controlled region spans some adjacent livehoods (19) with different life activity patterns, including East Liberty and Bakery Square.

We choose two zones, $Z_1$ in East Liberty and $Z_2$ in Bakery Square as shown in Figure 1b, for closer observation. For the two zones, accurate departure vehicle flows can be detected at the side-street exits of the intersections F and H, and of the intersections O and M, respectively.

Figures 9a and 9b show the seasonal average vehicle flow patterns for zones $Z_1$ and $Z_2$ respectively. In the zone $Z_1$, the traffic flow in weekends is significantly higher than that in weekdays; While in the zone $Z_2$, the traffic flow in weekends is significantly lower than that in weekdays. The zones $Z_1$ and $Z_2$ respectively contain a major store (Target) and a major company (Google), the two figures therefore are consistent with the the common human behaviors of shopping in weekends but working in weekdays.

In the participatory sensing data, there were 2390 checkins from 1054 users in Zone $Z_1$, and 5349 checkins from 1260 users in Zone $Z_2$. There were only 253 users appeared in both zones.

We further find the transitions of users for the zones. For each zone, each user transition
is defined as two consecutive checkins, one is within the zone and another is outside of the zone (or vice versa), made by a user in a time threshold (four hours in this paper). If the later checkin is in the zone, the transition is an in-zone transition, otherwise it is an out-zone transition. For each zone, the in-zone and out-zone transitions are more related to arrival and departure traffic. The transitions also provide origin and destination (O-D) information for each zone.

Figures 9c and 9d give respectively the checkin and transition patterns for the two zones. In Z1, most checkins can be classified into transitions, which might due to the fact that users do not stay too long after shopping. In Z2, the number of transitions is significantly lower than that of checkins, since users might stay long and send checkins in their workplaces. For weekdays and weekends, the temporal checkin patterns are roughly similar to that of traffic flow patterns (see the comparisons of Figures between 9a and 9c, and between 9b and 9d). Both the numbers of checkins and transitions shown in Figures 9c and 9d are now not large enough to support a statistical analysis of correlation with the traffic flow data shown in Figures 9a and 9b. But the situation should be improved in the near future based on the rapid increasing trend in participatory sensing data as shown in Figure 2a.

Figure 10 gives the user transition information for zones Z1 and Z2. The origin/destination (O-D) checkins of transitions have been distributed broadly in the participatory sensing region. As shown in 10c and 10d, the O-D locations are highly clustered, and the majority of them are covered by a few clusters of sources. For the two zones, they share most of the O-D sources (clusters in blue), even though they have an apparent difference in both the checkin and flow patterns. It implies that urban traffic might be mostly generated among a few clustered sources.
FIGURE 9: Vehicle Flow and Checkin Patterns for Two Zones Z1 and Z2.
This O-D information might be used to further understand the traffic demands, based on some modeling frameworks, e.g., the gravity model (31) or the radiant model (2). These clustered O-D locations can be used to construct popular routes (32), and the information beyond the traffic control region might then be useful for recommending time-sensitive alternative routes (33) to help reducing traffic congestion within urban traffic control systems (34), especially if traffic anomaly (35) has been quickly identified or predicted. It is also possible to provide carpooling recommendation for users based on the similarity in their O-D transition patterns.

CONCLUSIONS
In this empirical study, we explored the potentials of combining participatory and physical sensing data for better understanding and improving urban mobility networks. The participatory sensing data were collected in the Pittsburgh metropolitan area with the APIs provided by the current location-based services, and the physical traffic flow data were collected in an area controlled by
the smart and scalable urban traffic control system. For both the sensing data, spatial and temporal characteristics were displayed.

We first investigated human mobility patterns for a better understanding on urban mobility network. By examining user checkins, we disclosed the distribution of user behaviors, a fundamental statistical properties of mobility pattern. With geo-location based cluster analysis, we identified personal favorite places of users in the studied regions. By measuring user entropy, we revealed the degree of predictability of user activities. By analyzing time-dependent mobility patterns, we showed the regularity of user behaviors.

Next we informed how we might take advantage of the combined participatory and physical sensing data to improve urban mobility applications. We presented three examples to illustrate the value of the combined data for urban mobility networks: (1) anomaly traffic detection and reasoning, (2) topic-based nontrivial traffic information extraction, and (3) traffic demand analysis. The study might shed some lights on further research for enhancing urban mobility.

Acknowledgements. This research was supported in part by the T-SET University Transportation Center at CMU, the CMU Robotics Institute, and the CMU Physics Department.

REFERENCES


