Mapping Social Media for Transportation Studies

Emmanouil Chaniotakis[*], Constantinos Antoniou[**], and Francisco Pereira[***]

[*]Technical University of Munich
[**]Technical University of Munich
[***]Technical University of Denmark

November/December 2016

1 Introduction

During the past decade, social media (SM) has emerged as a prominent trend in social communication. Online platforms such as Facebook and Twitter conquer the Internet with millions of visitors per day. In April 2015, Alexa.com ranked Facebook and Twitter as the third- and tenth-most visited websites in the world, respectively, in terms of average daily visitors and page views. Those ranks are much more indicative of the overall trend when we consider that half of the remaining Top-10 websites are search engines.

Although the wide use of SM emerged in the 2000s, the idea of a virtual space for opinion exchange and messages is not at all new. In 1979, Tom Truscott and Jim Ellis created Usenet, a virtual space for users to post public messages [1]. Internet Relay Chat debuted in 1988 as a way for friends and strangers to communicate, followed by the widespread use of modern forums for user profiles and discussions. According to Danah Boyd and Nicole Ellison [2] SM websites first started as a space for communication, user profiles, and updates with SixDegrees.com in 1997. Later, many less-than-successful attempts were made (such as Friendster and Cyworld) until the deployment of Myspace and Facebook, which initiated an exponential increase of SM users.

These days, SM usage generates an astonishing amount of information (and consequently, data). Every minute, almost 250,000 tweets are posted on Twitter, almost 300,000 Facebook statuses are updated, and about 136,000 photos are uploaded to Facebook. Usage over the past 10 years illustrates that the number of active SM users is increasing (www.statista.com), with 74 percent of online adults using SM sites[1]. Interestingly, most users use their smartphones for accessing SM applications (88% of Facebook, 83% of Twitter users on mobile).

The emergence of SM in conjunction with the availability of APIs for accessing data has prompted researchers to study SM and the applications that data have in various fields. Although much research has been done in transportation, the literature is fragmented and mostly based on data availability rather than data requirements that can support research hypotheses. This is likely a consequence of both the difficulty in acquiring the data and the quickly changing landscape of SM data availability and services offered to users. In this article, we attempt to map the landscape of SM in transportation by identifying the main streams of research; analyzing the taxonomy of the dominant SM; presenting a strengths, weaknesses, opportunities, and threats (SWOT) analysis; and examining the applicability of data collected from SM sources for transportation.

[1]Source: http://www.pewresearch.org

2 Social Media and Transportation

The emergence of SM has led to many definitions that attempt to capture the diverse services offered. Here, we adopt the most generic one, from Andreas Kaplan and Michael Haenlein, who define SM as "a group of Internet-based applications that are built on the ideological and technological foundations of Web 2.0 and allow the creation and exchange of user-generated content." [1]. Jim Sterne [3] has proposed categories that confine the different SM, including forums and messages boards, review and opinion sites, social networks, blogging, microblogging, bookmarking, and media sharing.

SM-based research has been conducted in many scientific disciplines, such as social sciences, economics, politics, and tourism. The most commonly exploited SM-originated information for transportation is based on the use of the spatial information accompanying posts (that is, geotags) and the language processing of posted content. SM has been exploited for its continuous streaming of information, used to identify disruptions or special events, and for forecasting. Transport service providers are using SM to directly communicate with customers, which differs from the use of cached/downloaded users data from SM. The mind map in Figure 1 presents a schematic depiction of the identified uses of SM in transportation research and practice.

