Studying Traffic Conditions by Analyzing Foursquare and Instagram Data

Anna Izabel J. Tostes†, Thiago H. Silva†, Fátima Duarte-Figueiredo†, Antonio A. F. Loureiro†
†Computer Science, Universidade Federal de Minas Gerais (UFMG), Brazil
Computer Science, Pontifical Catholic University of Minas Gerais (PUC-MG), Brazil
annatostes@gmail.com, {thiagohs,loureiro}@dcc.ufmg.br, fatimafig@pucminas.br

ABSTRACT
Traffic jam is a contemporary society problem in urban areas. There are specific sources of information about traffic conditions in the Web, such as Bing Maps. This system presents real-time information about the traffic conditions (e.g., free or congested). Recently, participatory sensing systems, such as Foursquare and Instagram, are becoming very popular. Data shared in these systems have the active participation of users using their portable devices ubiquitously. In this case, these systems can be seen as a kind of sensor network, where users can be considered a social sensor, because the data shared by them are associated with their habits and routines. Thus, can we use data from social sensors, specifically from Foursquare and Instagram, to better understand traffic conditions? This paper shows that data from social sensors and traffic conditions, provided by Bing Maps, are surprisingly very correlated. The social data distribution is equal to the traffic condition distribution, shifted by an offset that can be easily calculated. This information can be extremely valuable, for example, to build more efficient traffic condition predictors.

Categories and Subject Descriptors
J.4 [Computer Applications]: Social and Behavioral Sciences

Keywords
Vehicular Network, Traffic Jam, Participatory Sensing, Social Networks

1. INTRODUCTION
Despite the technological advances in vehicles aimed to improve the experience of drivers and passengers, traffic problems, specially traffic jams, lead to prohibitive economic losses, decrease the overall productivity and negatively impact the environment [1]. Daily users of vehicular networks travel from home to work, hoping to reach their destination within an estimated time, when in most cases they are faced with traffic jams. The individual decisions about which routes to follow with no updated information about traffic leads to an unbalanced use of the roads that can worsen the traffic conditions. However, a more balanced vehicle distribution is of collective interest, since it tends to reduce average trip times, the fuel consumption and the emission of carbon dioxide [5].

Intelligent Transportation Systems (ITS) and the opportunity to have wireless connectivity in vehicular environments can assist drivers in making faster and safer trips, reducing the negative effects of traffic problems on the roads [3]. ITS use infrastructure sensors to monitor the traffic condition in a vehicular environment and provide applications, such as collision detection systems, with traffic-related information and ubiquitous connectivity to the Internet [4, 10, 13].

There are many kinds of sensors used by ITS such as road detectors (inductive loop detection), in-road reflectors, infrastructure-to-vehicle and vehicle-to-infrastructure electronic beacons. These are some examples of “traditional sensors”. However, today we have many other emerging source of data that can be very useful for ITS. For instance, there are information about traffic conditions in the Web, such as Bing Maps. This system presents real-time information about the traffic conditions (e.g., free or congested), which can be acquired using APIs or other methodologies [20]. Besides that, the use of participatory sensing systems, such as Foursquare1, Instagram2, and Waze3, are becoming very popular. For example, in 2014 Foursquare registered 45 million users, Instagram 200 million users, and Waze 50 million users. Data shared in these systems have the active participation of users. In this case, these systems can be seen as a sort of sensor network, also known as participatory sensor network (PSN) [17]. In this network users plus his/her mobile devices can be considered a sensor, because users carry mobile devices that are able to sense the environment and make relevant observations at a personal level.

Data from sensors in a PSN are associated with users’ habits and routines [17]. Thus, these sensors can be considered social sensors, which provide valuable to better understand city dynamics and the urban behavioral patterns of their inhabitants. Since this is a new type of data many questions emerge. For instance, can we use data from social sensors, specifically from Foursquare and Instagram, to better understand traffic conditions? In fact, answer this question is the main goal of this paper.

To tackle this question we developed a methodology to correlate data from social sensors and traffic conditions available on the Web, using, as an example, data from Bing Maps. We found that data from social sensors and traffic conditions, provided by Bing Maps, are surprisingly very correlated. The social data distribution is equal to the traffic condition distribution, shifted by an offset that can be easily calculated. That is, the time difference between the

measurements is \( \delta \) minutes, which indicates that social data signalize the traffic flow \( \delta \) minutes latter. For Manhattan, \( \delta \) is equal to 36 minutes. This information can be extremely valuable in several scenarios. For instance to build more efficient traffic condition predictors or to build new VANET protocols to control and to minimize the traffic problems, such as jams.

