Location-Centric Social Media Analytics: Challenges and Opportunities for Smart Cities

Dingqi Yang, Bingqing Qu, and Philippe Cudre-Mauroux

Abstract—With the proliferation of increasingly powerful smartphones, location-centric social media platforms, such as Foursquare, have attracted millions of users sharing their physical activity online, resulting in an invaluable source of fine-grained, semantically rich, spatiotemporal user activity data. Such data provides us with an unprecedented opportunity for analyzing urban dynamics and developing smart city applications. In this article, we first systematically discuss the unique characteristics of location-centric social media data, which consist of four data dimensions, i.e., spatial, temporal, semantic and social dimensions. We then highlight three key challenges relating to data analytics, i.e., data heterogeneity, data quality and privacy. Finally, we discuss the opportunities of leveraging location-centric social media data for urban analytics and smart cities, including both data analytics within and across the four data dimensions, and data fusion with further urban data.

Index Terms—Location-centric social media, Location based social networks, urban analytics, smart city

1 Introduction

The worldwide popularity of smartphones in the past decade has started a revolution in social media services. These GPS-embedded smartphones introduce a novel dimension of location to online social media, which bridges the gap between the virtual online social networks and our physical world. Specifically, user-generated content on social media gets associated with geo-tags, such as GPS coordinates, an address, or even a geographical region. By incorporating geo-tagged media content, so-called Location Based Social Networks (LBSNs) have attracted millions of users. A broad definition of LBSN [1], [2] includes any social media services with geo-tagged media content, such as Twitter[1] for geo-tagged Tweets, Flickr[2] for geo-tagged photos, or Facebook[3] for geo-tagged posts.

In this study, we discuss a particular (and also the most popular) type of LBSNs, namely location-centric social media [3], such as Foursquare[4], Facebook Places[5], the US-based Yelp[6], or the Chinese-based Jiepang[7]. Different from other LBSNs that associate user-generated content with GPS coordinates (e.g., geo-tagged Tweets/photos), location-centric social media puts a particular focus on the location dimension by representing a location as a Point of Interest (POI). A POI typically refers to a specific place that may interest users, such as a bar, a supermarket, or a gym. Such a POI-centric mechanism allows locations to be associated with rich semantic information including name, category, contacting information, opening hour, services, etc. In addition, it also supports a fine-grained location resolution. For example, two different POIs may have the same GPS coordinate, as they are located on different floors but in the same building. Despite the broad definition of LBSNs given by [1], [2], the term "LBSN" mostly refers to such location-centric social media in the current literature [3]. Therefore, we do not distinguish these two terms in the rest of this article.

In a typical LBSN like Foursquare, users can share their real-time presence at POIs with their friends, which are called "check-ins". Moreover, rich media content, including messages, photos, or even videos, can also be associated with a check-in. Due to the successful gamification strategy of LBSNs, millions of users are interacting with each other by sharing their check-ins at POIs, generating a tremendous volume of fine-grained, semantically rich, spatiotemporal user activity data online. Such large-scale user activity data thus become an invaluable source for urban analytics and building smart city applications.

However, exploiting check-in data introduces several challenges. First, the rich media content naturally brings data heterogeneity issues. For example, a user's preference on POIs could be expressed by both check-ins frequency and review sentiment. Second, as a self-reported user activity data source, check-in data from LBSNs has inevitable limitations in terms of data quality. Specifically, as users voluntarily share their presence at POIs with their friends on LBSNs, check-in data is intrinsically a biased sample of the users' daily activity, which may limit its practical applications in some use cases. Third, as check-in data contains fine-grained spatiotemporal information about user activities, it also raises privacy concerns for data analytics. Therefore, it is of critical importance to fully understand the characteristics of check-in data and the challenges of check-in data analytics before exploiting them to develop smart city applications.