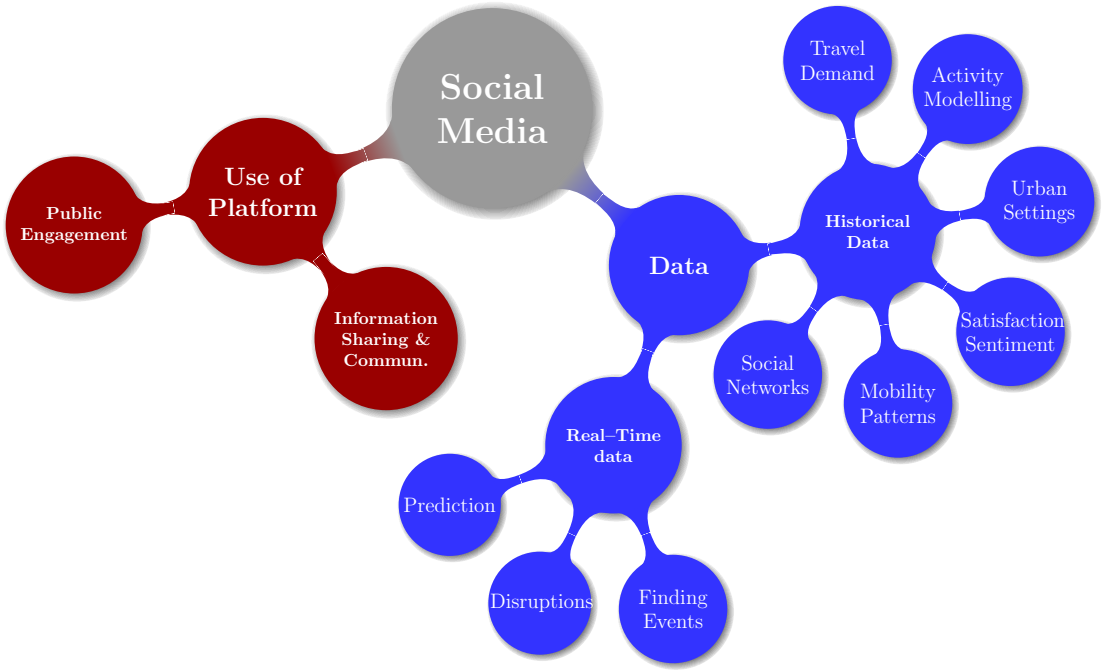

Figure 1: Mind–Map of Social Media uses in Transportation

The merits of real-time data collection are mainly in the identification of events, either in the form of social events, system disruptions, and traffic conditions, [4,5] or in the prediction of locations to be visited by individuals. Several methodologies have been developed that aim to identify those events, in some cases used by traffic centers to get information [6]. For example, Wanichayapong et al. developed a methodology to extract and classify traffic information from SM [7], Kumar et al. used Twitter to detect road hazards [8] and Axel Schulz et al. used semantic analysis to identify small road incidents [9]. The core of those methodologies is the definition and training of classifiers that can recognize events.

On the other hand, researchers have explored historical time datas use for modeling applications and as supporting actions for gaining theoretical insights for the transportation system. The main applications examined can be distinguished based on the purpose of using SM-originated data. Common methods and applications include the following goals: Common

methods and applications are found in the identification of Spatial and Temporal mobility patterns, the investigation of the applicability of the SM-originated data for travel demand modelling, the identification of user activities, the definition of urban settings and related characteristics (such as Points of Interest (POIs), Urban Boundaries, Land uses), the investigation of riders satisfaction and the examination of the relationship between Social Networks and mobility.

In the field of transportation research, SM-originated data can be used to identify the movement of population [10], define the boundaries of cities [11], derive Origin Destination matrices [12], investigate users' mobility patterns [13] and investigate users' social networks and the effect they have on transportation-related behaviour [14].

The use of SM platforms for information sharing and communication forms a two-way communication channel between transportation operators and individuals. Operators use SM to announce updates and information on their websites, advise users of service disruptions, handle complaints and travel questions, respond to questions about SM, and share seasonal messages for goodwill. Users use SM to communicate with transportation operators, share their opinions about a transportation service, report an incident, and ask transportation-related questions [15]. The reasons to use SM from operators were studied by Bregman [16], who performed a survey and found that the goals of using SM can be summarised in communication with current riders, improvement of customer satisfaction and improvement of agency image. In some studies, SM were used as a means towards public engagement and governance [17].

3 Transportation oriented Social Media Taxonomy

The ability to extract information concerning the transportation system from SM is influenced by two factors: the functionalities and focus of the SM and the data availability. Concerning the functionalities and focus of SM, the leading companies are built on market strategies shaped to fulfill different needs of their costumers. Smith[2], Webb[3] and Butterfield blogged their ideas on the needs fulfilled by each SM, developing in an essence a mapping of their market strategy pursued within a 7 blocks honeycomb functionality framework. In this article, we focus on each functionalitys practical implications for transportation research. On data availability, indicators – such as the researchers ability to collect or acquire data, the availability of georeferenced locations, and social networkrelated or textual information – provide a detailed mapping of the capabilities offered.

