Analysis of Mobile User Behavior in Vehicular Social Networks

Victor R. Neto, Dianne S. V. Medeiros and Miguel Elias M. Campista
Universidade Federal do Rio de Janeiro - PEE/COPPE/GTA - DEL/POLI - Rio de Janeiro, Brazil
Email:{vrneto, dianne, miguel}@gta.ufrj.br

***Abstract*—Participatory social networks can provide great amount of data about users and their surroundings. When properly crafted, this data can be used as an important source of information about human behavior. In this paper, we use a vehicular social network aiming to evaluate the impact of external factors from vehicular environments on users' contributions to social networks. Results from a publicly available Waze dataset show that users are mostly motivated to post traffic jam information and that they do it during rush hours on weekdays and during the afternoon on weekends. We also observe that users who receive low reliability on their posts tend to keep low scores in the following. Finally, results additionally indicate that users at higher speeds do not contribute to the network and that posts experiencing longer delays until published are poorly evaluated.**

I. Introduction

The growth of social networking combined with new market trends makes understanding the human behavior a key strategy for business improvements [1]. Nevertheless, retrieving useful information from social networks is not trivial. Data must be carefully crafted, which is even more challenging in dynamic environments such as vehicular networks. In these networks, participants face intermittent connectivity, further introducing timing constraints [2].

Typical analyses of social networks do not consider the existence of networking issues. These works often aim to solve deficiencies in recommendation systems, detect users' routine behavior, or semantically identify users' mood in a post [3], [4], [5]. In vehicular environments, more specific challenges are addressed, such as the proposal of new mobility models or social tie recommendations [6], [7], [8]. To accomplish that, information can be obtained from mobile applications for vehicular social networking. In this direction, Waze is one of the most popular applications for vehicular environments [9] with main focus on information exchange of traffic conditions. Each user posts information regarding the traffic and this information is evaluated by other participants. Based on the evaluation, the post has a reliability score assigned.

The joint analysis of users' behavior and the influence of networking issues, herein called external factors, can possibly reveal different insights concerning information reliability. Even though, apparently, users' behavior can be considered as the sole responsible for information reliability, vehicular mobility and the network connectivity can also interfere. In vehicular scenarios, disconnections during social networking interactions can affect the reliability of some alerts as they are injected late into the social network. These alerts are susceptible to low reliability evaluations, as events may no longer exist. Yet, depending on the impact on users' routine and on the expected duration of the event, users' can feel more motivated to contribute. In any case, the data obtained can be composed of unreliable information as a consequence of network conditions or event characteristics.

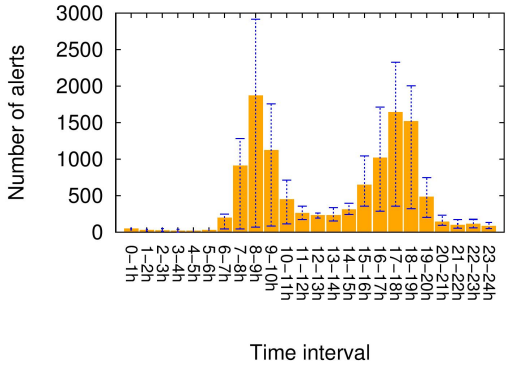
In this paper, we aim to evaluate the impact of external factors on users' behavior. To accomplish that, we analyze a publicly available Waze dataset [10], which suits well our main purposes. From the Waze dataset, it is possible to verify the frequency of users' contributions and the main alerts triggered, the influence of users' speed and transmission delays on information reliability, and the impact of the first evaluation on the final reliability of an alert. With all these possibilities in hands, we show that users feel more motivated to contribute to Waze during rush hours on weekdays and during the afternoon on weekends. Yet, traffic-jam alerts are the preferred ones. Results also show that the speed of vehicles and the transmission delays indeed have a great influence on the information reliability. We observe that the number of alerts becomes negligible above 30 km/h and that alerts with higher delay have lower reliability. We also found that recently received alerts are assigned lower reliability and, once poorly evaluated, the evaluation does not change. Lastly, our dataset shows that users do not regularly contribute to the social network, but when they do, the time interval between the first and last contribution is short.