The rest of this work is organized as follows. Section 2 presents the related works. Section 3 discusses the usefulness of social data. Section 4 describes our new methodology to obtain and process traffic condition data. Section 5 shows the results. Finally, Section 6 presents the conclusion and future work.

2. RELATED WORK

We divide our related work in two classes. The first one is related to the study of traffic conditions. In our previous work [20], we used traffic flow information from Bing Maps in order to predict the traffic conditions. We have only used information about the traffic flow intensity (green – free, red – congested) from the previous week to forecast the future traffic condition. An accuracy of 64% with a precision of 60% was achieved.

In [16], the authors developed a device for predicting traffic jams on roads. The prediction method correlates traffic jam with the traffic jam pattern and predicts the current traffic jam degree based on the up-to-the-minute traffic jam information and the current traffic state. In [15], a Bayesian network has been projected to predict the mobility of vehicles in a grid map. The model is based on location information and the direction that vehicles take when they arrive at an intersection.

The second class of related work considers participatory sensing, obtained, for example, from Twitter, in order to better understand city dynamics. The authors in [9] used a dataset from Twitter and proposed a technique to determine the type of activities that is most common in a city by studying tweeting patterns. [2] studied whether collective mood states derived from Twitter feeds are correlated to the value of the Down Jones Industrial Average over time. [14] studied the real-time interaction of events in Twitter (e.g. earthquakes), and propose an algorithm to monitor tweets to detect a target event. [12] present a geo-social event detection system to identity local events (e.g., local festivals) by monitoring crowd behaviors indirectly via Twitter.

In the same direction [19] considers the traffic alert system Waze, another example or participatory sensing source, as a sensor network in order to verify its properties. Among the results, the authors show that data sharing is correlated with users’ routines, and that Waze data can improve traffic condition understanding. Correlated with [19], [7], the authors used a dataset from Twitter and proposed an information extraction technique to get the data of traffic. The traffic data is presented in map view as a mobile application of Android.

Although much has been done, no previous effort studied the correlation of information from participatory sensing systems and real traffic conditions data. This is the point that differentiates our work from the previous ones. We here investigate whether we use can use participatory sensing data (or social data) to better understand traffic conditions.

3. SOCIAL SENSING

Participatory Sensing Systems (PSSs) provides a mobile interface that allows people to share data about the environment (context information) they are in any time and place using any mobile device. Examples of PSSs deployed and functioning at global scale are location-sharing services, such as Foursquare, and photo sharing services, such as Instagram. They can provide valuable information about an aspect of a given city or society in almost real-time, such as its traffic and weather conditions, local parties and festivals, riots, among others integrate user interactions [19].

Data shared in participatory sensing systems have the active participation of users. These systems can be seen as a sort of sensor network, also known as participatory sensor network (PSN) [17]. In this network users plus his/her mobile devices can be considered a sensor, because users carry mobile devices that are able to sense the environment and make relevant observations at a personal level.

PSNs are an example of the interplay between technological networks and social networks, since a key element in a participatory sensor network is the human being. As shown in [17] data shared by the sensors in a PSN are associated with users’ habits and routines. Thus, these sensors can be considered social sensors, which provide valuable to better understand city dynamics and the urban behavioral patterns of their inhabitants. Many questions emerge when using this new type of data, for example: can we use them to better understand traffic conditions? In fact, this is a very interesting question and it is the central question that this paper is answering.

4. METHODOLOGY

In this work, we analyze two social sensing sources, one derived from Foursquare, and another derived from Instagram. Recall that in our work every data shared in Foursquare or Instagram is called check-in. We want to investigate if we can use data from social sensors (check-ins) to better understand the traffic condition, that is, the traffic flow from Bing Maps. The objective is to verify if check-ins can be used as a hint of traffic conditions changes or current situation.