In this article, we systematically discuss location-centric

• *Dingqi Yang is with the University of Fribourg, Switzerland and the University of Macau, SAR China. Bingqing Qu and Philippe Cudré-Mauroux are with the University of Fribourg, Switzerland*
E-mail: {Firstname.Lastname}@unifr.ch

Manuscript received April 19, 2005; revised August 26, 2015.

1. https://twitter.com/
2. https://www.flickr.com/
3. https://www.facebook.com/
4. https://foursquare.com/
5. https://www.facebook.com/places/
6. https://www.yelp.com/
7. http://jiepang.com/

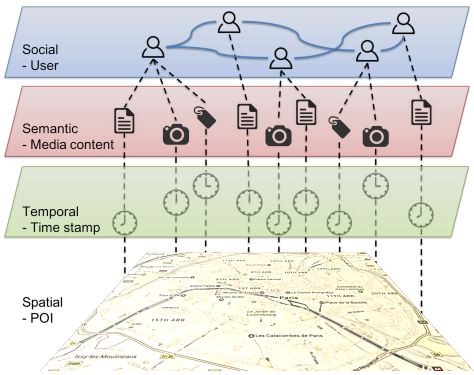
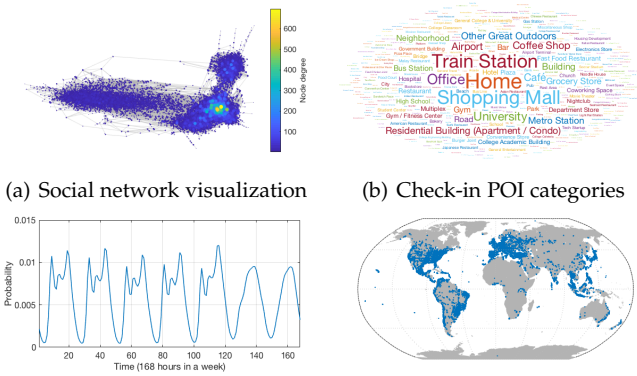


Fig. 1. Four data dimensions of LBSNs



(a) Social network visualization (b) Check-in POI categories

(c) Temporal traffic of check-ins (d) World map of check-in POIs

Fig. 2. Visualization of the collected LBSN data

social media mining for urban analytics in smart cities. To this end, we first discuss the unique characteristics of LBSN data, and then highlight challenges in LBSN data analytics. Finally, we discuss opportunities relating to LBSN data analytics and its applications in smart cities. Different from existing studies focusing either on spatial trajectory analytics [2] or on the board definition of LBSNs [1], this article puts a particular focus on location-centric social media, which comprise fine-grained, semantically rich, spatiotemporal user activity data.

2 Characteristics of LBSNs

There are four data dimensions in a typical LBSN, i.e., spatial, temporal, semantic and social. Each check-in activity can be regarded as a link across these four dimensions. As shown in Fig. 1, a check-in indicates a user's presence at a specific time and at a POI along with certain media content. In this study, we collect and analyze a global-scale LBSN dataset over about two years (from Apr. 2012 to Feb. 2014) from Foursquare[8]. It contains 90,048,627 check-ins from 2,733,324 users on 11,180,160 POIs.

The social dimension consists of users and their social relationships. Each user is associated with a profile including her name, photo, home city, gender, contact, and a short bio, etc. The social links can be established in a unidirectional or bidirectional manner, where Foursquare supports both types. Friendships can be established between users online. Fig. 2(a) visualizes the friendship network with a color bar indicating user node degrees, where we observe a heavy-tailed node degree distribution, implying that only a few nodes with high degrees.

The semantic dimension consists of different media content that can be associated with a check-in, such as messages, tags, tips/reviews, photos and videos. Taking Foursquare as an example, messages and tips are two primary types of media content that can be associated with check-ins. While check-in messages tend to express real-time personal feelings, tips are more like customer reviews. Users can also add tags to a POI to describe its features, such as "free Wi-Fi", "wine" and "vegan" for a restaurant. Such tags are often used to index POIs for search purposes [3]. In addition, Foursquare also allows users to add photos to POIs.