3.1 SM functionalities and Transportation

We examined this frameworks building blocks in the context of transportation research. The description is based on the work of Jan Kietzmann et al. [18]. These seven building blocks (presence, sharing, relationships, identity, interactions, groups, and reputation) indicate why individuals use each SM.

Presence This describes the functionality that tells another user when someone is accessible and where that individual is located. In the virtual world, presence lets other users know that the individuals in question are online, whereas in the real world it lets other users know the location of their friends (or others if they share information in public). This functionality, and especially its reflection on the real world, is interesting in transportation research because it lets researchers analyze traces from individuals to study mobility and activity patterns.

Sharing This refers to the extent that a SM lets individuals share content. Depending on the platform, content can be almost anything, from documents, pictures, and videos to executable files and compressed folders. Some platforms (such as Flickr, Instagram, Myspace,

[2] http://nform.com
[3] http://interconnected.org/home/

and YouTube) are built around a particular type of content, whereas others (such as Facebook and Twitter) embed content within other functionalities. To the best of our knowledge, transportation studies have not explored SM content sharing, although it could allow for valuable information on the individual identity and also on some of the personality characteristics. However, such an endeavor would require the use of advanced image and video processing algorithms and would raise privacy issues

Relationships SM platforms commonly allow for the definition of social networks. Some let users define a group of acquaintances or individuals they generally want to connect with. Others use the strength of ties by letting users define groups of family, friends, and coworkers and then choose which content to share with those different groups. The inclusion of relationships as a key functionality provided by SM platforms allows for the exploration of online social network structures and interactions, which are factors that affect mobility.1 [19, 20].

Identity This describes the extent to which SM allow and require users to reveal their true identity by including information, such as their real name, age, gender, education, place of birth, and profession. Identity can also be perceived as the act of disclosing thoughts and feelings that describe an individuals preferences [1]. Some SM (such as Facebook and LinkedIn) allow the creation of profile pages wherein individuals can describe themselves. Identity characteristics are particularly interesting for transportation research, for the identification of SM user samples for modeling purposes. Most studies on SM for transportation do not account for this aspect of SM, and instead present their work on the SM sample space from which it is not always clear how to depict the generalization to the population.

Interactions These let users communicate (via messages, pokes, and posts) with people in their SM social network and with strangers (in some cases, from groups defined within each SM platform). SM conversations can be performed in public or with private messages. Depending on the nature of the SM posting process and other functionalities, conversations can vary significantly from one SM platform to another. For example, Facebook allows lengthy posts and comments that can facilitate an exchange of political views, whereas Twitter allows only 140 characters (including links for sharing content) and is most commonly used for microblogging. The interactions that take place among individuals in SM platforms and their tone, form, and content can reveal each individuals social network and, consequently, can allow transportation researchers to get insights in the social network and mobility relation. They also allow for the identification of identity and personality characteristics that have been found to be factors affecting mobility.

Groups In many cases, SM platforms let users form communities and groups to exchange information with the community members privately or publicly. User management tools let group managers maintain a community that might be open to everyone, invitation based, or based on an approval process.

Reputation In SM platforms, reputation can be evaluated on two levels: the individual and the content posted. In many cases, SM platforms provide statistics on reputation used to provide suggestions to users –for example, the number of likes a picture receives or the number of friends or followers a user can indicate the individual's reputation. The reputation indicators that a SM platform provides can be indicative of the SM use and the users real-life profile, adding information on the personal characteristics of users (identity).

Figure 2 presents the categorization of SM for the 10 dominant SM platforms. Darker colors indicate a stronger emphasis paid by SM providers to a specific functionality.

