Using Data from Location Based Social Networks in Human Mobility Analysis

Daniel Marsh-Hunn
Westfälische Willhelms Universität
Münster, Germany
d_mars02@uni-muenster.de

ABSTRACT
Location Based Social Networks (LBSN) have experienced a large growth in their communities in recent decades and online presence in LBSNs has become the norm for citizens. Data created on LBSNs has proved to be a valuable data source for studying human mobility patterns and location prediction. While there are various types of LSBNs which handle and exploit users' locations differently, data quantities are usually large, allowing research to be done in unprecedented spatial and temporal dimensions. This review provides insight into the potential of crowdsourced LBSN data by presenting a selection of applied studies done in the field of human mobility analysis. Although some challenges must be faced, and limitations considered, the outlook for the usage of LBSN data shows great potential in contributing new knowledge in this field of study.

Author Keywords
Location Based Social Networks (LBSN); Volunteered Geographical Information (VGI); urban environment; urban planning; Location Based Services (LBS)

INTRODUCTION

Models of human mobility patterns have proven to be useful in a variety of application fields. They are commonly used for forecasting and can be applied to domains such as traffic, transportation networking, human activities, epidemiology and mobile network simulation. Traditionally, data used for mobility analysis originated from travel surveys questionnaires and interviews. Although these data may contain richer information including more details, this approach limits the analysis' temporal resolution and included data quantity profoundly [1]. More recent approaches of movement data acquisition have been the inclusion of cell phone data, smart card transactions in public transport networks and the tracking of bank notes [19].

The spread of internet connectivity and the high availability of devices featuring GPS has facilitated an increased production of spatial data and enabled a stronger engagement of citizens with everything involving location [3]. In the emerging concept of Web 2.0, also known as the Social Web, large numbers of users influence central websites by interacting with them and providing information [4]. This phenomenon in the domain of geographical data has been defined as volunteered geographic information (VGI) [5] or crowdsourced geographic information. Services like OpenStreetMap and WikiMapia have shown how powerful crowdsourced data can be as a complementary or standalone data source.

The use of Web-based Social Networks like Facebook and Twitter have become the norm in citizen's lives. Twitter reports a user community increase from 30 to 335 million users from 2010 to 2018 [2]. Many social networks can be configured to use a device's location, and consequently, Location-Based Social Networks (LBSN) like Foursquare or Gowalla have become very popular. These services allow users to "check in" at locations of interest, uploading their location to the web for other users to see [7]. Other features allow users to geo-tag uploaded messages, videos or audios. In the US almost three quarters of smartphone users use location-based directions and information on their phones and almost 20% use geo-social services [6]. The resulting unprecedented amount of geo-tagged data offers a new dimension of information related to human activity and has enabled many novel applications in a variety of fields such as location recommendation systems, travel demand modelling, epidemiology and urban planning [8].

This review aims at providing a summary of analysis methods using crowdsourced data from location-based social networks. In an initial section the types LBSNs will be introduced, naming the most influential ones and including what kinds of geographical data is produced by them. The main section of the paper provides a set of examples how LBSN data can be applied to human mobility analysis. In the final section the topic is discussed, focusing on opportunities and threats, and ends with a conclusion on the topic.

LOCATION-BASED SOCIAL NETWORKS

Enhancements in location-acquisition technology like GPS and Wi-Fi has enabled people to add a location dimension to online social networks in a variety of ways. Zheng (2010) [9] defines three categories of LBSNs based on the role location plays: *Geo-tagged-media-based, Point-location-driven* and *Trajectory-centric.*

- *Geo-tagged-media-based*: This category comprises LBSNs that provide services to add a location label to media content like text or photos and videos generated in the physical world, which other users can comment on. These LBSNs add or enrich information from media content, which is still the focus rather than the location.

Representatives of this category are Flickr, Panoramio, Twitter and Facebook.