The temporal dimension consists of check-in times, which are often used to characterize user dynamics over time. Fig. 2(c) plots the weekly temporal traffic pattern of check-ins over 168 hours in a week. We observe not only a clear daily periodicity, but also the difference between weekday (with three peaks in the morning, afternoon and evening, respectively) and weekend (with only one flat peak during the daytime).

The spatial dimension consists of POIs. Each POI refers to a specific place that may interest users, such as a bar or a restaurant. It could contain rich information about the place, including its name, GPS-coordinates, contact information, category, opening hours, menu and prices (for business POIs), etc. Fig. 2(d) visualizes the GPS coordinates of POIs on a world map. We observe that POIs from our dataset spread all over the world, with urban areas concentrating many of the POIs. To efficiently manage a large set of POIs, LBSNs often associate each POI with certain predefined categories. For example, POIs on Foursquare were organized as a four-level ontology tree, which consists of 10 root categories that are further classified into 451 categories on a second level, and so on. Fig. 2(b) shows the categorical distribution of check-in POIs. We find that daily-routine related activities including "Shopping Mall", "Home", "Train station" and "Office", are among the most frequent categories.

3 Key Research Challenges for LBSN Data Analytics

The above characteristics of LBSNs intrinsically bring three key challenges to data analytics. First, the four data dimensions of check-in data (in particular the rich media content on the semantic dimension) make LBSN data heterogeneous. Second, the user self-reported check-in data often suffer from data quality issues. Third, the fine-grained spatiotemporal check-in data often raises serious privacy issues.

3.1 Data Heterogeneity

The four data dimensions in LBSNs naturally imply data heterogeneity, as rich media content can be associated with

8. https://sites.google.com/site/yangdingqi/home/foursquare-dataset

check-in data. For example, users can report a check-in with a tag (in the form of keywords), with a tip (in the form of some text), with a photo (in the form of images), or even without any media content. We find that 66.75% and 32.75% of POIs are associated with tags and tips, respectively (28.4% of POIs have both tags and tips). Such heterogeneous data has great potential when analyzing patterns of user check-in activity, as different types of media content characterize different aspects of users' activity. For example, pure check-in counts can be regarded as "foot-voting", which implies a user's preference on POIs; tips can be regarded as user reviews on POIs, which imply detailed opinions of users on POIs. However, how to effectively extract user preference from both types of data and integrate them together to better model user preference on POIs and build high-quality POI recommendation applications is not straightforward.

Existing techniques on analyzing heterogeneous networks (such as LinkedIn and DBLP networks) cannot be adopted to LBSN data to fully capture its spatiotemporal properties. More precisely, the spatiotemporal regularity of human activities implies not only a strong inter-correlation between spatial and temporal dimensions, but also a strong intra-correlation within individual dimensions. For example, compared to a DBLP Scholar network (author-paper-venue-topic) where each author/paper/venue/topic are considered as independent, the POIs in an LBSN graph are highly correlated by their distance in space, which is a key factor to identify urban hotspots from POIs [4], for instance.

3.2 Data Quality

As self-reported data, LBSN data has an intrinsic limitation in its quality. Here we discuss two main issues about LBSN data quality, i.e., data sparsity and data credibility.

3.2.1 Data Sparsity

LBSN data is sparsely sampled user activity data. Specifically, different from the passively and continuously sampled user activity data (e.g., the Lausanne Data Collection Campaign dataset [5]), users actively and voluntarily report their check-in activity in LBSNs. Users are often willing to share only part of their activities by checking-in at POIs. In other words, a user might visit some POIs without checking-in there, due to several reasons such as uninteresting places, privacy concerns, or simply forgetting to do it. Therefore, check-in data is thus a biased sample from users' daily activities. Due to the social sharing purpose of check-ins, we regard check-in data as a *social representation* of users' daily activities.

