Mining Complaints for Traffic-Jam Estimation: A Social Sensor Application

Theodore Georgiou[1], **Amr El Abbadi**[1], **Xifeng Yan**[1], and **Jemin George**[2]

[1]Department of Computer Science, University of California, Santa Barbara
[2]U.S. Army Research Laboratory
{teogeorgiou, amr, xyan}@cs.ucsb.edu, jemin.george.civ@mail.mil

***Abstract*— Physical events in the real world are known to trigger reactions and then discussions in online social media. Mining these reactions through online social sensors offers a fast and low cost way to understand what is happening in the physical world. In some cases, however, further study of the affected population's emotional state can improve this understanding. In our study we analyzed how car commuters react on Twitter while stuck in heavy traffic. We discovered that the online social footprint does not necessarily follow a strict linear correlation with the volume of a traffic jam. Through our analysis we offer a potential explanation: people's mood could be an additional factor, apart from traffic severity itself, that leads in fluctuations of the observed reaction in social media. This finding can be important for social sensing applications where external factors, like sentiment, also contribute on how humans react. Ignoring the existence of such factors can lead in reduced quality and accuracy of a regression analysis.**

We propose a novel traffic-congestion estimation model that utilizes the volume of messages and complaints in online social media, based on when they happen. We show through experimental evaluation that the proposed model can estimate, with higher accuracy, traffic jam severity and compare the results with several baselines. The model achieves at least 38% improvement of absolute error and more than 45% improvement of relative error, when compared with a baseline that assumes linear correlation between traffic and social volume. To support our findings we combined data from the California Department of Transportation (CALTRANS) and Twitter, for a total of 6 months, and focused on a major traffic-heavy freeway in Los Angeles, California.

I. Introduction

Since the establishment of online social media, real life events frequently trigger a social reaction on the web. This has led to an era where Big Data and social media content are strongly tied together [3]. Utilizing this vast, but publicly available, amount of information to mine the correlation between physical events and postings on Twitter or Facebook has proven to unveil hidden behavioral patterns or validate social and psychological theories that once required extensive and expensive surveys [21]. Additionally, the discovery of what is happening in the real world is now feasible through purely automated and algorithmic tools that only require access to the Internet. However, due to the noisy nature of the data, its size, and in many cases our own lack of better understanding, the quality of any data mining or machine learning product will just approximate the actual reality.

In the current study we focus in particular on the fact that the sentimental state or mood of the analyzed population (in the context of social sensors and event discovery) is seldom attributed. Most algorithms measure the levels of a disaster or the magnitude of an event as a simple function of the corresponding social media discussion volume. This simple function can be anything from a linear model to an exponential distribution. But what gets usually ignored is the state of the people that participate in the online discussion. For example, an overly enthusiastic crowd might give a false idea of the size of a political demonstration. A shy demographic might lead to the perception that a specific music trend is not as popular as it really is. People complaining about their jobs during a very hot day might give the false sense they are generally unhappy with their work environment. To avoid arriving to such false conclusions based on online social signals, a better understanding is needed of when people publish on social media, what emotional state they are in, and which factors might have led them there.

For our experiments, a specific user behavioral pattern was examined: complaining in social media while stuck in traffic jams. We combined two publicly available datasets, one for traffic in California and one for Twitter content, to study how car drivers react in social media while driving during increased traffic congestion. Driving a car is already known to be a stressful activity for many and things can be much worse during traffic jams; frustration and boredom may lead drivers to make irrational decisions or behave relatively abnormally due to anger. Unfortunately, both behaviors can increase traffic, be dangerous, and cause accidents. Social Media have already been utilized for some time now to help with traffic decongestion. From specialized social media apps like Waze [26] - a crowd-sourced community that monitors traffic, accidents and other events in real time - to regular use of Twitter to automatically or manually publish reports and alerts of the street conditions [8]. The purpose of such information tools is for drivers to inform themselves about traffic conditions before getting in their car and make the necessary choices to optimize their commuting route and time. In reality, a non trivial amount of smartphone owners are observed to use their handheld devices while driving, despite laws that render the use of handheld devices by drivers for texting purposes illegal, for obvious safety reasons [7].

Apart from getting informed about traffic, users resort to social media to also complain or update their Twitter/Facebook status about being stuck in traffic. Most frequently, such status updates include humorous remarks, swearing, frustration, and the occasional warning about traffic congestion on specific freeways (for others to see). Some Twitter users state humorously that the best time to tweet is during rush hour traffic, or that the 405 freeway is the only freeway where there's enough traffic to stop and tweet about the traffic (I-405 is a freeway in Los Angeles, California). We use this signal as a social sensor to model the circumstances and traffic conditions, and how the drivers' frustration may have an impact in the observed social discussion volume.

Indeed, we discovered that social reaction fluctuates in a non trivial manner. Different circumstances lead to different volumes of complaining about the traffic severity instead of following a strictly linear correlation. And while in many cases correlation is not equal to causality, for this particular experiment, the observed correlation between the real world (traffic) and the social reaction (tweets) is actually a causal relationship. The measured social reaction - tweets made by drivers stuck in traffic - is strictly caused by traffic congestion and the two variables are strictly dependent.