- *Point-location-driven*: dedicated LBSNs like Foursquare, a location of interest recommendation platform, rely on location data to serve their purpose. Foursquare's popularity derives from the advantages of focusing on location-based services, hence users are not distracted by other communication [10]. "Check-ins" at venues of interest inform other users about the user's current location and can be commented or "liked". "Check-ins" earn badges and points for the user, allowing them to compete with friends in a gamified environment. Gowalla follows the principle of Foursquare on location recommendation, but includes more gamification, like virtual item collection and exchange. A further point-driven LBSN is Tinder, which relies on the user's location to locate other people searching for romantic relationships [11].
- *Trajectory-centric*: These LBSNs allow users to share routes, meaning a sequential connection of point locations. This application is frequently used in the health and fitness sector. People record and share their jogging or biking trajectories and can add further information about the routes, like comments, photos and opinions. This category includes services like Strava, Garmin Connect, GeoLife and Bikely [9].

LBSN applications from different categories generate distinct types of datasets, which in return can be used for different kinds of spatial analysis.

HUMAN MOBILITY ANALYSIS APPLICATIONS USING LBSN DATA

The quantity and quality of data generated by current LBSNs bring with them a vast range of analysis approaches for human mobility and movement behaviour. Long-term human mobility research in the past was mostly conducted by doing costly and time-consuming surveys on a comparatively small number of test individuals, showing strong relations of mobility patterns with land-use and city environments and demonstrating great regularity in daily travel patterns. More recent studies show more interest in location-based data from the web, which often contains more information about the relations between places and people and therefore is effective data for analysis [8].

An example for LBSN data usage for human mobility research is shown in a study done by Hasan et al (2013) [8]. The researchers use a large dataset of **Foursquare** check-ins from New York city tweeted on **Twitter**, which were obtained from Twitter's gardenhouse streaming API (~1% of the entire Twitter public timeline). A total of 3256 users leaving 504,000 check-ins are included in the study. The tweets have added coordinates, timestamp and a weblink to the affiliated Foursquare venue. Activity purposes for tweets could be identified since Foursquare places all registered venues into categories. On a map grid of 200x200 m per cell,

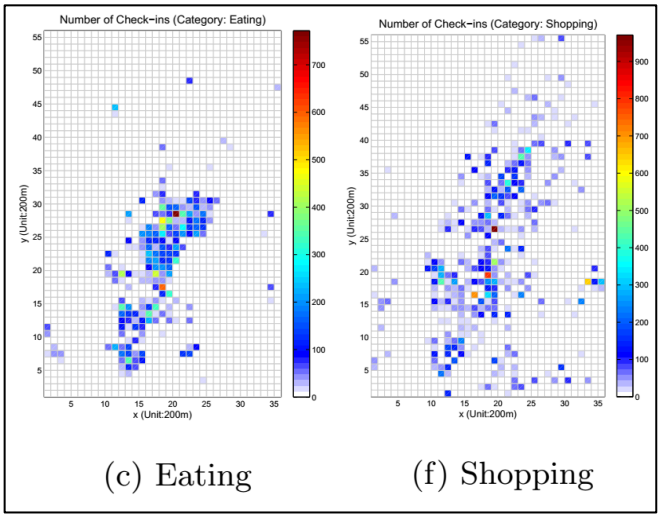

Figure 1: ***Check-in density for Eating (c) and Shopping (f)*** **[8]**

cells were ranked according to the number of check-ins, rank 1 being the cell with the highest number of check-ins, which is considered a measure of cell popularity. From data it emerged that the probability of visiting cells decreases with higher cell ranks in a truncated power law distribution, since popular places are much more likely to attract new and repeated visitors. Furthermore, different categories show different spatial distributions. *Figure 1* shows that shopping check-ins are scattered over a large area, while eating check-ins are concentrated in certain areas of the map. Investigating temporal mobility patterns of the dataset, it emerged that distinct activity purposes have varied daily and weekly concentrations, depending on the nature of the category (e.g. shopping occurs more on weekends, entertainment activities occur peak around late night time etc.). A limitation of check-in data discovered in this research was that certain categories of visits are more likely to receive check-ins than others (e.g. home and work-related check-ins are less frequent than entertainment and recreation), resulting in varying quantity of data and possible accuracy disparities in the analysis. The overall result of the study demonstrates how the additional activity category information provides richer insights into human mobility research.