4.1 Overview

As people move around in cities, they share their location, pictures and videos throughout their routine through the check-in. Between one check-in and the next one users have to move in a certain way, for example by bus or by car. Thus, since check-ins are correlated with inhabitants routines [17], they might implicitly have embedded information about the traffic conditions. Consider the following scenario: a person makes a check-in at home, gets his/her car and drives to work, gets stuck in a traffic jam, and arrive at work and makes another check-in. Imagine how many people on Foursquare follow the same routine. Therefore, the information of check-in can reflect the behavior of the traffic. Do people make more check-ins when traffic is congested? Do check-ins can be a sign of heavy traffic? In other words, do the information of check-ins are correlated with heavy traffic? In summary, can we use the check-ins as a sensor to indicate and to predict traffic jams? Answer those questions is a fundamental step in order to use check-ins as a complementary source of information to improve intelligent traffic systems.

To investigate our hypothesis that check-ins and traffic jams are correlated, we analyze two datasets from the same period of time at Manhattan, New York City: (1) check-ins from Foursquare and Instagram; (2) traffic flow from Bing Maps.

We chose New York City, Manhattan region, because it a popular region, having a high number of people performing check-ins, as previous verified in [18]. Besides that, according to the NYMTC report [6], it has the third lowest daily vehicle miles traveled\(^4\) (VMT) per capita due to high population density and high

\(^4\)Vehicle Miles Traveled (VMT) is the sum of distances traveled by all motor vehicles in a specified region [6].
proportion of transit use. Manhattan region has 10,702,575 VMT daily. Figure 1 shows that New York and Los Angeles have the highest travel volumes in comparison with other cities and a very large area, which is a metropolitan statistical area with over three million residents.

Figure 1: Vehicular travel volumes in New York and comparable metro areas [6].

In order to evaluate the correlation between traffic flow and check-ins, we have developed the following methodology:

1. As the time interval between check-ins is high, we aggregate the check-ins into a 3-hour period. Thus, for the 24 hours of a day, we have 8 periods of 3 hours. However, if we aggregate the traffic flow in the same time interval we can miss valuable information, as the transit varies widely. Traffic flow in Bing Maps varies from fast traffic (represented by green color or the integer 1), moderate traffic (represented by yellow color or the integer 2), and slow traffic (represented by red color or the integer 3). Knowing that, we categorized traffic flow (provided by Bing Maps) for a particular street as it follows:
   - Bing = 1, when the average traffic flow of that specific street segment is less than 1 (recall that we also represent traffic conditions by an integer);
   - Bing = 2, when the average traffic flow of that specific street segment is between 1 and 1.5;
   - Bing = 3, when the average traffic flow of that specific street segment if higher than 1.5.

2. For each street segment\(^3\), we calculate the mean value and standard deviation of check-ins. These information will be used to map the check-ins category into 1, 2, or 3 (the same as traffic flow categories).

3. For each time interval (1–8), we perform the following steps:
   - Capture the traffic flow category (1, 2 or 3);
   - Calculate the total number of check-ins in which its location is in the street segment;
   - Based on the total number of check-ins every street segment, the mean and standard deviation calculated on step 2, apply the following steps to categorize check-ins:
     - If the number of check-ins in the street segment is between the confidence interval (based on the mean and one third of the standard deviation), then the check-in category is 2;
     - If the number of check-ins in the street segment is higher than the mean with one third of the standard deviation, then the check-in category is 3;
     - Otherwise, the category is 1.

4. To analyze the correlation between the traffic flow categories and the check-ins category, five groups have been created:
   - **Group 1**: Check-ins category is less than the traffic flow category, which means that when the number of check-ins is low, the traffic flow is more congested. In this case, the traffic flow is 2 or 3 (yellow or red) and the number of check-ins is low (respectively 1 or 2).
   - **Group 2**: Check-ins category is higher than the traffic flow category, which means that when the number of check-ins is high the traffic flow is free. When the traffic flow category is 1 or 2 (green or yellow), the number of check-ins is high (respectively 2 or 3).
   - **Group 3**: Check-ins category is less than or equal to the traffic flow category, in which we aggregate data from Group 1 and Group 3.
   - **Group 4**: Check-ins category is less than or equal to the traffic flow category, in which we aggregate data from Group 2 and Group 3.
   - **Group 5**: Check-ins category is equal to the traffic flow category, which means that the check-ins category is equal to the traffic flow category (the correlation is 1).

5. Finally, to analyze if check-ins are a good signal to indicate traffic flow condition (free or congested), we aggregated the data according to a time interval of \(\delta\) minutes. We separate the analysis in weekdays days (Monday to Friday) and weekends (Saturday and Sunday). For each dataset, two distributions were plotted, one with the frequency of check-ins during 24 hours, and another with the frequency of congested traffic flow (average traffic flow category for all streets). With these distributions, we want to demonstrate that they have similar shapes, i.e., even if a certain threshold shifts them, acting then as a sort of signal about traffic condition changes. This could help to improve the prediction of, for example, problems such as jams.