The data sparsity issue of check-in data needs to be carefully considered when performing data analytic tasks [6]. In practice, the data sparsity issue is often handled by selecting only "active" users and POIs as the subject of study. For example, in our collected dataset, we find 30.75% or 7.88% "active" users who have at least one check-in per month or per week, respectively; we also find 13.3% or 0.89% "active" POIs which have been checked at least 10 or 100 times, respectively. These thresholds defining 'activeness' here are often application-specific. For example, urban dynamic analytics often use active POIs [7], while mobility modeling mostly resorts to active users' mobility traces [8].

3.2.2 Data Credibility

LBSN data also contains fake check-ins. Specifically, although the virtual and financial rewards in LBSNs (e.g., "Mayorship" and coupons on Foursquare) can effectively incentivize users to share more activities, such a gamification mechanism is a double-edged sword as the rewards may motivate users to report fake check-ins [6], i.e., checking-in at POIs without being present there. To alleviate this problem, Foursquare had to update their rules of "Mayorships" several times by adding terms including "check-in only counts when you're in close proximity (via your phone's GPS) to the POI" and "only one check-in per day counts and ties go to reigning mayor", etc. However, malicious users can still use Foursquare APIs to report fake check-ins. Here we highlight one particular type of fake check-ins that are commonly recognized by the research community, namely "sudden-move" check-ins [9], referring to the case that consecutive check-ins imply a speed faster than 1200 km/h (i.e., movement faster than the common airplane speed). An empirical study on our dataset shows that the "sudden-move" users represent about 1.1% of all the users, while their check-ins represent about 3.4% of all the check-ins in the collected data. We found that these "sudden-move" users indeed generate three times as many check-ins as average users; they probably want to report more check-ins for getting the rewards provided by Foursquare.

3.3 Data Privacy

As typical spatiotemporal data, check-in data contains sensitive information about fine-grained user activities, which may cause serious privacy leakage, e.g., revealing information about a user's identity, home address, health status, income level, friendship, etc. Despite many existing privacy protection techniques, privacy protection mechanisms need to be carefully designed considering the practical use cases of LBSNs. For example, to protect a user's location privacy, existing methods for location-based services often slightly alter or generalize the user's location data in order to avoid revealing her real position. However, such a mechanism may not be appropriate for LBSNs as it hinders the key benefits for the LBSN users [10], who intendedly share their presence within their friends by checking in at POIs. Hence, they might not appreciate the fact that the service hides their location information (even any part of their check-in data) from their friends. For example, when a user is watching a football match at a bar and wants to share her presence to potentially attract some nearby friends, she does not want her location to be obfuscated in any sense. Therefore, this factor needs to be carefully considered for designing an effective yet practical privacy protection mechanism for LBSNs.

4 Opportunities and Applications in Smart Cities

LBSN data contains fine-grained, semantically rich, spatiotemporal user activity information, which provides us

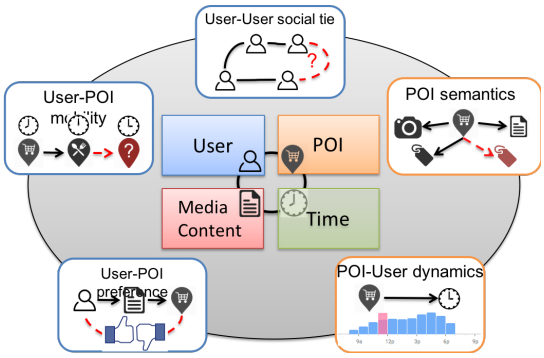


Fig. 3. Data analytics within and across the four data dimensions

with an unprecedented opportunity to understand urban dynamics. On one hand, we could mine LBSN data to understand the internal relationship among users, POIs, time and media content (within and across the four data dimensions). On the other hand, we could also integrate LBSN data with further urban data to discover its external correlations with various factors of urban environments, and thus reveal rich urban dynamics.