The **Twitter** Streaming API can not only be used in combination with Foursquare check-ins. Between June 2013 and March 2014 geo-located tweets were captured for a study done in Kenya by Blanford et al (2015) [15], investigating human movement patterns across political borders. A total of over 720,000 tweets from over 28,000 users were collected. For each tweet coordinates, timestamp and user ID were extracted for analysis. While mobile phone data depends on phone providers, which do not necessarily extend beyond country boundaries, Twitter data uses unique user identification globally, and therefore can capture cross-border movement. The researchers argue that, although Twitter data may be limited by people's access to technology

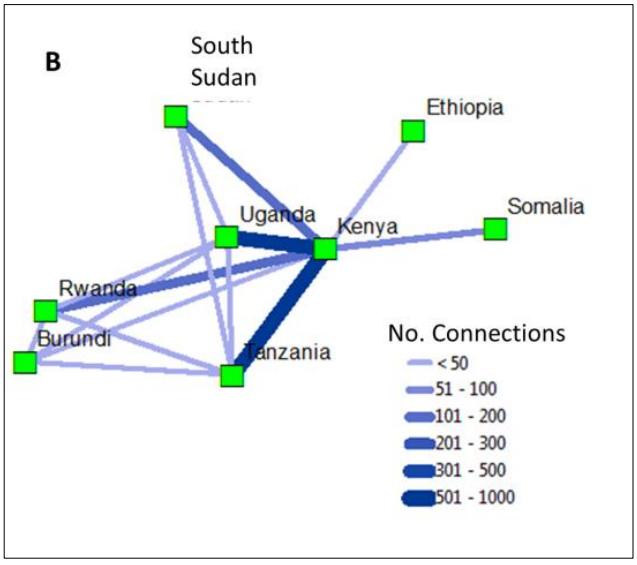

Figure 2*: Connections between Kenya and the surrounding countries based on Twitter activity* **[15]**

for network connectivity, they can still serve as a "good enough" proxy data source for mobility in regions where data is scarce. The results of the study show human movement streams between Kenya and surrounding countries (*Figure 2*). More importantly, travel hubs could be identified from the data analysis, which can bring further insight into the spread of diseases and epidemics, which at the time of the study was a heavily researched topic due to the Ebola outbreak in 2014. The authors also see great potential in the usage of Twitter data for determining causes of human displacement at different spatial and temporal scales. A further application field proposed by the authors are using Twitter data as a complement to cell-phone-based data for population distribution estimates.

In an example of exploiting point-location-driven LBSN data, a study undergone at Arizona State University by Gao et al (2013) [14] demonstrate how publicly available **Brightkite** and **Foursquare** datasets can be used to create predictive models about individual users' movement. The datasets include 2,000,000 and 4,600,000 venue check-ins by 18,000 and 27,000 users, respectively. The researchers propose a general framework for modelling temporal cyclic patterns and their relationships with spatial and social data. Modelling the temporal effects of user mobility is done considering the spatial preferences and the temporal correlations. An advantage of the resulting modelling framework is that these two aspects can be included into various prediction algorithms from various model combinations. Temporal cyclic patterns can affect mobility behaviour and can be modelled as a Gaussian mixture distribution, capturing the user's temporal preferences more accurately. Results show that the temporal-spatial correlation models produced in the study outperform corresponding spatial only models, providing complementary information to spatial context and improving prediction accuracy. A further observation showed correlation between users' temporal preferences and their friends on the social network, indicating preference overlapping.