The remainder of this paper is structured as follows. Section II overviews the Waze application and describes the dataset used. Section III discusses the obtained results. Section IV presents the related work. Finally, Section V concludes this work and presents future directions.

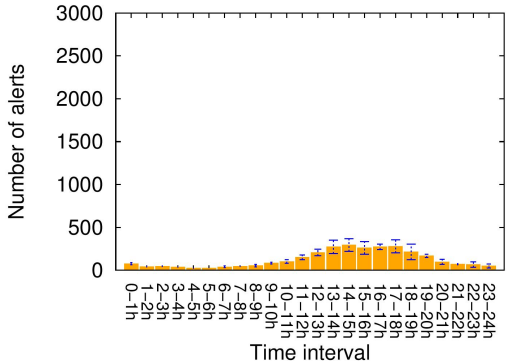
II. Waze Dataset

Waze is an application for mobile devices based on satellite navigation. It provides real-time information about users and trajectory details, according to the device location. It differs from traditional GPS navigator due to its features related to social networking such as circle of friends and messages exchange. In addition, Waze learns how users drive to provide trajectories and traffic updates in real time. People can report

978–1–5090–4671–3/16/$31.00



(a) Weekdays: from Monday to Friday.



(b) Weekend: from Saturday to Sunday.

Fig. 1. Hourly distribution of the average number of alerts. The dotted vertical bars show the maximum and the minimum number of alerts during the entire period.

accidents, congestion, road upgrades, landmarks, house numbering, etc. Alerts posted by users are evaluated using a "like" button or comments. Based on the evaluations received, Waze can build the reliability of a post.

A. Dataset description

Only a few datasets from social networks are publicly available. According to our knowledge, the only public dataset from Waze available is the one provided by the Massachusetts government [10]. This dataset contains data captured during one week, between February 22th and March 1st 2015, totaling 100 MB of data. The data is arranged in tables, where we find information about users and corresponding alerts such as: user ID (anonymized), geographical position, speed, street and city names, alert type and subtype, start and injection time information, delay, and reliability. It is important to highlight that Waze classifies these alerts into four types (jam, weather hazard, road closed, and accident), further divided into 24 subtypes. Additionally, Waze assigns reliability to a post varying from 5, the lowest reliability, to 10, the highest.

The dataset documentation does not precisely describe all existing information. We deduce the meaning of some, based on the expected semantics. We assume the start time is the moment when a user interacts with the application, e.g., when the user sends information; the injection date contains the time when the social network publishes the result considering the received information and finally, the reliability is the evaluation received by a post. The speed does not explicitly inform the unit, but we assume it is given by meters per second as the maximum speed allowed in the Massachusetts area indicates. We also adjust the dataset timezone to represent the timezone of Massachusetts in February 2015.

TABLE I
TOTAL NUMBER OF ALERTS PER TYPE.

Type	Total
Jam	50,895
Weather Hazard	16,013
Road Closed	3,320
Accident	1,277

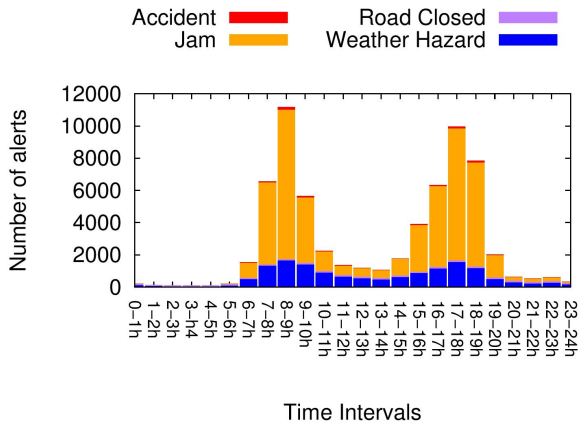
We additionally sanitized the dataset, removing alerts produced by vehicles at unfeasible speeds. As a consequence, we removed from the dataset all alerts with unrealistic speeds, representing 0.016% of the total. Investigating these *outliers*, we found that the IDs of all users producing records with unfeasible speeds are different. Therefore, we assume this is a consequence of technical issues, probably during GPS information upload to the social network.