4.1 Data Analytics within and across the Four Data Dimensions

The four data dimensions of LBSN encompass users, POIs, time and media content, respectively. Check-in data thus creates links across these four dimensions. Based on such data, researchers design and develop novel models to capture the inherent patterns and to predict "missing links" within and across the four data dimensions. Here we adopt a broad definition of missing links, referring to any links within and across the four data dimensions that have not been established. From the perspective of users and POIs (physical entities in LBSNs), we first introduce five types of such "missing links" as shown in Fig. 3, followed by a discussion on data analytics across different types of "missing links".

4.1.1 User-POI Preference

User check-in data massively implies their preference on POIs. Based on users' check-in history, we are able to extract user preference on POIs, and then predict a user's preference on unvisited POIs (missing user-POI preference) for recommendation [11]. The most straightforward way to model user preference from check-in data is based on its count, using methods like repeat customer theory (a POI is preferred by a user if she has visited it more than twice), mapping the check-in count to a classical 1-5 rating scale, or coming up with a ranking list of POIs according to the counts [3]. With the help of rich media content, in particular user reviews, we could also model and predict fine-grained user preference on certain specific aspects of POIs (e.g., a certain dish in a restaurant). In addition, considering the spatiotemporal dynamics of user check-in data, we can further make context-aware predictions of user preference on POIs. Effectively modeling user-POI preferences can enable high-quality personalized location-based services for both customers and business owners. On one hand, end users could receive personalized POI recommendations. On the other hand, business owners of POIs could identify potential customers for marketing purposes.

4.1.2 User-POI Mobility

Mobility patterns characterize the spatiotemporal properties of users. In LBSNs, despite the sparsity of check-in data, one can still discover the regularity of user movements over space and time (e.g., staying at home during the night and working in the office during the day). Previous studies have shown that 50% to 70% of all check-in data can be explained by periodic behaviors [12]. In addition, social relationships are also a key factor to understand user mobility. For example, a user may check in at a POI near one of her friends' homes when visiting her. It has also been shown that such social relationships could further explain about 10% to 30% of all check-ins [12]. By understanding and modeling user-POI mobility, we are able to not only recover missing check-ins from the past, but also predict future user check-ins (where a user will be in the future) [8], [13]. In particular, knowing a user's next move is critical to enabling many proactive smart city applications, such as automated reservations and personalized advertisements.

4.1.3 User-User Social Ties

Similar to classical online social networks, social ties (both unidirectional and bidirectional) also exist among users on LBSNs. In addition to online social networks that imply proximity in the virtual world, check-ins at POIs imply users' proximity in the physical world. Such links across the virtual and physical worlds re-define the community structures on LBSNs. Specifically, beyond traditional social tie prediction methods that are mainly based on social network structures, check-ins on LBSNs create a novel type of relationship, i.e., "place-friends", who have collocated with the user at certain POIs. It has been shown that about 30% of the new social ties are created among such "place-friends" [14], which could significantly improve social ties prediction performance. In practice, social ties recommendation to a user could help the user extend her social circles by making new friends, or get interesting information feeds by following other users.

4.1.4 POI-Semantics

User-generated rich media content associated with a POI can be used to characterize the POI. Traditionally, linking user-generated content to locations [15] is a key task in geographic information systems. Fortunately, user-generated media content is intrinsically connected to locations in LBSN data. One can directly extract rich POI semantics from a specific type of media content. For example, by analyzing and aggregating user reviews on a POI, one can analyze the POI using not only keywords (topics) that indicate the characteristics of the POI, but also the collective sentiments that users have on specific aspects of the POI. Second, heterogeneous data can also be combined for POI semantic analysis. For example, photos of a POI on LBSNs often contain both related scenes of the POI and irrelevant selfies. The semantics of the POI can be enriched by extracting and

(a) New York City (b) Tokyo

Fig. 4. Categorical distribution of POIs in different cities.