Twitter may also be a crucial data provider in the field of human mobility in disaster management. Research conducted in 2015 in Tokyo, Japan suggests using geo-tagged tweets to identify and visualise paths of high evacuation risk in the event of a disaster [16]. In the event of a disaster crowded areas strongly affect people's behaviour. For this study tweets were extracted in crowded areas every hour and then mapped on a two-dimensional grid with a resolution of 500 meters. Tweets were calculated for each cell and normalised using a Gaussian filter. After filtering, cells are selected when tweets including photographs exceed a pre-defined threshold, classified as "crowded areas". Using OpenStreetMap network data, multiple paths to evacuation sites are then calculated, considering the number of people that must be evacuated from crowded areas. The evacuation paths are then assessed considering possible traffic congestion and the Emergency Response Difficulty Assessment (ERDA) degree, which includes factors such as fire risk, road width and building stability. The assessment results show the highest risk evacuation paths, which can be taken into consideration in the case of an emergency. This methodology can visualise how people's evacuation paths change according to the time of the day in a fin-grained manner. In a final remark the authors argue that their system is effective, but they would like to include models based on estimating the number of evacuees for different evacuation sites.

Geo-tagged **Twitter** message analysis is taken even further by Steiger et al (2011) [20]. In their study the authors develop a methodology framework for the identification of human spatial behaviour and of underlying mobility clusters within mass events. The framework processes the tweet text and breaks it down into single words, then selects words that occur more frequently. Tweets are then compared spatially using local spatial autocorrelation, temporally by tweet timestamp and semantically by using Latent Dirichlet Allocation (LDA). Results show successful identification of mobility flows towards event venues. The authors state they see great potential for this methodology applied during mass events in a (near)-real time manner, and especially to events where official data for the event is scarce. The results also imply some limitations though, especially concerning the accuracy of LDA and the heavy reliance on Twitter, assuming that there are enough in-situ tweets to perform an analysis with good accuracy.

In a further case study for human mobility analysis and LBSN data Sun et al (2017) [12] use a large **Strava Metro** dataset for Glasgow (UK) urban area to determine whether commuting cyclists and cyclists riding for recreational and other purposes are exposed to different levels of air pollution. Strava is a very popular trajectory-centric social network for

runners and cyclists and hosts an extremely large user base. Strava records distance, time, average speed and the GPS trajectory of each activity. More importantly for this study, users can add a "commute" flag to their tracks, indicating journeys to and from work. Strava Metro is a suite of data services aiming to produce high quality spatial data products. The dataset included in this study consists of 50,057 cycling activities. In a spatial context, the data consists of edges (streets) and nodes (street intersections), with the Strava attributes added to the nodes (e.g. a node will count the total number of "non-commuting" activities). Taking the "commute" flags into consideration, clusters of high non-commuting rate were calculated. These were then compared to air pollution data (PM10, PM2.5). At node level, independent variables influencing cyclist behaviour were identified, mainly represented by distances to certain features (e.g. water bodies, green spaces, bus stops, etc.). Results of the study show that non-commuting activities are more likely to happen in the outskirts of the city. Moreover, non-commuters are spatially less likely to be exposed to high levels of air pollution. Strava Metro data have their limitations though, the main one in this study being the small temporal resolution of the dataset of one year. To research the impact of air pollution on a cyclist's health, the dataset would have to include a larger time frame. Another issue with Strava Metro datasets is that they don't include individual-level trips, neither including the trip duration nor counting the number of trips cyclists make.