Our hypothesis is that these two functions are shifted \(\delta\) minutes. For that we propose the Equation 1, in which \(y_t\) is the value of traffic flow in time \(t\) and \(y^*_k\) the value of check-ins in time \(k\). To discover the value of \(\delta\), we calculated the discrepancy value through the sum of the difference between the \(y\) value (traffic flow) in time \(t\) and the other \(y^*_k\) value (check-ins) in time \(t - \delta\), as shows Equation 1:

\[
D(\delta) = \sum_{t} \left( |y_t - y^*_k(t-\delta)| \right)
\]  

(1)

It is possible to calculate the above function for \(\delta = 1, 2, 3, \cdots\) in order to find the \(\delta\) value that minimizes the discrepancy. That equation represents the error between the two input distributions, which is also a random variable.

### 4.2 Foursquare and Instagram Dataset

In this study we focus on users’ check-ins given on Foursquare and Instagram. Foursquare is a very popular location sharing service, class of system that enables users to share their location with friends. To give an idea of the popularity of Foursquare, this year it gathered over 45 million users worldwide, which shares millions of locations every day [8].

In location sharing services the main object are venues, which represent any physical location, such as a restaurant, an university, or a gas station, that a user can be at. Thus, the basic activity users can perform in location sharing services is called check-in, which is an action to announce in the system the venue you are at a certain moment. Foursquare also maintains a set of eight pre-defined venue

\(^3\)Street segment is the part of the street between consecutive intersections, considering a specific direction.
categories, namely, “Arts & Entertainment”, “Colleges & Universities”, “Food”, “Great Outdoors”, “Nightlife Spots”, “Travel Spots”, “Shops”, “Home, Work and Others”. Other actions could also be allowed. For instance, in Foursquare users can post tips in specific places aiming at sharing information on any aspect related to the venue.

Instagram, created in 2010, is a photo sharing service that allows users to take pictures, and share them on a several social networking services, such as Twitter. Currently, Instagram users can create Web profiles featuring recently shared pictures, biographical information, and other personal details. Instagram is a very popular photo-sharing service. In February 2013 Instagram announced that they had 100 million users, and in 2014 this number reached 200 million users [11]. In photo sharing services the main object are photos. Photos could be taken anywhere, e.g., at a restaurant or at home. Thus, the basic action is the sharing of a photo. In Instagram the user can also associate a location with each photo. The action of sharing a photo with a location is also called check-in, because, as in location sharing services, the user is announcing where he/she is at a certain moment. We have collected data from Foursquare and Instagram directly from Twitter, since Foursquare and Instagram check-ins are not publicly available, by default. Only geo-tagged tweets shared in New York City were considered. The interval of collection was the same for both of them: between June 24 and August 22 of 2013. In both datasets each check-in consists of the latitude, longitude, venue’s id, and time. We have extracted from Twitter 65,293 geo-tagged tweets containing check-ins shared on Foursquare. Data were shared by 14,575 unique users on 18,896 unique venues. The Instagram dataset is composed of 12,7185 geo-tagged tweets containing check-ins shared by 41,205 users in 30,738 unique venues.

To give an idea of the coverage of our dataset, Figure 2 shows a a heatmap of all check-ins in the area of New York City considered: Manhattan. In the heatmap, the darker the color, the higher is the number of check-ins in that area.

Figure 2: Check-ins heatmap over Manhattan region.

4.3 Bing Traffic Flow Dataset

A traffic flow dataset of Manhattan was developed using the following methodology, proposed in previous works [20].

Aiming a citywide traffic flow discovery to establish inferences and big data analysis for patterns discovery, a methodology for acquiring traffic flow data from distinct sources was developed. Any GIS map service can be input of this methodology. Figure 3 presents the process of traffic flow acquisition. Through the API map service, a city flow web crawler has been developed. Then, a bash script was designed in order to collect the traffic flow image from the selected city. We used virtualization to print the screen and save it into the database. A image processing software was developed for extracting each road traffic intensity, saving the percentage of green pixels, yellow pixels, and red pixels, which correspond to the flow intensity (green is free while red is congested). Each image from the image database was processed and its flow intensity was saved to a specific date and time into the Traffic Flow Acquisition.