Finally, due to privacy reasons, the social data used in this study got anonymized, especially since as stated above, there are legal issues involved when tweeting while driving. It should be noted here that the processed social postings (publicly available tweets) are made by Twitter users with non private accounts and are openly provided by Twitter through the streaming API. However, to satisfy privacy and ethical concerns, we are not publishing any names, usernames, or content that could lead to the identification of specific users.

Contributions: The contributions of this work are listed below:

- We propose a novel model for traffic-severity regression based solely on the generated social volume. The proposed model exploits the fact that people complain in different levels throughout the day and **can be used to estimate traffic congestion in areas that lack proper traffic monitoring resources**.
- We offer a better understanding of human behavior when it comes to drivers and their social media actions while behind the wheel.

II. Related Work

There are two research fields related to the subject of the current work: 1) Studying and modeling of Traffic Congestion and 2) Social sensors utilized on online social media to mine information about physical events.

Traffic Analysis: There has been a lot of work and many studies that focus in the general analysis of traffic. They deal with questions like: How does traffic correlate with urbanization and economic growth? What causes traffic when there is no apparent reason? How does human behavior contribute in traffic congestion?

Traffic is studied in a plethora of areas, and a few of them are listed here: (a) Financial/Political: measuring urban growth [5], (b) Psychological: measuring human behavior, DUIs etc [22], [14], (c) Transportation: improving roadway conditions [11], and (d) Mathematics/Statistics: modeling traffic using statistical and mathematical frameworks [13].

Online Social Sensors: Social sensors and the discovery of what is happening in the real world through social media is a well studied area. Some representative works are listed here (list is not exhaustive): Kryvasheyeu et al. [15] examine how social sensors performed during hurricane Sandy (disaster control), García-Herranz et al. [10] utilized the social friendship network to quickly detect viral diseases, Zhaoet al. [28] use social media content to discover physical events in real time with a focus on sports events, and finally, Aggarwal et al. [2] wrote a book chapter that describes the current developments and challenges on social sensing in the context of data mining.

Studies that focus on social sensors specifically for the improvement of traffic reporting are closer to the problem tackled in our work [18], [12], [19], [23], [9], [20], [17]. In an ongoing Microsoft Research project [17] researchers try to combine the vast amount of historical data (both social and traffic) to create a single model for traffic prediction. Both works from Daly et al. and Ribeiro et al. [9], [20] mine the social sphere to identify/explain traffic conditions and events. In a publication by Pan et al. [18] a system is proposed for monitoring traffic via mobile cell signals in order to identify anomalies in the usual traffic flow. In a work authored by Jingrui He et al. [12], a way to improve traffic prediction is proposed, by combining social data from Twitter and historical traffic data. The authors use a raw, but localized, tweet stream to discover the users' future destinations and combine it with historical traffic data to produce a near-term (5 minutes to 1 hour) traffic prediction. The results show an improvement of the mean absolute percentage rate by almost 2% from the baseline model that only utilizes historical traffic data. Finally, [19] appears to be the only work that studies the correlation between social volume and traffic, at different hours of the day, but does not offer a model that captures their observations.

Approaches like the ones above can be improved with a more fine grained modeling that improves the correlation between social volume and traffic congestion. We propose such a model and show in Section V how this kind of traffic prediction can be potentially improved. To the best of our knowledge, all models in the mentioned publications ignore latent social factors that could skew the social volume related to traffic.

III. Data

A. California Traffic Data

The first step towards a combined traffic and social analysis is to obtain the necessary traffic congestion information and establish the ground truth. We focused in the area of California where the Department of Transportation (CALTRANS) collects a wide range of traffic statistics and publishes them online on the PEMS website [4]. CALTRANS maintains a plethora of physical stations known as Vehicle Detector Stations (VDS) on freeways across the state of California. Many sparsely inhabited areas have no stations but most metropolitan areas

like Los Angeles, San Francisco and San Diego are very well monitored. Each VDS is located next to a freeway and reports data like lane occupancy (if there are more than one lanes), speed in each lane, and health status, with a frequency of 5 minutes. For the purposes of this analysis, we did not use the raw data from the VDS stations since the PEMS website does not provide a programmatic way to download data for many stations. Instead, a very useful tool was utilized, provided by PEMS, that computes and reports all traffic bottlenecks on a daily basis.

Definition A **traffic bottleneck** occurs where the traffic demand exceeds the available capacity of the roadway facility.

More specifically, a bottleneck between two station detectors on the same freeway is observed under the following conditions:

- There is a speed drop of at least 20 mph (32 Km/h).
- The overall speed is less than 40 mph (64 Km/h).
- The distance between the two stations (minimum extent of a traffic jam) is at least 3 miles (4.8 km).
- The speed drop is observed for at least 70% of a 35 minute duration.

Note that these conditions have been chosen by CALTRANS. It's beyond the scope of this work to validate the above numbers, conditions, and semantics of traffic congestion. Since we are using the same definition across the whole analysis, there is no bias that could skew our observations.