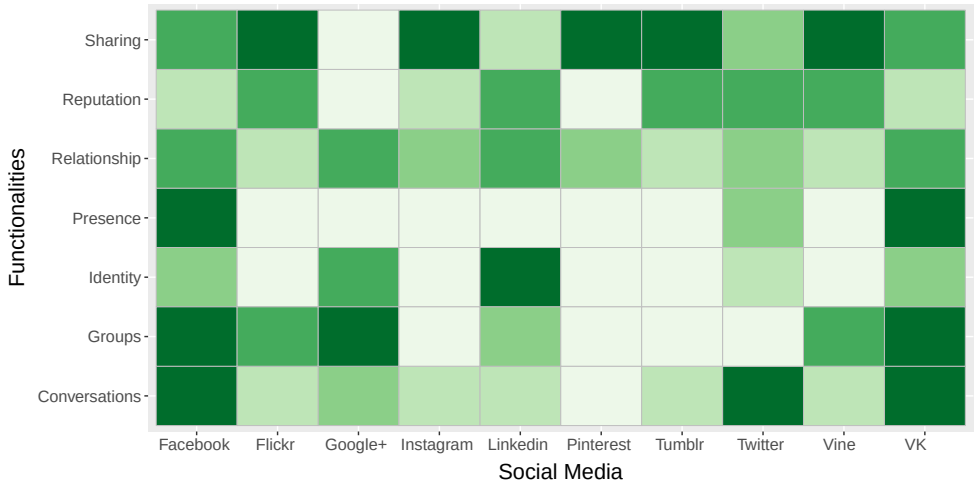

Figure 2: Social Media Functionalities summary table

3.2 Data Availability

The second factor to consider when designing a strategy to use SM-originated data is the datas availability. Data availability is defined as the ability to access sample data from an SM platform and use the data for research. A collection of indicators is specified to understand the data typology and the current limitations of SM platforms. Most SM platforms have created APIs through which researchers can access the data. However, the degree to which this is possible, and the type of information that someone can use, can differ significantly from one SM platform to another and is also subject to terms of use that can change without prior notice.

Although most of the research performed includes the collection of data from APIs, companies owning SM platforms in some cases share datasets for research. Chaniotakis and Antoniou presented a generic methodology for collecting SM data from the available APIs [21]. They illustrated that in some cases a public API can allow for collection of random data, whereas in other cases, researchers must build an application and ask users to join and give privileges to the application in order to collect data from those specific users. Each SM platform generally has supporting libraries (such as Twitter4J (http://twitter4j.org) and Facebook4J) in a wide range of computer languages (for example, Java, Python, and R) that allows for data collection and the use of available data.

Figure 3 presents the data availability indicators for the 10 dominant SM platforms. Darker colors indicate more possibilities to extract the indicated type of information or perform indicated actions.

4 The future of Social Media Research

In light of the two critical factors that define the exploitation potential of SM in transportation studies, we focus in this section on the SWOT of using SM data in transportation studies (see Figure 4). This analysis is based on findings from the pertinent scientific literature [1,18]) and online blogs that discuss SM-related matters, such as Buffer Social (www.blog.bufferapp.com), Jenns Trends (www.jennstrends.com), and Socially Sorted (www.sociallysorted.com.au). Starting with strengths, SM offers an opportunity to obtain a combination of user-generated textual, temporal, and spatial information (user-generated content and information provision) that does not rely on user recollection, while also expressing the current state of mind at generation. The activity-oriented use, within a scope of rich data provision, strengthens researchers ability to investigate the activity space, taking into account the social network dimension. Furthermore, the intense streaming of information allows for dynamic approaches on an unprecedented

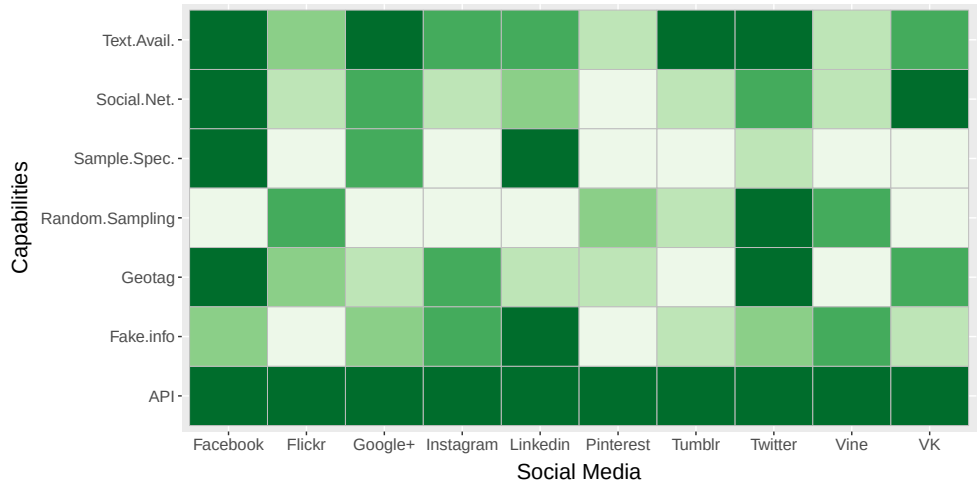

Figure 3: Social Media Data Availability summary table, darker indicated higher scoring on data availability

amount of data produced. Finally, the cost of collecting data and performing research can be significantly reduced while ensuring a comparatively large sample size.