![Figure 3: Traffic flow acquisition methodology through a GIS map web crawler.](image)

In this work, Bing Map has been used as input. Algorithm 1 presents the procedure for loading Bing Map in a webpage, with the traffic layer on, and without the illustrations over the map. First the map is set to the specific city center location (geo-code of New York is 40.783094, -73.980324), with the specific zoom level (12). Then, the traffic layer is enabled and its opacity is turned off (1). We display only the map, disregarding everything else. To know when the maps and the traffic layer are loaded on the browser, we start two event handlers. When the map and the traffic layer are loaded, two events `tiledownloadcomplete` are triggered.

**Algorithm 1: Bing Map Traffic Module Load**

```
setMapView();
showTrafficLayer();
opacityTrafficLayer();
eventHandlers();
```

The web crawler algorithm is presented by Algorithm 2. For each HTML scenario established as Algorithm 1, we get current hour, open the web browser with the created HTML file, and wait for the event signals indicating that all map layers were loaded. Then, we get a screenshot. This process is repeated every second. This approach was implemented with PhantomJS, an open-source headless browsers.  

**Algorithm 2: Traffic Flow Web Crawler**

```
DELAY = 10 seconds
for each Scenario s do
    get current hour;
    open web browser with the code from Algorithm 1;
    print the screen;
    kill web browser;
    sleep DELAY;
end
```

After collecting image data, the next step is the processing. As input, the algorithm needs the image data and the masks for each road. The mask is a binary image with white background and black street line. Figure 4 illustrates a street mask for one street segment and a translucent image of Chicago map overlapped, which were used in this work. For such scenario, 100 street masks were manually drawn.

Algorithm 3 presents the steps followed to process one map image. For each black pixel in the street mask, the counter for each flow category (green, yellow, red, or no category – error) was increased according to its color. To establish a band for each flow intensity, HSL (Hue, Saturation, and Lightness) was used. Similar to HSV (Hue, Saturation, Value), HSL is one of the most common cylindrical-coordinate representations of points in an RGB color model. The variation of the hue corresponds to the values 0–360, [21][21][21].
Algorithm 3: Traffic Flow Image Processing

Input: Image file \( i \), Set of Road Masks \( k_r \)

GreenPixels = 0;
YellowPixels = 0;
RedPixels = 0;
NoCategoryPixels = 0;

foreach Road Mask \( k_r \) do

foreach Pixel \( p \) in the \( k_r \) image do

if \( p \) is black then

\( \theta \) Increase the counter of its respective color if \( \text{hue}(p) < 30 \) or \( \text{hue}(p) \geq 330 \)
RedPixels++;

end

else if \( \text{hue}(p) < 70 \) then
YellowPixels++;
end

else if \( \text{hue}(p) < 150 \) then
GreenPixels++;
end

else
NoCategoryPixels++;
end

each Pixel \( p \) in the \( k_r \) image end
foreach Road Mask \( k_r \) do

end

end

in which \( 0 \) is a red band, followed by a yellow band, and other band colors, and finally a red band again. So it is possible to identify color bands to Bing’s traffic flow intensity.

In our dataset, data was collected from June 24 of 2013 to August 22 of 2013. Bing map acquisition occurred every 1 minute. Next, the database has the following information: (I) date; (II) hour; (III) street number; (IV) number of green pixels; (V) number of yellow pixels; (VI) number of red pixels; and (VII) number of no category pixels. The dataset has 12,394,175 data.

Figures 5a and 5b presents a temporal analysis of the traffic flow for Manhattan region. We present the frequency of green, yellow and red traffic flow during working days (Monday to Friday) and during weekend (Saturday and Sunday). One can notice that the traffic congestion has two peaks of congested traffic flow, at \( x = 85 \) (approximately 7am) and at \( x = 192 \) (4pm), corresponding to rush hours (when people drive to work and to home). During weekends, one can notice that we have only one peak when \( x = 175 \) (approximately 3pm), but the free traffic is more usual that does not occurs on working days in which the congested traffic is more usual (yellow and red flows).

Figure 6: Most likely traffic flow in time intervals of 5 minutes.

To make a spatial analysis, we generated videos presenting the most likely traffic flow for Manhattan region considering working days and weekends. These videos are available online\(^7\). Intervals of 5 minutes have been used in order to consider most of the traffic flow transitions on streets. Figure 6a and 6b show two snapshots during the instant 7am and 4pm, respectively. We can see which street segments are more congested and which streets have free flow. Comparing with Figures 6c and 6d, we can see that at 7am the yellow traffic flow is more likely than at 4pm during working days. In weekends, the free flow is more likely in such hours than during working days.