For each analyzed day, the full list of all reported bottlenecks in California is obtained. Each bottleneck consists of a location (VDS latitude and longitude), extent, duration, and delay. Extent is the distance, in miles, of the reported traffic jam. Delay is the total duration, in minutes, of the congestion. Finally, delay is an artificial composite metric that describes the total loss of time due to the bottleneck and is measured in "vehicle-hours":

$$TotalDelay = N \times extent \times duration \times \left(\frac{1}{speed} - \frac{1}{35} \right)$$

where N is the total number of cars affected by the congestion and speed is the reported speed during a bottleneck. Note that this is a simplified version of the total delay formula [16]; PEMS is actually using the non publicly available knowledge of each lane's occupancy and corresponding speeds to increase the accuracy of the delay computation. In any case, due to the nature of this formula to combine all the other metrics (extent, speed, duration) as well as the total number of affected drivers, it is commonly used by traffic analysts [6], [27] as the indicator of how severe a traffic jam is. We will also be referring to it as "traffic volume" or "bottleneck severity".

One drawback of the PEMS-generated bottleneck report is that it does not provide an accurate time for each bottleneck (only the exact location). Instead, CALTRANS provides a low granularity time attribute named "shift" which takes the values AM, PM, and NOON. Therefore, bottlenecks can only be studied on a shift basis, which for the purposes of our paper is enough as shown later on. The AM shift includes the hours between 5am and 10am, the NOON shift between 10am and 3pm, and the PM shift between 3pm and 8pm. Bottlenecks that occur during the night or after hours are not reported and based on the raw traffic data, traffic-jams during those hours are extremely rare and would not be useful for a statistical analysis. As we will show in Section IV, very low traffic periods may occur even during the day, especially during weekend mornings or national holidays.

Daily traffic data was collected for every day within the period from May 2014 to October 2014. In order to match traffic jams with a physical location, we use the corresponding VDS station that observed each bottleneck, to identify the county/city and more importantly the exact freeway the station is measuring. Through the freeway name and number (e.g. US-101) we can then process the social data and collect tweets that correspond to a specific freeway's traffic jam.

B. Social Data

We use Twitter as the social sensor platform to study traffic jams. To obtain the necessary data we used the Streaming API [25]. Another explored alternative was the use of the Search API [24] which however does not provide any guarantees on the distribution or the completeness of the search results and therefore introduced statistical bias.

Using the streaming API, however, while guaranteeing completeness, has two drawbacks when compared to the search API. First, one can only collect data starting from the time the api calls begin and on (no historical data access). Second and most important, the streaming API does not support geo-enabled queries in a form that would be helpful to the current analysis. One may query for all tweets from California OR all tweets about traffic, but not their intersection. Alternatives exist, like collecting all tweets from Los Angeles and separately all tweets about traffic and then join them but due to the rate-limiting imposed by Twitter it would not be feasible to get all tweets from Los Angeles, given the large number of Twitter users living there. Not having the ability to filter tweets by location led us to collect any tweet that mentions the keyword "traffic" and then proceed to filter down the collected tweets using other heuristics. Specifically, only the tweets that mention the freeway name we are studying are kept, tweets from automated or traffic reporting accounts (like police departments and radio stations) are removed, and finally, human judges manually go through all remaining tweets and keep only those that were made by people stuck in traffic. The last step is consistently performed using basic rules like: tweet text contains phrases with temporal hints like "this traffic" or "on my way", tweet contains a picture of other cars in traffic jam taken from inside a car, tweet contains a self-taken picture of the driver (known as selfie).

The last filtering step to keep only tweets from people that drive in traffic, is the only one that needs human assistance to complete. It is still the most error prone step, since Twitter users will not always be explicit about being behind the wheel while tweeting. It's important to note here that interacting with a (smart)phone for purposes like texting, checking social media, tweeting etc. while driving, even during stand-still traffic, is considered illegal in California [7]. However, this

does not discourage people from posting selfies (self-portrait photographs) on Instagram, or tweeting about the annoying traffic. Still, the fact that such actions are deemed illegal makes it an interesting signal to study.

The final product of the social data collection is a set of tweets (including all meta-data provided by Twitter), grouped by date and shift (AM, NOON, PM), made by people while stuck in traffic jams. In the rare cases where a user made more than one tweets during a specific time period we counted only one of them. We will be referring to the number of tweets as "social volume" in this analysis.

In total, we gathered 3.2k tweets for the studied period of 6 months. Table II shows a more precise view of how these tweets are distributed in an average week. While the numbers might appear to be low, they are consistent in the duration of these 6 months.

C. I-405 Freeway

Given the mentioned limitations posed by the collection of social data, we focus on one major freeway, infamous for its devastating traffic jams: San Diego Freeway I-405. I-405, founded on 1964, has a length of 72 miles, passes through the whole city of Los Angeles and is used by hundreds of thousands drivers daily and there is always stand-still traffic reported during rush hours. People even call it the "monster" [1] as a humoristic acknowledgment of its size and severe traffic. Traffic congestion on I-405 is not evenly distributed but instead there are some specific points where traffic jams mostly occur, which makes traffic at these points even more severe during rush hour. We chose I-405 over US-101 (another popular candidate) because it is limited in the area of Los Angeles while US-101 covers the whole west coast of the United States. However, we made sure that the traffic patterns observed in I-405 are not unique. The traffic volume between the two freeways was compared and we found that they follow the exact same patterns for all days of the week and all shifts of the day. Therefore, it is safe to say that the choice of I-405 does not introduce any freeway-specific traffic anomalies.