Numerous opportunities arise: researchers can increase their knowledge on transportation-related behavior, supplement or redesign explanatory and prediction models, and use SM data for transportation management. Finally, SM could provide a space for tailored surveys that would require less time to be completed and would be accompanied with users characteristics.

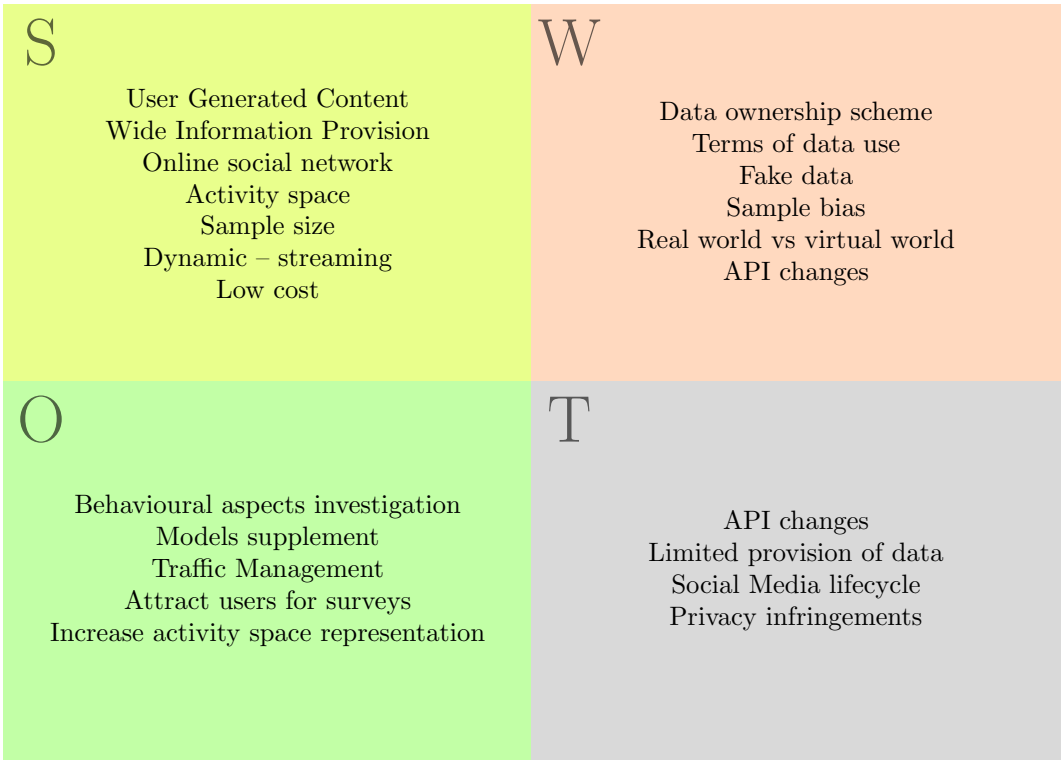

Figure 4: SWOT Analysis

However, the use of SM data for research could have disastrous weaknesses. For example,

the data ownership scheme may contradict conventional expectations that data belongs to public entities. SM providers generally support research, but data collection, storage, and use is a gray area, and data exchange is prohibited. Moving along, SM users post intentionally misleading information (fake data) that degrades the information quality that can be achieved. Additionally, researchers have to deal with high information noise (such as nonsense posts) to acquire useful information. On another level, SM users form a rather misrepresentative sample of the population (sample bias). Researchers should consider the online/offline life differences and the ego bias that user-generated content bears; often, users can have a second (online) life that is different than their real life. Finally, SM providers in many cases redesign the API or the data it provides. This could result in quickly outdated research contributions that might be based on API data provision.

The main threats identified express issues of privacy, SM usage, data provision that might limit the use of SM, or make it inaccessible for research. SM emerged at a rather dramatic pace in a very limited time. We cannot yet be sure that they are going to be able to form a representative sample of the entire population or even if its use will reduced or abandoned. Furthermore, although research is considered as a neutral cause of collection, storing and processing data, it is not certain that SM would continue data provision as it might be identified by many as an invasion of privacy.