III. RESULTS

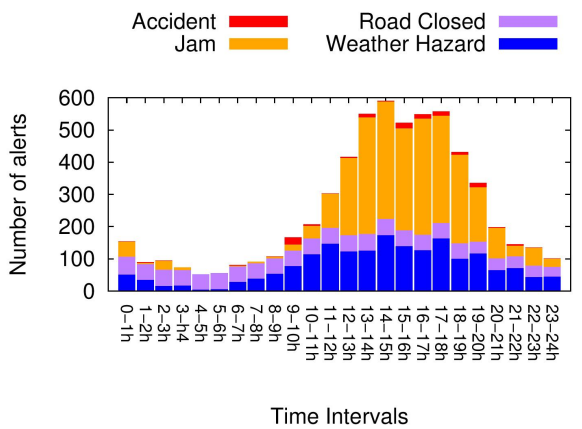
We divide our analysis into five parts. Firstly, we derive users' preferred hours for posting information. We also verify alert types and subtypes triggered more often by users in an hourly basis. Secondly, we investigate the variation of the reliability assigned to a post. We want to observe the typical first evaluation to a post and if this evaluation changes. Thirdly, we check if drivers' speed influences on the number of alerts triggered and on their reliability. The hypothesis is that drivers at higher speeds interact less with the mobile device. Fourthly, we also verify the relationship between the reliability and the posting delay. We expect that alerts with higher delay have lower reliability and lose value due to temporal aspects. Fifthly, we analyze the interval of consecutive contributions of each user, examining if the number of contributions has any relation to the time interval between consecutive contributions. Each of these relationships adds knowledge about Waze users' behavior.

A. Hourly distribution of alerts

In our first analysis, we investigate the average number of alerts triggered by users per hour during weekdays and weekends. Fig. 1 shows, besides the average, dotted vertical bars with maximum and minimum values in the corresponding hour during the whole measurement. In Fig. 1(a), we note peaks of alerts around 8-9h and 17-19h, which demonstrate that users contribute with the vehicular social network mostly during "rush hours". In Fig. 1(b), on the other hand, the same trend does not exist and the number of alerts increases at noon until 18h. This shows a completely different behavior from



(a) Weekdays: from Monday to Friday.



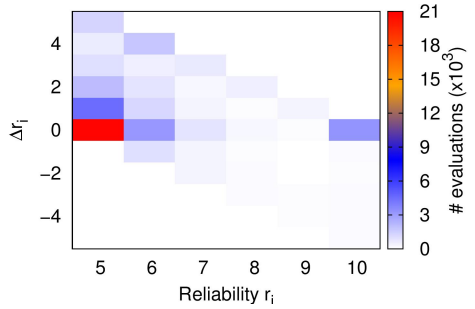
(b) Weekends: from Saturday to Sunday.

Fig. 2. Hourly distribution of the cumulative number of alerts separated by type.

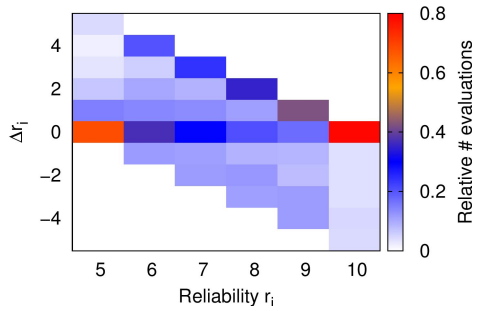
users compared with weekdays, as rush hours do not affect as much.

We further investigate the number of alerts, separating the values by type. Table I shows the cumulative number of alerts separated by type, considering all the valid posts in the Waze dataset.

The most representative type of alert is "Jam", which occurs three times more than the second most triggered one, "Weather Hazard". This indicates that users are more motivated to contribute in case of vehicular traffic issues. Figs. 2(a) and 2(b) show the cumulative hourly distribution grouped by type, during weekdays and weekends, respectively. In Fig. 2(a), we note that jam alerts are predominant over the other types, mainly during rush hours. In Fig. 2(b), during weekends, even though jam alerts are still predominant, the relative difference to the other types are not that high. Taking a closer look at the subtypes, we confirm that the most triggered alerts are of the type jam, subtypes "Heavy Traffic" and "Still Traffic", with more than 10,000 alerts each.