D. Tweets from Drivers

As explained in subsection III-B only tweets made by people driving during traffic jams are counted, instead of every tweet mentioning traffic and the freeway name. Utilizing the latter as the social volume, would introduce cases where the raw volume of noisy tweets is misleading for estimating the actual traffic. There are two categories of "noisy" tweets. First, there are tweets made by automated accounts (e.g. police dispatch, highway patrol) or news agencies that report traffic on Twitter [8]. Such tweets are published whenever traffic bottlenecks occur and are usually agnostic of the exact severity of the bottleneck or how much it really annoys the drivers. The second category consists of tweets that are potentially about traffic, posted by normal users, but not during their commute. The problem posed by both categories is that those tweets are not part of a direct social reaction to a traffic jam. Any traffic jam estimation that utilizes those tweets would introduce excessive noise and predictive bias. As an example of a case where the raw volume of all tweets is misleading, on Friday the 23rd of March 2014 a new carpool lane opened for freeway I-405 which caused an abnormally high volume of discussion among Twitter users. Most of this discussion included chatter about the potential usefulness of this new lane or excitement about it. On another similar case, a celebrity Twitter user made a tweet about being stuck on traffic which triggered many replies from fans and followers. In both cases, any conclusions or traffic modeling based on the generated "social reaction" will be very biased unless the data is correctly processed and filtered.

In Section V we will compare the traffic regression error between a model that uses tweets only from drivers and a model that uses tweets from every normal Twitter user that talks about traffic (automated accounts, news stations, and bots are still removed).

IV. Analysis

The purpose of the current analysis is to discover hidden features that could yield better results for estimating the magnitude of traffic congestion through social media. Our basic assumption is that there are cases where the size of an event may be different from how humans perceive it. Perception is a complicated process and there are many factors that play their role (e.g. mood, enthusiasm, weather, family status, political beliefs, etc). We assume that complaining about traffic falls under the umbrella of such events and study the correlation between traffic and complains to show that indeed there are other latent factors that contribute in non-trivial fluctuations of the social reaction volume. Traffic jams are measured with high accuracy by automated traffic monitoring stations but the human perception of a bottleneck may vary under different circumstances. To the best of our knowledge this is the first work to study how fluctuations, potentially due to psychological factors like mood or sentiment, can improve the accuracy of a social sensor.

A. Basic Data Statistics

To begin the analysis, a better understanding of the two datasets (traffic volume and social volume) is necessary. As mentioned at the end of Subsection III-B our analysis is focused on the California freeway I-405. Table I and Figure 1 show the traffic volume on I-405, by day of the week and shift of the day. Only weekdays are shown since traffic congestion during weekends is extremely low. Note again that these numbers describe the total delay and not the amount of cars traveling. Close to zero traffic volume in our context means that there is no introduced delay since the cars in the freeway are running at a speed close to the limit and not that there is no traffic at all. Our proposed model works for weekends as well but they are omitted from the current analysis for simplicity. The reader can view the actual statistics for weekends in tables I and II. These tables also provide the standard deviation for each average.

The first observation based on the traffic volume data is a clear traffic increase towards the end of the day (PM). Also,

Day of the week	AM mean	AM stdev	NOON mean	NOON stdev	PM mean	PM stdev
Monday	16840.25	2927.09	3462.38	1271.59	21234.75	3234.97
Tuesday	18747.29	1907.23	5299.43	3212.63	27126.57	3705.78
Wednesday	19708.20	2741.86	5451.60	2473.53	34484.80	2725.18
Thursday	19167.00	3225.76	7585.11	1764.80	40134.67	4830.19
Friday	11997.67	2857.14	13364.78	2270.96	41370.00	3769.05
Saturday	200.25	50.77	7038.50	2332.95	9759.00	2767.73
Sunday	54.50	91.71	2893.25	1231.31	3020.25	907.63

TABLE I. TRAFFIC VOLUME (TOTAL DELAY) STATISTICS FOR I-405 (LOS ANGELES) BY DAY OF THE WEEK. TO MEASURE TRAFFIC VOLUME WE SUM UP THE TOTAL DELAY OF EACH REPORTED BOTTLENECK ACROSS I-405 DURING EACH DAY'S SHIFT.

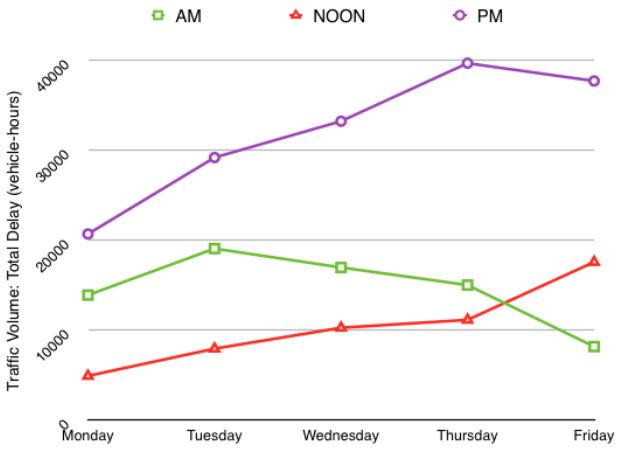

Fig. 1. Traffic averages for each day of the week and shift of the day. The general trend for most of the weekdays (no weekend) is that PM traffic is always worse than AM and NOON and AM is worse than NOON except on Fridays.