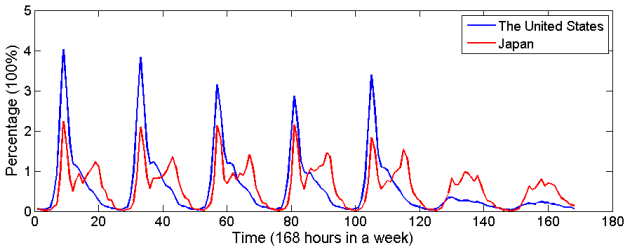


Fig. 5. Traffic pattern for "Office" POIs in the United States and Japan.

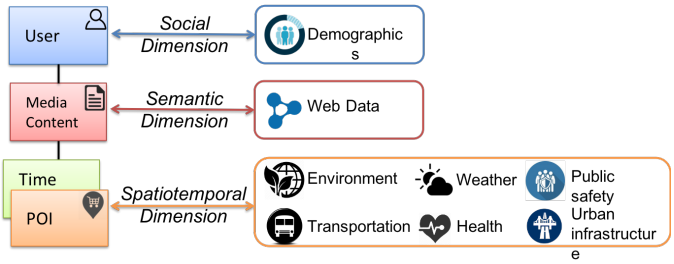


Fig. 6. Data fusion with further data

filtering visual concepts from those photos (e.g., the appearance of a specific scene/object) and then linking the relevant visual concepts to the textual information (keywords extracted from user reviews) [16]. A semantic-enriched POI database is the cornerstone for enabling high-quality search and recommendation applications. Finally, collective POI statistics are also an important ingredient for urban analytics. For example, the categorical distribution of POIs has been used in computational social science for identifying cultural differences between cities [7]. Fig. 4 demonstrates the categorical distribution of POIs in New York and Tokyo, where we observe clearly the cultural differences between the two cities, i.e., New York users usually check in at home, bars, gyms, outdoor places while Tokyo users often check in ramen/noodle houses, convenience stores, and Japanese restaurants.

4.1.5 POI Dynamics

From the point of view of a POI, user check-in data intrinsically implies its temporal traffic dynamics. Specifically, by analyzing the collective check-in count at a POI over time using time series analysis techniques, one can model the temporal traffic dynamics of the POI. For example, Fig. 5 compares such traffic patterns for "Office" POIs in the United States and Japan. We observe that Japanese daily working time is obviously longer than that of Americans, and a large number of Japanese work particularly in the evening and also during the weekend. Such temporal traffic patterns are also key ingredients to predict the volume of check-ins at a certain time in the future, which can be regarded as an indicator of the volume of customers. This can help not only the business owner of the POI to prepare for the popular time, but also customers to plan their visits to avoid peak hours, for example. Moreover, it has been shown that the volume of check-ins is indeed a strong indicator of sales for commercial POIs. For example, by analyzing the temporal traffic dynamics of check-ins at Chipotle Mexican Grill, a popular Mexican food chain, Foursquare has successfully predicted that its sales would plummet 30% in the first quarter of 2016[9].

4.1.6 Data Analytics across "Missing Links"

Considering multiple types of the above "missing links", we are able to better understand different aspects of urban dynamics. For example, combining POI-semantics and POI-dynamics, urban dynamics is analyzed via an embedding model showing spatiotemporal activity dynamics in an urban environment [4]; considering user-POI mobility and user-user social ties, an asymmetric impact of mobility and social relationships on predicting each other is revealed [17].

9. http://fortune.com/2016/04/15/chipotle-foursquare-swarm/

4.2 Data Fusion with Further Data

The key idea of data fusion is to combine the advantages of LBSN data with other urban and Web data for urban analytics purposes. In particular, the recent movement of open data has made large repositories of urban data publicly accessible, such as demographics, environment, public transportation, urban infrastructure data, etc. These different types of urban and Web data often share one or more common dimensions with LBSN data. Fig. 6 illustrates the common data dimensions shared between LBSN data and some representative urban data. Linking LBSN data to urban and Web data via these common data dimensions can give us insight into the correlation between citizens' social activity on LBSNs and other factors of urban environments, which can benefit both citizens and urban authorities by empowering smart city applications.