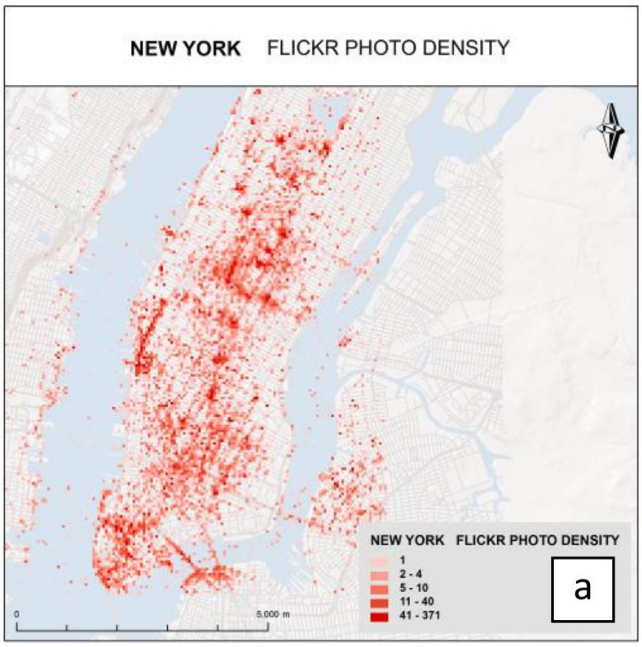

Figure 3: ***Flickr photo density in New York (USA)*** **[13]**

A further LBSN providing useful data for research human mobility patterns is **Flickr**, a geo-tagged photo sharing platform. A study on collective human behaviour patterns by Sagl et al (2012) [13] use photo coordinates and timestamps to identify spatial-temporal tourist mobility in multiple example cities, but also to identify seasonal variations in touristic activities in rural and urban areas. *Figure 3* shows the original Flickr point data in New York City (USA) in five density categories. This research puts various applications of crowdsourced LBSN data applications on display, but also argues that, taking Flickr data as an example, LBSN data often is poor on information, providing little complementary information about uploaded feature.

DISCUSSION

The previous section provided a general insight into applications of LBSN data for human mobility analysis purposes. Although most research produced useful results, the examples also shed light on some limitations of crowdsourced social media data, which always should be kept in mind when doing future research. A major limitation to the LBSN approach to mobility analysis is the varying accessibility to technology [15]. LBSN data analysis can only be conducted if there is a user community for the social network in question. This limits the approach geographically to areas with the necessary prerequisites, such as network connectivity, phone availability and public activity in social networks. Fortunately, these problems are fading as global technology accessibility increases and the user communities on social media grow, but nevertheless they should not be neglected.

Twitter results to be the most frequently used social platform to extract data for mobility analysis. Twitter's streaming API exposes 1% of their tweets, which results in large amounts of geographical point data. These data can be used to perform mobility analysis for various sub topics, showing its effect particularly in social hub identification and human mass quantification. But as mentioned before, analysis can only be performed on members of the social platform community, meaning that exact quantification remains impossible. Furthermore, the quality of the data is low, providing little more than raw coordinates and time stamps. Twitter combined with third party social networks such as Foursquare show more promising results regarding mobility purpose, enriching Twitter data with valuable information. The potential for richer data extraction in social network combination is very promising since it connects individual user locations with specific venues in space. This provides unprecedented, large scale information about human movement linked to specific activities.

One emerging threat to the LBSN data analysis approach may be the growing awareness of the importance of data privacy. With more data privacy scandals hitting the news in recent years, people may become more suspicious and reluctant to upload their location data. On May 25th, 2018, the European Union General Data Privacy Regulation (GDPR) [17] took effect, s forcing companies including social network to inform their users about the usage, storage and accessibility of their private data and granting them more data control. This may cause LBSN users to rethink

uploading their location data even further. While in the past many LBS and LBSN users reported indifference when sharing location data due to hidden location privacy settings or simply because they were unaware about the consequences [18].

CONCLUSION

The field of LBSN data analysis is bound to grow with fast developing web technology and inflating communities on social networks. Previous research has shown there is great potential in transferring LBSN data to knowledge and has great potential to contribute profoundly to human mobility research. Even though crowdsourced data from LBSNs lacks data richness the mere quantity generated daily enables analysis of unprecedented dimensions. The spatial expansion of LBSN data research to a larger variety of urban environments and higher temporal resolution due to more consistent streams of data will enlarge the scale of mobility analysis of crowdsourced data. New types of location-based services and social networks may also have an impact on data quantity and quality in the future.

ACKNOWLEDGMENTS

This review on Location Based Social Network data in Human Mobility research was composed in the scope of the course “Location-Based Services - 2018” held by Prof.. Dr.. Christian Kray at the Institute for Geoinformatics at the University of Münster, Germany.