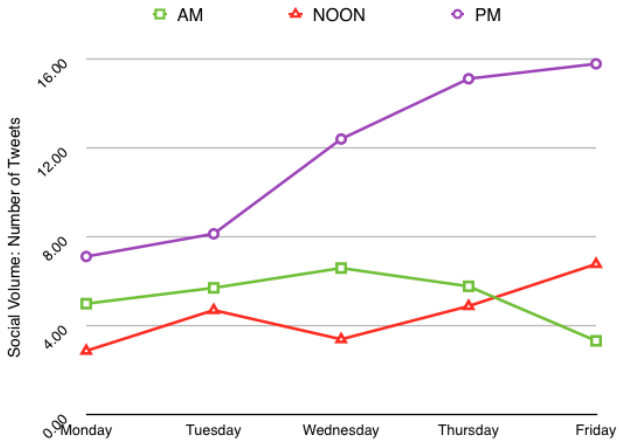

Fig. 2. Social volume averages by day of the week and shift of the day. The general trend of social reaction appears to be in sync with the traffic volume (if compared with the plots in Figure 1.

for every weekday, the morning and noon traffic fluctuate far less than the evening's. The second observation is that evening traffic gets worse towards the end of the week (Thursday and Friday) as can be seen in Figure 1. There are many potential explanations as of why these patterns occur. Arguably, the reasons why most people drive during rush hours are work related. Therefore, most patterns could be explained by how people schedule their work hours. For example, it could be that during Fridays people tend to leave earlier from their work and cause a more concentrated traffic congestion around 4pm and 5pm. Regardless of the reason, the fact remains that traffic volume is higher on evenings and towards the end of the week, and lower in the noons and mornings.

Similarly to the traffic volume, Table II and Figure 2 show statistics about the social volume (number of tweets), again on a day-of-the-week and shift-of-the-day basis.

The social volume statistics confirm our intuition that social reaction is proportional to the traffic volume. Same as with the traffic, during morning and noon hours social volume is generally low across all days of the week but peaks up during the evening hours. Also, the evening social volume becomes higher towards the last days of the week (Thursday and Friday).

B. Naive Approach: Linear Model

From the basic statistics we listed in Subsection IV-A it would be reasonable to expect a linear relation between traffic volume and social volume. It makes absolute sense that social reaction becomes stronger when traffic jam conditions worsen. Based on this hypothesis we can use linear (least squares) regression to compute a linear model that can estimate traffic based on the number of generated tweets (a typical example of social sensors). Figure 3 depicts the linear model as a straight line:

$$TrafficVolume = 1850.0 \times SocialVolume + 5299.1$$

Note that the model's coefficient of determination (R^2) is 0.6597 which can be considered high depending on the application and desired level of regression precision. We list in Section V the absolute and relative errors yielded by this model when trying to estimate (predict) traffic.

We also tried to fit a second degree polynomial model to the data. The result was a minor improvement of the R^2 value but unfortunately, due to physical limits, greater traffic volume values that could validate a polynomial model do not exist. Without loss of generality or introducing any bias for further findings, we assume a linear fit for the purpose of this study.

While the linear model appears to be relatively accurate, certain underlying patterns exist, which are ignored. Plotting the same data from Figure 3 and grouping datapoints by shift of the day in Figure 4, makes it clear that each group extends in it's own space in the graph. The conclusion from this observation is that latent features might describe the connection between traffic and social reaction in a better way. This conclusion lead us to the hypothesis that a different model that exploits such patterns could fit better than the naive linear model.

Day of the week	AM mean	AM stdev	NOON mean	NOON stdev	PM mean	PM stdev
Monday	5.00	1.51	2.88	1.64	7.12	2.47
Tuesday	5.71	2.36	4.71	3.64	8.14	2.41
Wednesday	6.60	2.51	3.40	2.30	12.40	2.19
Thursday	5.78	2.54	4.89	1.69	15.11	3.41
Friday	3.33	2.06	6.78	3.07	15.78	5.63
Saturday	1.25	1.04	4.88	2.36	5.38	4.24
Sunday	0.25	0.71	2.00	0.93	1.00	1.07

TABLE II. SOCIAL VOLUME (NUMBER OF TWEETS) STATISTICS. THESE ONLY INCLUDE TWEETS MADE BY DRIVERS STUCK ON I-405 TRAFFIC JAMS.

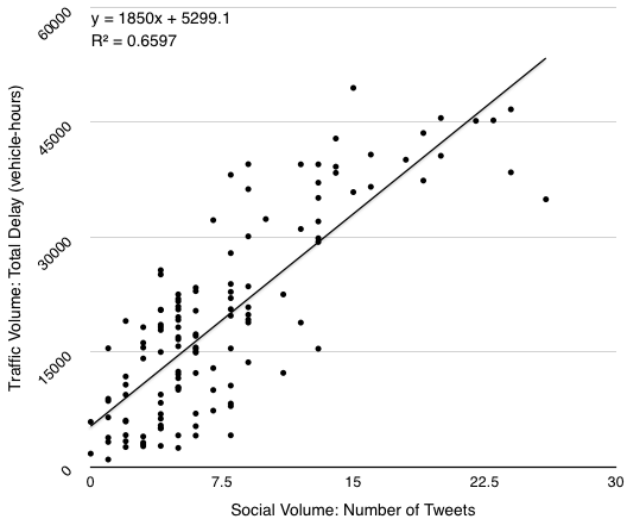

Fig. 3. Plot of traffic volume vs social volume. Each point describes the data of a single day and shift. The x-axis measures the social volume (number of tweets) and the y-axis measures the traffic volume as total delay (vehicle-hours). We can fit a linear model with R^2 value of .6597.