(a) Absolute viewpoint.



(b) Relative viewpoint.

Fig. 3. Reliability variation after receiving the first rating.

B. Reliability variation

We show in Figs 3(a) and 3(b) the trend of the reliability scores after the first rate. In these figures, the X-axis represents the first rate given to a post (r_i), and the Y-axis shows the difference between the reliability of the following evaluations compared with the first (Δr_i = r-r_i). In Fig. 3(a), the color denotes the absolute number of evaluations with the same Δr_i given the first reliability r_i. We observe that the reliability 5 is predominant and remains unchanged in the following evaluations. This is observed in the plot with the reddish color in $(x, y) = (5, 0)$. Hence, Waze is more likely to evaluate received posts with the minimum score, which is not often changed. When they change, it is only a small variation from reliability 5 to 6, for instance. The second most used reliability score is 10. We observe that posts well evaluated also remain considered reliable. The other scores, from 6 to 9, are barely used. This indicates a user bimodal behavior as they evaluate posts in such a way that makes Waze either consider the post reliable with 10 or unreliable with 5. Taking a look at the score variation, we note that there is a greater tendency of improvement rather than deterioration, as shown by the higher concentration of blueish colors above $\Delta r_i = 0$.

Fig. 3(a) hides the reliability variation when users begin with r_i different than 5. In Fig. 3(b), we highlight such variation by computing $\Delta r_i / \sum_{-5}^{5} \Delta r_i$ for the same r_i and setting the color accordingly. We observe that users tend to maintain their scores when starting with 5 and 10. Nevertheless, when they start with reliability between 6 and 9, they have a higher improvement trend.

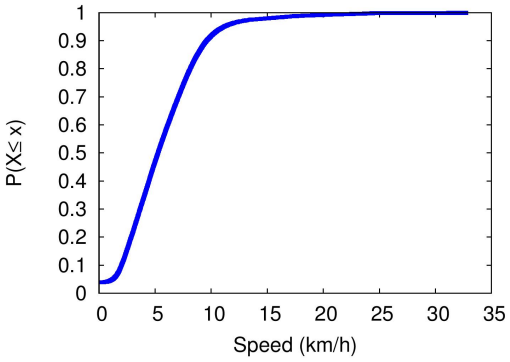


Fig. 4. CDF of the speed of vehicles when the user triggers an alert.

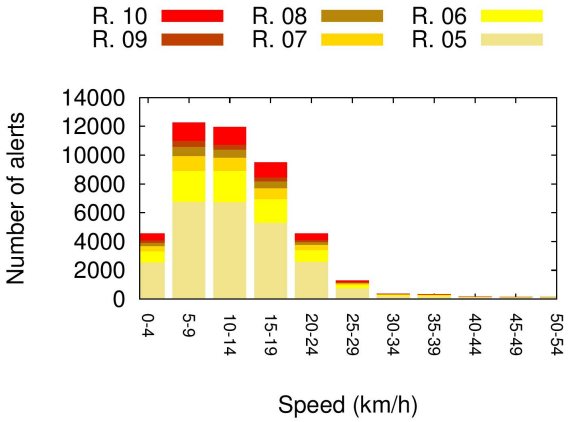


Fig. 5. Relationship between users' speed, number of alerts, and the reliability of each post.

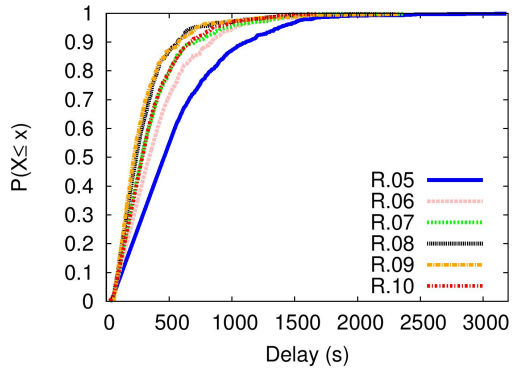


Fig. 6. Relationship between delay, reliability and total number of alerts.