REFERENCES

1. Hasan, S., Schneider, C. M., Ukkusuri, S. V, & Gonzalez, M. C. (2013). Spatiotemporal Patterns of Urban Human Mobility. *Journal of Statistical Physics*, *151*(1–2), 304–318. https://doi.org/10.1007/s10955-012-0645-0
2. Statista. (2015). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2015 (in millions). The Statistic Portal, 2018, 2018. Retrieved from http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
3. Oliveira, M. G. De, Baptista, C. D. S., Campelo, C. E. C., Moura, J. A., Filho, A., Gabrielle, A., & Falcão, R. (2015). Producing Volunteered Geographic Information from Social Media for LBSN Improvement. *Journal of Information and Data Management*, *6*(1), 81–91.
4. Goodchild, M. F. (2007). Editorial : Citizens as Voluntary Sensors : Spatial Data Infrastructure in the World of Web 2 . 0. *International Journal of Spatial Data Infrastructures Research*, *2*, 24–32. https://doi.org/10.1016/j.jenvrad.2011.12.005
5. Goodchild, M. F. (2007a). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y
6. Zickuhr, K. (2012). Three-quarters of smartphone owners use location-based services their location with friends, 27.
7. Gao, H., Tang, J., & Liu, H. (1987). Exploring Social-Historical Ties on Location-Based Social Networks, 114–121.
8. Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp ’13*, 1. https://doi.org/10.1145/2505821.2505823
9. Zheng, Y. (2010). Location-based social networks: Users. *Computing with Spatial Trajectories*, 243–276. https://doi.org/http://dx.doi.org/10.1007/978-1-4614-1629-6_8
10. Chorley, M. J., Whitaker, R. M., & Allen, S. M. (2015). Personality and location-based social networks. *Computers in Human Behavior*, *46*, 45–56. https://doi.org/10.1016/j.chb.2014.12.038
11. David, S., & Brennan, S. (2015). How Location-based Social Network applications are being used. Retrieved from https://jyx.jyu.fi/bitstream/handle/123456789/45700/1/URN%3ANBN%3Afi%3Ajyu-201504221654.pdf
12. Sun, Y., & Mobasheri, A. (2017). Utilizing crowdsourced data for studies of cycling and air pollution exposure: A case study using strava data. International Journal of Environmental Research and Public Health, 14(3). https://doi.org/10.3390/ijerph14030274
13. Sagl, G., Resch, B., Hawelka, B., & Beinat, E. (2012). From Social Sensor Data to Collective Human Behaviour Patterns – Analysing and Visualising Spatio-Temporal Dynamics in Urban Environments. Jekel, T., Car, A., Strobl, J. & Griesebner, G. (Eds.) (2012): GI_Forum 2012: Geovizualisation, Society and Learning. © Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/Offenbach. ISBN 978-3-87907-521-8., 54–63.
14. Gao, H., Tang, J., Hu, X., & Liu, H. (2013). Modeling temporal effects of human mobile behavior on location-based social networks. Cikm, 1673–1678. https://doi.org/10.1145/2505515.2505616
15. Blanford, J. I., Huang, Z., Savelyev, A., & MacEachren, A. M. (2015). Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. PLoS ONE, 10(6), 1–16. https://doi.org/10.1371/journal.pone.0129202

16. Kanno, M., Ehara, Y., Hirota, M., Yokoyama, S., & Ishikawa, H. (2016). Visualizing High-Risk Paths using Geo-tagged Social Data for Disaster Mitigation. Proceedings of the 9th ACM SIGSPATIAL Workshop on Location-Based Social Networks - LBSN16, 2011, 1–8. https://doi.org/10.1145/3021304.3021308
17. European Union. (2016). Regulation 2016/679 of the European parliament and the Council of the European Union. *Official Journal of the European Communities*, *2014*(October 1995), 1–88. https://doi.org/http://eur-lex.europa.eu/pri/en/oj/dat/2003/l_285/l_28520031101en00330037.pdf
18. Coppens, P., Claeys, L., Veeckman, C., & Pierson, J. (2014). Privacy in location-based social networks : Researching the interrelatedness of scripts and usage. Symposium on Usable Privacy and Security.
19. Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. Nature, 439, 462. Retrieved from http://dx.doi.org/10.1038/nature04292
20. Steiger, E., Ellersiek, T., Resch, B., & Zipf, A. (2011). Uncovering latent mobility patterns from Twitter during mass events. Journal for Geographic Information Science, (Zheng), 525–534. https://doi.org/10.1553/giscience2015s525