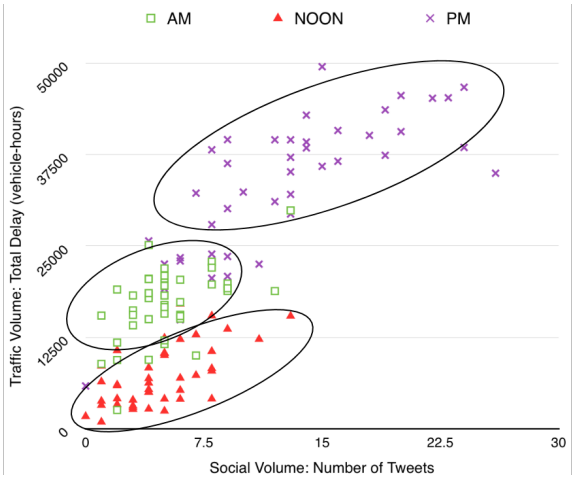

Fig. 4. Same datapoints from Figure 3 but grouped by shift of the day. PM datapoints are mostly located on the upper-right, AM datapoints at the center-left, and NOON datapoints at the lower-left.

C. Analysis by Time of the Day

To evaluate whether traffic is perceived differently under different circumstances we computed the ratio of Traffic Volume over Social Volume for different times of the day (averaged across all weekdays). We also tried to explore correlations with the day of the week or the weather (temperature) but the time

Shift of the Day	Social (SV) to Traffic (TV) Model	R^2
AM	$TV = 974.79 \cdot SV + 12122$	0.2761
NOON	$TV = 916.09 \cdot SV + 2928.6$	0.3990
PM	$TV = 1211.2 \cdot SV + 17849$	0.5941

TABLE III. SOCIAL-TO-TRAFFIC MODELING, BY SHIFT OF THE DAY (SHIFT-BASED MODEL).

of the day proved to be by far the strongest feature. The ratio of traffic volume over social volume measures how much drivers complain per traffic delay and lower values indicate higher complaining. Note that due to the limitation of the traffic jam dataset, the analysis is performed on a shift basis (AM shift: 5AM-10AM, NOON shift: 10AM-3PM, PM shift: 3PM-8PM). A plot of these ratios can be found in Figure 5. On the right-most column of the chart, the average ratios across all weekdays are shown.

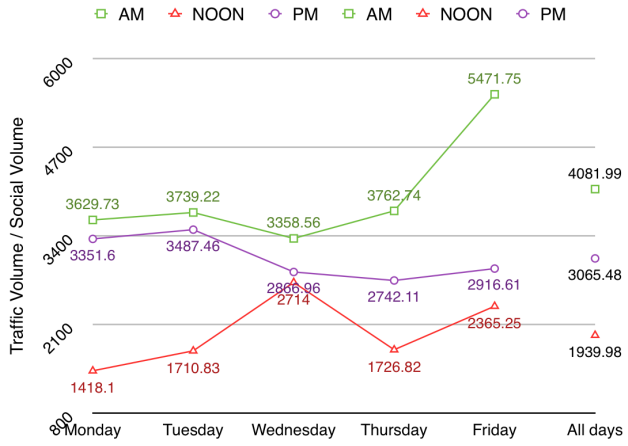

Fig. 5. Traffic volume / Social volume ratios. Lower values indicate heavier social reaction. Morning social reaction to traffic appears to be the lightest while noon reaction is the heaviest. Humans react to traffic congestion differently based on the hour of the day. Even though NOON traffic is the lightest (Figure 1) it causes the most severe social reaction.

Through these ratios the conclusion is made that a different time of the day indeed results in different levels of traffic reaction. In Figure 6 the datapoints are plotted based on the shift (AM, NOON, and PM). We can then fit individual models on each subset of the data. In Figure 6 the linear models are plotted using least squares regression. As with the naive liner model (subsection IV-B), we also tried to fit other models (polynomial, exponential) but the linear yields the best results even if not all individual R^2 values are high enough. The 3 individual sub-models for each shift of the day and the corresponding R^2 values are listed in Table III.

Note that the individual R^2 values for each shift are lower than the R^2 value of the linear model (which is 0.66). While

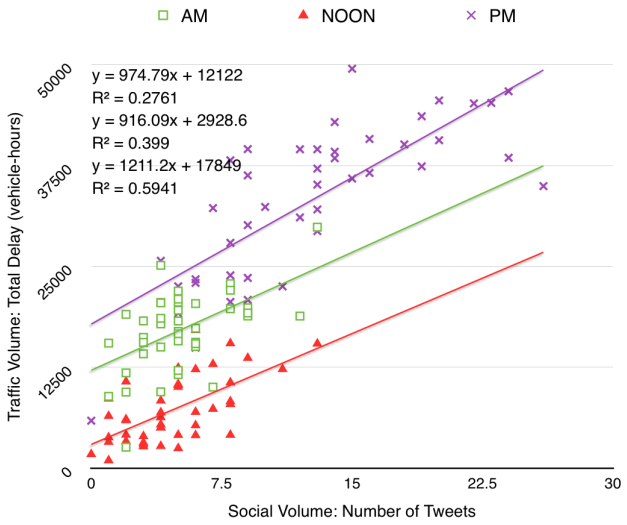

Fig. 6. Shift-based linear model: A mixture of 3 different linear models, one for each shift of the day (AM, NOON, PM).

this could be interpreted as a bad fit of the proposed model to the data, we will show in our experimental analysis in Section V how the proposed model compares to the naive linear model and other baselines, when used in the context of estimating traffic through social volume. Generally, R^2 values are not always the best indicator of well-fitness and in cases where residuals form specific patterns, can be misleading. In any case, the actual superiority of our model will be shown through its regression accuracy.