4.2.1 Spatiotemporal Dimension

The most popular way of integrating LBSN data with urban data is via its spatiotemporal dimension. In practice, many types of urban data include spatiotemporal information. For example, environmental monitoring data shows the quality of the environment at a certain time and place; Public transportation data shows the collective dynamics of citizens over space and time; Public safety data contains crime records associated with GPS coordinates and timestamps. LBSN data also contains spatiotemporal information, which could be considered from two perspectives. First, POI data can be used to characterize the functions of urban areas, such as a commercial district with many shops and restaurants; an industrial district with many offices and factories [18]. The categorical distribution of POIs is indeed a strong indicator of such functions, where we can analyze

its underlying correlation with other urban data over space and time. For example, we find that the categorical distribution of POIs over space is correlated with the crime rate in New York City [19]; the top positively correlated POI categories include "Argentinian Restaurant" and "Mexican Restaurant" while the top negatively correlated ones include "College Auditorium" and "College & University". Second, check-ins data on POIs characterizes human flows in urban areas, which can then be used to analyze crowd mobility. For example, collective check-in counts on POIs have been adopted to effectively predict traffic demand [20].

4.2.2 Semantic Dimension

The semantic dimension of LBSN data contains user-generated media content about POIs, such as tips, photos, tags, etc. Such POI-related information can be significantly enriched using Web data, as a large number of POIs have related information available online. For example, tourist spots often have their homepages with rich descriptions; popular restaurants and bars often have reviews from experts on Zagat[10]; hotels often have a photo gallery on Booking[11]. These external information sources on the Web can be leveraged to complete the user-generated media content on LBSNs, and thus provide users with richer information when searching for POIs.

4.2.3 Social Dimension

Users are the main subject of study on the social dimension. Although LBSNs often allow users to build a brief profile, the available attributes in the profile are very limited compared to rich demographic data. On the contrary, demographic data often lacks mobility information. Therefore, linking demographic information to a group of LBSN users could help us to discover correlations between users' mobility (characterized by their check-in behaviors) and their demographic and socioeconomic features. For example, it has been shown that people in wealthy cities (high average household income) travel more frequently to distant places than people in poorer cities [9]. Understanding these correlations is important for urban authorities for better resource management, for example.

5 Conclusion

Location-centric social media platforms have attracted millions of users sharing their activities (i.e., check-ins) on POIs, resulting in an invaluable data source containing fine-grained, semantically rich, spatiotemporal user activity information. This article discussed the challenges and opportunities in mining LBSN data for urban analytics. We highlighted three key challenges for LBSN data analytics, i.e., data heterogeneity, data quality and privacy. To efficiently leverage LBSN data for urban analytics, we first discussed data analytics based only on LBSN data, and highlighted five research directions. We then discussed the opportunity of fusing LBSN data with further urban and Web data for data-driven urban analytics. With more and more urban data available online, integrating LBSN data with these data is a promising research direction gaining increasing popularity recently, as it can shed more light on richer urban dynamics and empower intelligent smart city applications.

10. https://www.zagat.com/
11. https://www.booking.com

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 683253/GraphInt).