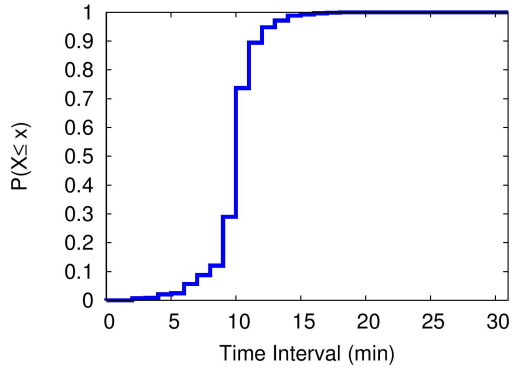


Fig. 7. CDF of the time interval between consecutive contributions of users.

C. Vehicle speed impact

We further investigate the influence of the speed on the number of alerts and on their corresponding reliability. Fig. 4 shows a cumulative distribution function (CDF) that allows us to check the most common used speeds for triggering alerts. We note that 90% of the interactions with the network happen with speeds lower than 30 km/h. Similarly, Fig. 5 shows that users interact with the social network at low speeds. The number of contributions increases until 10 km/h and, after that, it quickly decreases. This result corroborates our previous analysis, where we observed that users are more likely to contribute when the vehicular traffic is jammed. Furthermore, one could expect that users tend to interact less with the social network at higher speeds, as they need to pay more attention to the road. In addition, we note that the reliability of the received alerts is usually low and from 30 km/h the number of alerts with high reliability becomes negligible. This also confirms that users often evaluate received posts with low reliability.

D. Posting delay impact

We also evaluate the relationship between the delay and the corresponding reliability. Fig. 6 shows a cumulative distribution function (CDF) that allows us to observe that contributions with higher delay have lower reliability. The rationale behind is that alerts that spend more time to be published on the social network may lose timing. In opposition, we can note that the contributions with higher reliability are those with lower delays, as they are quickly published.

E. Posting frequency

We also evaluate how often a single user contributes to the social network. To accomplish that, we compute the interarrival time between consecutive posts at the social network. Fig. 7 shows the CDF of the time interval, in minutes, between consecutive contributions triggered by the same user. Most interarrival times, approximately 90%, are near 10 minutes. This means that most users send a new post after 10 minutes from the previous contribution.

Fig. 8 shows the time interval, in hours, between the first and the last contribution from the same user, and the number of alerts triggered by this user. The obtained results indicate that users with many contributions tend to have longer intervals between their first and last posts, whereas users with fewer contributions send consecutive posts within shorter intervals. It is possible to note, however, that the time interval between the first and the last contribution of the same user increases approximately in a linear fashion. This means that users tend to maintain a regular frequency between consecutive contributions. In addition, we can observe that most users contribute with less than 20 posts in a week. We can infer

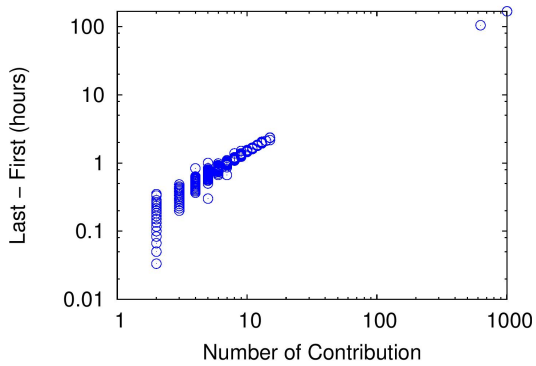


Fig. 8. Relationship of the time interval between the first and last contributions from the same user, and the number of contributions from this user.

that users do not participate on a daily basis, but when they do participate they can make consecutive contributions.

At the present stage of this work, we identified users with too many contributions, around 0.01% of the total. Even though this number of contribution is too big, we do not have enough arguments to consider them as outliers.