V. Traffic Congestion Prediction through Social Sensors

In this section we describe the exact details of the shift-based model and provide comparisons between the proposed model, the naive linear approach and some additional baselines. Note that the term prediction is used in the context of statistical regression and social sensors and not predicting future traffic.

A. Models

To measure the regression improvement of the proposed shift-based model we introduce some baseline models. The first baseline model is the naive linear model that was described in Subsection IV-B (denoted as NAIVE). Since the shift-based model is practically splitting the datapoints in three categories, it should be compared with a 3-random-partitions model that just picks 3 random partitions and fits a linear model on each one (denoted as RAND3). Random partitioning makes sense as a baseline because if the proposed shift-based model had no statistical significance, then it should yield similar results with the random partitioning.

Similarly to the shift-based model we also tried to fit the data on a daily basis – one linear fit for each day, from Monday to Friday (baseline denoted as DAY-BASED). Finally, two more models are introduced that use the naive linear model (NAIVE) to fit the datapoints of each day of the week (NAIVE-DAY) and the datapoints of each shift (NAIVE-SHIFT). The last two models are fixed, not generated by training data, and don't require cross validation; we measure their fitness purely for comparison purposes.

Model	Mean Error			R^2
	Squared	Absolute	Relative	
Naive Linear	5.5856	**6062.0**	**0.6619**	0.6596
Random 3 Partitions	5.8604	6209.8	0.6783	0.6611
Day-based	5.9272	6374.8	0.6658	0.6064
Naive by Day	5.406	5983.5	0.6607	0.6064
Naive by Shift	5.3690	5958.2	0.6584	0.4230
Shift-based	2.4245	**3739.8**	**0.3598**	0.4230

TABLE IV. Error comparison for each regression model. Squared error values are $\times 10^7$.

Model	Mean Error			R^2
	Squared	Absolute	Relative	
Naive Linear	8.5558	7378.3	0.7259	0.4948
Shift-based	3.5027	4527.4	0.3919	0.2571

TABLE V. Error comparison for the linear and shift-based model with all tweets about traffic (no driver-based filtering). When tweets are not coming directly from drivers in traffic jam the error is significantly higher.

The shift-based model (denoted as SHIFT-BASED) is a composite model that consists of three linear submodels, one for each shift of the day (AM, NOON, PM). Since the number of datapoints for each shift is equal, the overall precision of the shift-based model can be defined as the average precision of each submodel. For example, when measuring the squared error of the model we need to compute the squared error for each submodel and then get their average.

B. Model Comparison

To compare the predictive power of each model the following cross validation setup is used: Repeated random subsampling validation. For each model, the data points are randomly ordered and then the first 80% of the datapoints is picked as training dataset and the rest 20% as validation dataset. Using least square regression we fit a linear model to the training data and then calculated the estimation error on the validation data. This process is repeated 1000 times and the average errors across all splittings are calculated. For the relative errors, all cases where the expected traffic volume is close to 0 were ignored, since it was introducing very large values.

The average squared, absolute, and relative errors for each model are listed in Table IV. For the sake of analysis-completeness we also provide the coefficient of determination in each case. The Shift-based model significantly outperforms all the baseline models which proves that focusing on the different shifts of the day has a statistically significant effect while other approaches like day-based perform poorly. In terms of absolute error, we observe a 38% improvement between the Naive Linear approach and the shift-based model. In terms of relative error we observe more than 45% improvement.

In Table V we list the average errors for the linear and shift-based models without having applied the driver constraint (tweets must originate by drivers while they are stuck in traffic). Basic filtering that removes tweets from automated accounts and bots is still applied but all the rest of the tweets

from normal Twitter users remain. Using this raw dataset for regression, results in an increased error for both linear and shift-based models. The conclusion from this comparison is that filtering of social posting based on users directly affected by traffic congestion results in a better model and accuracy.

Note again, that even though the Coefficient of Determination is lower for the proposed model compared to the majority of the baselines, the Shift-based model can achieve a very significant improvement in traffic estimation which shows that R^2 is not a good measure of fitness when modeling this particular traffic/social dataset.

VI. Conclusions and Future Work

Social sensors offer a fast and low cost way to understand the physical world through online content on social media. Mining the correct correlation between the crowd's reaction and an event's magnitude can be very critical and improves our understanding of what is happening and how much it effects our lives. Using the correlation between traffic congestion and social reaction on Twitter as a showcase we show that exploring dimensions that have different psychological links, like the time of the day, can lead to a better grasp of the traffic severity. We propose a novel model to estimate traffic jams using social sensors, that utilizes three linear submodels, one for each shift of the day (AM, NOON, PM) and social posting from car drivers. We show that the proposed model can be at least 38% better than the naive linear approach and performed several comparisons with different baselines to prove that these findings are statistically significant. Finally, we offer the exact linear sub-models that describe the relation between complaints and traffic, for different times of the day.