References

[1] Y. Zheng, "Tutorial on location-based social networks," in *WWW*, vol. 12, no. 5, 2012.
[2] Y. Zheng and X. Zhou, *Computing with spatial trajectories*. Springer Science & Business Media, 2011.
[3] D. Yang, "Understanding human dynamics from large-scale location-centric social media data: analysis and applications," Ph.D. dissertation, Institut National des Télécommunications, 2015.
[4] C. Zhang, K. Zhang, Q. Yuan, H. Peng, Y. Zheng, T. Hanratty, S. Wang, and J. Han, "Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning," in *WWW*. International World Wide Web Conferences Steering Committee, 2017, pp. 361–370.
[5] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen *et al.*, "The mobile data challenge: Big data for mobile computing research," Tech. Rep., 2012.
[6] G. Wang, S. Y. Schoenebeck, H. Zheng, and B. Y. Zhao, ""will check-in for badges": Understanding bias and misbehavior on location-based social networks," in *Tenth International AAAI Conference on Web and Social Media*, 2016.
[7] D. Yang, D. Zhang, and B. Qu, "Participatory cultural mapping based on collective behavior data in location-based social networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, pp. 1–23, 2016.
[8] D. Yang, B. Fankhauser, P. Rosso, and P. Cudre-Mauroux, "Location prediction over sparse user mobility traces using rnns: Flashback in hidden states!" in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 2020, pp. 2184–2190.
[9] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.
[10] D. Yang, D. Zhang, B. Qu, and P. Cudré-Mauroux, "Privcheck: privacy-preserving check-in data publishing for personalized location based services," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 545–556.
[11] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, "Recommendations in location-based social networks: a survey," *GeoInformatica*, vol. 19, no. 3, pp. 525–565, 2015.
[12] 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*. ACM, 2011, pp. 1082–1090.
[13] N. A. H. Haldar, J. Li, M. Reynolds, T. Sellis, and J. X. Yu, "Location prediction in large-scale social networks: an in-depth benchmarking study," *The VLDB Journal*, vol. 28, no. 5, pp. 623–648, 2019.
[14] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1046–1054.
[15] M. Gritta, M. T. Pilehvar, and N. Collier, "Which melbourne? augmenting geocoding with maps," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 1285–1296.
[16] Y.-L. Zhao, Q. Chen, S. Yan, T.-S. Chua, and D. Zhang, "Detecting profilable and overlapping communities with user-generated multimedia contents in lbsns," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 10, no. 1, p. 3, 2013.

[17] D. Yang, B. Qu, J. Yang, and P. Cudre-Mauroux, "Revisiting user mobility and social relationships in lbsns: a hypergraph embedding approach," in *WWW*. ACM, 2019, pp. 2147–2157.
[18] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 186–194.
[19] D. Yang, T. Heaney, A. Tonon, L. Wang, and P. Cudré-Mauroux, "Crimetelescope: crime hotspot prediction based on urban and social media data fusion," *World Wide Web*, vol. 21, no. 5, pp. 1323–1347, 2018.
[20] L. Chen, D. Zhang, G. Pan, X. Ma, D. Yang, K. Kushlev, W. Zhang, and S. Li, "Bike sharing station placement leveraging heterogeneous urban open data," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 571–575.

Dingqi Yang is a senior researcher at the University of Fribourg in Switzerland. He received the Ph.D. degree in computer science from Pierre and Marie Curie University and Institut Mines-TELECOM/TELECOM SudParis, where he won both the CNRS SAMOVAR Doctorate Award and the Institut Mines-TELECOM Press Mention in 2015. His research interests include big social media data analytics, ubiquitous computing, and smart city applications.

Bingqing Qu is a researcher at the University of Fribourg in Switzerland. She received her Ph.D. in Computer Science in University of Rennes 1 in 2016. Her research interests include historical document analysis, multimedia content analysis, social media data mining and computer vision.

Philippe Cudre-Mauroux is a Full Professor and the director of the eXascale Infolab at the University of Fribourg in Switzerland. He received his Ph.D. from the Swiss Federal Institute of Technology EPFL, where he won both the Doctorate Award and the EPFL Press Mention. Before joining the University of Fribourg he worked on information management infrastructures for IBM Watson Research, Microsoft Research Asia, and MIT. His research interests are in next-generation, Big Data management infrastructures for non-relational data.