5 Empirical Analysis

In this section we present findings from the data collection performed for the city of Thessaloniki, Greece, in order to practically illustrate the merits and the context of using social media in Transportation studies. The data used is collected through the public APIs from Facebook, Foursquare and Twitter. Facebook, Foursquare provided information about specific venues within an area, defined by collecting the number of check-ins measured per collection period. Twitter provides information about individual users who post geolocated tweets. The data collection period has been defined as three months for Facebook and Foursquare and one year for Twitter. The spatial and temporal distribution of SM data was examined and compared to traditional travel surveys performed in 2014 for the city of Thessaloniki, Greece.

On the temporal dimension, the average percentile distribution of within day variations for each data source is presented in Figure 5. A clear difference is identified in the percentile distribution of the performed check-ins for all social media on a temporal level, in comparison to the arrival times from the conventional travel survey. As Chaniotakis et al. [23] discussed this difference can be attributed to the primary use of social media which is linked to recreational or social activities.

On a spatial level, locations characterized by high SM activity are found to be characterized by recreational land uses. This is evidenced in data from three SM platforms. On the other hand, the conventional data collected illustrates a more evenly distributed concentration of attractions with a low amount of fluctuations. This tendency of SM use at locations characterized by recreational land uses is further supported by the work by Chaniotakis et al. [23] who found that for the case of Athens, Greece, there is a high spatial correlation of the points identifying areas of interest and the positioning of recreational Points of Interest (POIs).

6 Conclusions

Social media can be a game-changer in many research fields, including transportation. Its characteristics –such as the continuous stream of information; user-generated content; combination of temporal, spatial, and textual information; and existence of a social network representation– have created a stream of research that focuses on the exploitation of the information that can directly be extracted. However, the study of SM requires the connection of users online and offline worlds so that their identified behavior and patterns can be generalized and delineated. Finally, empirical evidence suggests that SM cannot displace existing survey methods used,

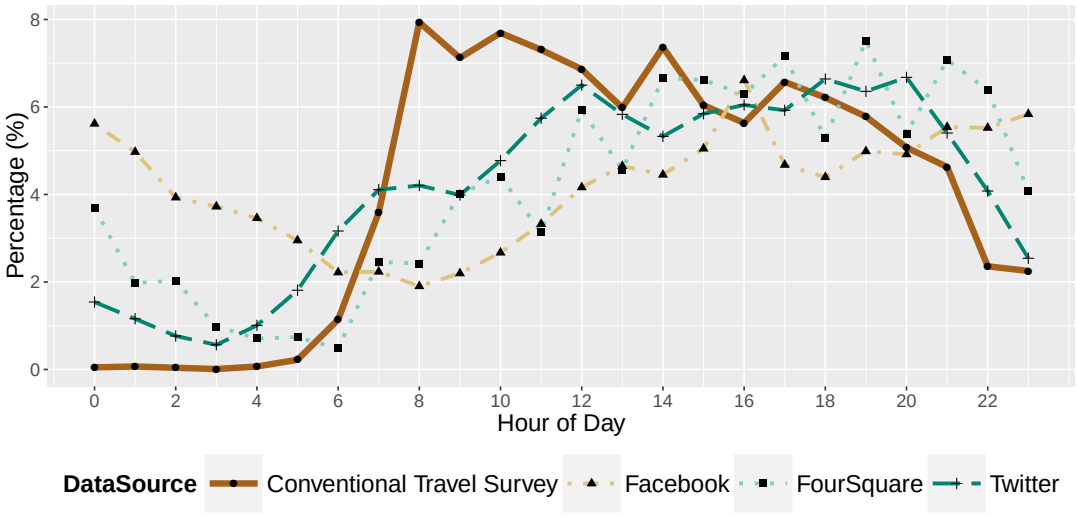

Figure 5: Percentile Distribution of within a day variations for Facebook, Twitter, Foursquare and the arrivals of the Conventional Travel Survey

with the most applicable area of research being the fusion of various data sources based on identified misrepresentations and commonly known issues.

Acknowledgements

We thank the TU1305 COST action participants for the valuable discussion and comments towards the use of Social Media in Transportation, and to thank the Centre for Research and Technology Hellas – Hellenic Institute of Transportation (CERTH-HIT) for support on the Thessaloniki case.