The next step for this study is to create an automated tweet classifier that could identify tweets made by drivers during traffic jams which for the current paper had to be done manually by human evaluators. Another interesting direction is to analyze additional signals that may alter the traffic complains patterns, like the time of the day does. Namely, the temperature in the area of the traffic bottleneck could have a negative effect and make drivers complaint more intensely, although our initial experiments have shown no strong correlation.

VII. Acknowledgments

This research was sponsored in part by the Army Research Laboratory under cooperative agreements W911NF-09-2-0053, NSF IIS 0917228, and NSF IIS 1135389. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

References

[1] Los angeles freeways: The great 405. http://www.davestravelcorner.com/guides/losangeles/LA-Freeways.

[2] C. Aggarwal and T. Abdelzaher. Social sensing. In C. C. Aggarwal, editor, *Managing and Mining Sensor Data*, pages 237–297. Springer US, 2013.

[3] D. Boyd and K. Crawford. Critical questions for big data. *Information, Communication and Society*, 15(5):662–679, 2012.

[4] Caltrans performance measurement system (pems). http://pems.dot.ca.gov.

[5] R. Cervero. Road expansion, urban growth, and induced travel: A path analysis. University of california transportation center, working papers, University of California Transportation Center, 2001.

[6] C. Chen, A. Skabardonis, and P. Varaiya. Systematic identification of freeway bottlenecks. In *Proceedings of 83rd Transportation Research Board Annual Meeting*, 2004.

[7] Text messaging law, california, 2008. https://www.dmv.ca.gov/cellularphonelaws.

[8] Twaffic - will twitter and tweets about traffic change the way we drive? http://www.slate.com/articles/life/transport/2011/04/twaffic.html.

[9] E. M. Daly, F. Lecue, and V. Bicer. Westland row why so slow?: Fusing social media and linked data sources for understanding real-time traffic conditions. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 203–212, New York, NY, USA, 2013. ACM.

[10] M. García-Herranz, E. M. Egido, M. Cebrián, N. A. Christakis, and J. H. Fowler. Using friends as sensors to detect global-scale contagious outbreaks. *CoRR*, abs/1211.6512, 2012.

[11] T. Golob and W. Recker. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of Transportation Engineering*, 129(4):342–353, 2003.

[12] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence. Improving traffic prediction with tweet semantics. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI'13, pages 1387–1393. AAAI Press, 2013.

[13] Kai Nagel and Michael Schreckenberg. A cellular automaton model for freeway traffic. *J. Phys. I France*, 2(12):2221–2229, 1992.

[14] W. Knospe, L. Santen, A. Schadschneider, and M. Schreckenberg. Human behavior as origin of traffic phases. *Phys. Rev. E*, 65:015101, Dec 2001.

[15] Y. Kryvasheyeu, H. Chen, E. Moro, P. V. Hentenryck, and M. Cebrián. Performance of social network sensors during hurricane sandy. *CoRR*, abs/1402.2482, 2014.

[16] J. Kwon, B. McCullough, K. Petty, and P. Varaiya. Evaluation of PeMS to improve the congestion monitoring program. Technical report, Final Report for PATH TO 5319, UC Berkeley, Berkeley, CA, 2006.

[17] Microsoft azure helps researchers predict traffic jams. http://blogs.msdn.com/b/msr_er/archive/2015/04/02/microsoft-azure-helps-researchers-predict-traffic-jams.aspx.

[18] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 344–353, New York, NY, USA, 2013. ACM.

[19] A. I. J. a. T. Ribeiro, T. H. Silva, F. Duarte-Figueiredo, and A. A. Loureiro. Studying traffic conditions by analyzing foursquare and instagram data. In *Proceedings of the 11th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks*, PE-WASUN '14, pages 17–24, New York, NY, USA, 2014. ACM.

[20] S. S. Ribeiro, Jr., C. A. Davis, Jr., D. R. R. Oliveira, W. Meira, Jr., T. S. Gonçalves, and G. L. Pappa. Traffic observatory: A system to detect and locate traffic events and conditions using twitter. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '12, pages 5–11, New York, NY, USA, 2012. ACM.

[21] R. Singleton, B. Straits, and M. Straits. *Approaches to social research*. Oxford University Press, 1993.

[22] H. Summala. Accident risk and driver behaviour. *Safety Science*, 22(1-3):103 – 117, 1996. Risk Homeostasis and Risk Assessment.

[23] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson. Vtrack: Accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, SenSys '09, pages 85–98, New York, NY, USA, 2009. ACM.

[24] Twitter search api. https://dev.twitter.com/docs/using-search.

[25] Twitter streaming api. https://dev.twitter.com/docs/streaming-apis/streams/public.

[26] Waze: Community-based traffic app. https://www.waze.com.

[27] C. Winston. On the performance of the u.s. transportation system: Caution ahead. *Journal of Economic Literature*, 51(3):773–824, 2013.

[28] S. Zhao, L. Zhong, J. Wickramasuriya, and V. Vasudevan. Human as real-time sensors of social and physical events: A case study of twitter and sports games. *CoRR*, abs/1106.4300, 2011.