IV. Related Work

Previous works have already analyzed the data provided by social networks to model human mobility. The idea is to associate human movements to social ties. Cho et al. [6] use location data from cellphones and online social networks to infer human movement. Authors claim that short-range and periodic trips in space and time do not have clear relationship with social networks, but long-distance trips do. Backstrom et al. [8] investigate the relationship between geographic position and friendship using user-supplied addresses and a network of associations between Facebook members. Authors introduce an algorithm that predicts the location of a person from a sparse set of located users with a performance that exceeds IP-based geo-location. Cranshaw et al. [7] propose a clustering model and a research methodology for studying the structure and composition of a city in large scale, based on the social network. User behavior in vehicular environments has not been tackled, as far as we know, using social networking data. Bouhout et al. [11], for instance, aim to model the driver behavior in the presence of different types of traffic information. To this end, the authors propose a new formal approach to develop a general driving behavior model that can be adapted to each driver.

Our work differs from the others as we analyze the impact of the vehicular networking environment on the reliability of the information available on social networks, more specifically in Waze. Up to now, social networking information has been used as an input of a given investigated problem. We conduct an opposite approach, as we aim to investigate the output of the considered problem at the social network.

V. Conclusions and Future Work

This work is the first to analyze the Waze dataset of Massachusetts. Our main goal is to investigate the posting behavior of users in vehicular social networks and the influence of speed and delay on this behavior. We also analyzed the reliability of users' contributions to the social network, which confirmed the expected human behavior and further derived a number of findings. We found that most contributions to the vehicular social network are triggered during rush hours, mainly on weekdays. On weekends, this trend changes, as users contribute more in the afternoon. Whether on weekdays or weekends, most contributions are of type "Jam", subtype "Heavy Traffic" and "Still Traffic". This shows that users are more motivated to contribute in case of traffic jam issues. Also, we observed that received posts are usually assigned low reliability and, after the first assignment, the reliability does not change. Yet, we found out that users moving at high speeds do not contribute to the vehicular social network and that the contributions with higher delay are usually considered unreliable. Finally, users do not contribute very often in a week time frame and, even when they do, consecutive contributions are triggered with approximately 10 minutes between each other.

As future work, we plan to further investigate the Waze dataset and to try to reveal any hint concerning which event should one pay attention first in a city based on social networking information.

Acknowledgment

The authors would like to thank CNPq, CAPES, and FAPERJ for partially funding this work.

References

[1] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *ACM SIGCOMM*, 2009, pp. 49–62.

[2] Y. Xia and C. K. Yeo, "Mobile internet access over intermittent network connectivity," *Journal of Network and Computer Applications*, vol. 40, pp. 126–138, 2014.

[3] J. He and W. W. Chu, *A social network-based recommender system (SNRS)*. Springer, 2010.

[4] F. Pianese, X. An, F. Kawsar, and H. Ishizuka, "Discovering and predicting user routines by differential analysis of social network traces," in *IEEE WoWMoM*, 2013, pp. 1–9.

[5] A. Mogadala and V. Varma, "Twitter user behavior understanding with mood transition prediction," in *ACM DUBMOD*, 2012, pp. 31–34.

[6] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *ACM SIGKDD*, 2011, pp. 1082–1090.

[7] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh, "The livehoods project: Utilizing social media to understand the dynamics of a city," in *AAAI ICWSM*, 2012, pp. 58–65.

[8] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in *ACM WWW*, 2010, pp. 61–70.

[9] Waze. Waze website - Outsmarting traffic, together. Last accessed 5 July 2016. [Online]. Available: https://www.waze.com/pt-BR/

[10] D. BOSTON. (2015, mar) Waze data from massachusetts state. Last accessed 16 May 2016. [Online]. Available: https://data.cityofboston.gov/Transportation/Waze-Jam-Data/yqgx-2ktq, https://data.cityofboston.gov/dataset/Waze-Point-Data/b38s-xmkq, https://data.cityofboston.gov/Transportation/Waze-Alert-Data/h8aq-6mw7

[11] A. Bouhoute, I. Berrada, and M. El Kamili, "A formal model of human driving behavior in vehicular networks," in *IEEE IWCMC*, 2014, pp. 231–236.