References

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, jan 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0007681309001232

[2] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[3] J. Sterne, *Social media metrics: How to measure and optimize your marketing investment.* John Wiley & Sons, 2010.

[4] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Using Data From the Web to Predict Public Transport Arrivals Under Special Events Scenarios," *Journal of Intelligent Transportation Systems*, no. June 2014, pp. 1–16, 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/15472450.2013.868284

[5] Z. Kokkinogenis, J. Filguieras, S. Carvalho, L. Sarmento, and R. J. Rossetti, "Chapter 12 - mobility network evaluation in the user perspective: Real-time sensing of traffic information in twitter messages," in *Advances in Artificial Transportation Systems and Simulation*, R. J. Rossetti and R. Liu, Eds. Boston: Academic Press, 2015, pp. 219 – 234. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123970411000121

[6] H. Nguyen, W. Liu, P. Rivera, and F. Chen, *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part I*. Cham: Springer International Publishing, 2016, ch. TrafficWatch: Real-Time Traffic Incident Detection and Monitoring Using Social Media, pp. 540–551. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-31753-3_43

[7] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-atikom, and P. Chaovalit, "Social-based traffic information extraction and classification," in *ITS Telecommunications (ITST), 2011 11th International Conference on*, Aug 2011, pp. 107–112.

[8] A. Kumar, M. Jiang, and Y. Fang, "Where not to go?: detecting road hazards using twitter," *Proceedings of the 37th international ACM ...*, vol. 2609550, pp. 1223–1226, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2609550

[9] A. Schulz, P. Ristoski, and H. Paulheim, "I see a car crash: Real-time detection of small scale incidents in microblogs," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7955 LNCS, pp. 22–33, 2013.

[10] Y. Liu, Z. Sui, C. Kang, and Y. Gao, "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data," *PloS one*, vol. 9, no. 1, p. e86026, 2014.

[11] B. Jiang and Y. Miao, "The Evolution of Natural Cities from the Perspective of Location-Based Social Media," *ArXiv e-prints*, Jan. 2014.

[12] F. Yang, P. J. Jin, X. Wan, R. Li, and B. Ran, "Dynamic origin-destination travel demand estimation using location based social networking data," in *Transportation Research Board 93rd Annual Meeting*, no. 14-5509, 2014.

[13] F. Rebelo, C. Soares, and R. J. Rossetti, "Twitterjam: Identification of mobility patterns in urban centers based on tweets," in *Smart Cities Conference (ISC2), 2015 IEEE First International*. IEEE, 2015, pp. 1–6.

[14] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and Mobility: User Movement in Location-based Social Networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1082–1090. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020579

[15] A. Gal-Tzur, S. M. Grant-Muller, E. Minkov, and S. Nocera, "The Impact of Social Media Usage on Transport Policy: Issues, Challenges and Recommendations," *Procedia - Social and Behavioral Sciences*, vol. 111, pp. 937–946, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877042814001293

[16] S. Bregman, *Uses of social media in public transportation*. Transportation Research Board, 2012, vol. 99.

[17] E. Bonsón, L. Torres, S. Royo, and F. Flores, "Local e-government 2.0: Social media and corporate transparency in municipalities," *Government Information Quarterly*, vol. 29, no. 2, pp. 123 – 132, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0740624X1200010X

[18] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," *Business Horizons*, vol. 54, no. 3, pp. 241–251, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.bushor.2011.01.005

[19] T. Arentze and H. Timmermans, "Social networks, social interactions, and activity-travel behavior: a framework for microsimulation," *Environment and Planning B: Planning and Design*, vol. 35, no. 6, pp. 1012–1027, 2008.

[20] K. W. Axhausen, "Social networks, mobility biographies, and travel: survey challenges," *Environment and planning. B, Planning & design*, vol. 35, no. 6, p. 981, 2008.

[21] E. Chaniotakis and C. Antoniou, "Use of geotagged social media in urban settings: Empirical evidence on its potential from twitter," in *IEEE 18th International Conference on Intelligent Transportation Systems, (ITSC)*. IEEE, 2015, pp. 214–219.

[22] Y. Yamamoto, "Java library for the twitters," 2007. [Online]. Available: http://www.twitter4j.org/

[23] E. Chaniotakis, C. Antoniou, and E. Mitsakis, "Data for leisure travel demand from social networking services," in *4th hEART Symposium*. European Association for Research in Transportation - hEART, DTU, Denmark, 2015.