5. RESULTS

This section presents the results about our investigation to discover whether check-ins and traffic flow are correlated, i.e., if check-ins are good signal of traffic conditions changes and matters. We made four analyses: (i) general analysis; (ii) spatial analysis; (iii) temporal analysis; and (iv) offset analysis. The general analysis presents the results of the 1–4 steps from our methodology (presented in Section 4). After confirming the correlation between check-ins and traffic flow, even if at certain times and categories, the next two analysis presents where (spatial analysis) and when (temporal analysis) the correlations occur, this is important in other

\(^7\)http://annatostes.azurewebsites.net/bing-maps-
to better understand the results. Finally, we present the offset analysis in order to verify how much distant are check-ins distribution from congested traffic flow distribution.

5.1 General Correlation

This section presents the results of general correlation. Steps 1–4 from our methodology have been applied with intervals of 3 hours due to the high inter-sharing times of check-ins. A positive correlation, between check-ins and traffic flow, means that the more check-ins the worse traffic condition.

Table 1: Correlation between check-ins and traffic flow.

<table>
<thead>
<tr>
<th>Correlation / Hour</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0am – 3am</td>
<td>0.364966</td>
<td>0.406984</td>
<td>0.24973</td>
<td>-0.252724</td>
</tr>
<tr>
<td>2</td>
<td>3am – 6am</td>
<td>0.154937</td>
<td>0.346152</td>
<td>0.471932</td>
<td>0.0697677</td>
</tr>
<tr>
<td>3</td>
<td>6am – 9am</td>
<td>0.141381</td>
<td>0.339984</td>
<td>0.468896</td>
<td>0.182969</td>
</tr>
<tr>
<td>4</td>
<td>9am – 12pm</td>
<td>0.46509</td>
<td>0.549677</td>
<td>0.58304</td>
<td>0.0542799</td>
</tr>
<tr>
<td>5</td>
<td>12pm – 3pm</td>
<td>0.38665</td>
<td>0.39853</td>
<td>0.34258</td>
<td>-0.021738</td>
</tr>
<tr>
<td>6</td>
<td>3pm – 6pm</td>
<td>0.53004</td>
<td>0.53456</td>
<td>0.427304</td>
<td>0.323002</td>
</tr>
<tr>
<td>7</td>
<td>6pm – 9pm</td>
<td>0.430769</td>
<td>0.635024</td>
<td>0.364643</td>
<td>0.092255</td>
</tr>
<tr>
<td>8</td>
<td>9pm – 12am</td>
<td>0.256889</td>
<td>0.672354</td>
<td>0.2050497</td>
<td>-0.027646</td>
</tr>
</tbody>
</table>

Table 1 presents the results of correlation for the five groups through different periods of the day. Each column represents a group (1–4) from left to right. As explained before, Group 5 was omitted because the correlation is always one, when we have data. We can see that the highest correlation is in Group 3 (“Check-in <= Traffic Flow” $\rightarrow$ {0.44, 0.33, 0.34, 0.55, 0.43, 0.42, 0.64, 0.67}), where the category of check-ins is lower than or equal to the traffic flow. In times when the traffic flow is more intense, we can see a higher correlation (after 6pm – 0.64 and 0.67).

Another interesting result is the comparison between the Groups 3 and 4 during 3am–6am and 6pm–9pm. We can see that the correlations are inverses. It suggests that in the morning, people makes more check-ins before going to work when the traffic is free, and then they move in the city and the traffic flow gets more intense (potentially congested). This can be seen by the greater correlation of Group 4 (0.47) compared to the correlation of Group 3 (0.33). At night, the opposite occurs, people tend to make more check-ins after getting stuck in traffic jams. This is an expected behavior because people do not have many incentives to perform check-ins while driving in intense traffic conditions. As we can see, the correlation of Group 3 (0.64) is higher than the correlation of Group 4 (0.36), enforcing our conjecture.

As we can see, those correlations are still low, suggesting that some regions might be more correlated than others. Even though, just by those results we are not able to prove that check-ins can be used as good signal about the traffic conditions. In other words, the usefulness of check-ins as a hint to predict traffic changes is still not clear up to this point. For this reason, we have made a spatial and a temporal analysis. For this, we applied Step 5 of our methodology.

5.2 Spatial Analysis

Here we present the results of the spatial analysis. We show the traffic flow and check-ins categories during the eight periods of time considered in this study (lines of Table 1).

Figures 7, 8 and 9 show the check-ins, the traffic flow and the check-ins categories in different periods of time, respectively, when we have different traffic flows. For a better visualization, videos featuring each result during all periods of time considered have been made. They can be found online.

![Figure 7: Results of check-ins in different periods of time.](http://annatostes.azurewebsites.net/check-ins-and-traffic-flow/)

![Figure 8: Results of check-ins category in periods of time.](http://annatostes.azurewebsites.net/check-ins-and-traffic-flow/)
The categories of check-ins in the same periods of time presented in Figure 7 can be seen in Figure 8. We can see that when we have less check-ins, we have more green streets. When the number of check-ins increases, the intensities of the streets (color) change from green to yellow and then to red. Regarding to the average check-ins in the street, the illustration enable the identification of where (street) there are more check-ins, beyond the usual, and where there are not. We can see that the red lines are more concentrated in downtown (see Figure 8b), but the period of 6pm–9pm is when people tend to make more check-ins, above the average, in the entire region (not just in downtown).

On the other hand, Figure 9 presents the traffic flow from the same periods. One can notice that the yellow traffic is more common at morning (see Figure 9c) and less common in other periods. This behavior is expected because it is the time when people go to work.

When we compare all graphics, we can see that when the traffic flow is higher (red lines), check-ins are lower (green lines). In the first look, it seems like the check-ins and the traffic flow are inversely correlated. But this is not what happens. The explanation is that the traffic flow behavior will only be seen in the check-ins category after $\delta$ minutes.

Despite that, a better analysis can be performed, because a period of 3 hours is too large to represent the dynamics of traffic. So we need to investigate shorter intervals of time, such as 5 minutes, 10 minutes, and so on. This analysis is presented in the next section.

5.3 Temporal Analysis

Here we present a temporal analysis of check-ins and traffic flow, describing how the check-ins and the congested traffic flow vary during the day.

Figure 10 presents three graphics where we can compare the real data from a specific day (07/25/2013 – see Figure 10a), the typical distribution for weekdays (see Figure 10b), and the typical distribution for weekends (see Figure 10c).

Comparing the two first graphics, we can see that the traffic flows and check-ins distributions seems to be almost the same, but shifted. And the typical distribution represents the specific day.

During weekends the behavior changes, presenting only one peak in the late afternoon for both check-ins and congested traffic flow. This finding is very surprising and shows that the social sensing might reflect real traffic conditions.

A natural question that emerges is the impact in the results when changing the time interval of our analysis. To evaluate the impact of choosing another time interval, we calculated the distribution for intervals of 5, 10 and 30 minutes, 1 hour and 3 hours, depicted in
5.4 Offset Analysis

In order to formally discover the value of the shift $\delta$ between the two distributions (check-ins and traffic flow), we calculated the distribution given by the Equation 1. It represents the discrepancy distribution, illustrated in Figure 12. The red line indicates the value of $\delta$ (time) that minimizes the discrepancy error. As we can see, for Manhattan, the $\delta$ should be equal to 36 minutes.

Then, we shifted the check-ins distribution in 36 minutes and Figure 13 presents both distributions but shifted. With this result, we can see that the check-ins distribution is equal do the congested traffic flow distribution. This indicates that check-ins can be used to forecast intense traffic flow.

6. CONCLUSIONS

This work has investigated whether Foursquare and Instagram check-ins can be used to signalize congested traffic flow. Through a methodology of five steps, we have showed that these information are correlated. Based on the temporal and the spatial analyses associated with the distribution Equation 1, we have shown that the distribution of check-ins is equal to the distribution of congested traffic flow with a discrepancy error that can be easily calculated, as demonstrated. In our case of study, we identified that check-ins signalize the traffic flow 36 minutes after. We believe that the 36 minutes threshold occurs because there is a time between the congestion and the time that the person arrives at the check-in point. The threshold can vary from place to place, but as shown in the paper, with our methodology this value can be easily calculated. As a future work we will investigate the thresholds for other cities around the world using our methodology.

Acknowledgments

We would like to thank CNPq, CAPES, FAPEMIG and Microsoft Research for the financial support.

7. REFERENCES


