Annals of the American Association of Geographers

Routledge
Taylor & Francis Group

The Missing Parts from Social Media Enabled Smart Cities: Who, Where, When, and What?

Journal:	*Annals of the American Association of Geographers*
Manuscript ID	AN-2018-1058.R3
Manuscript Type:	Special Issue 2020
Manuscript Category:	Geographic Methods
Key Words:	Location-based social media, Data quality, Sampling biases, Smart city

The Missing Parts from Social Media Enabled Smart Cities: Who, Where, When, and What?

Yihong Yuan[1,*], Yongmei Lu[1], Edwin T. Chow[1], Chao Ye[2], Abdullatif Alyaqout[1], Yu Liu[2],

[1]Department of Geography, Texas State University, San Marcos, TX 78666
[2]Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing, China, 100871

*corresponding author

Abstract. Social Network Sites (SNS), such as Facebook and Twitter, have attracted users worldwide by providing a means to communicate and share opinions and experiences of daily lives. When empowered by pervasive location acquisition technologies, location-based social media (LBSM) has become a potential resource for smart city applications to characterize social perceptions of place and model human activities. However, there is a lack of systematic examination of the representativeness of LBSM data. If LBSM data are applied to decision-making in smart-city services, such as emergency response or transportation, it is essential to understand their limitations in order to implement better policies or management practices. This study formalizes the sampling biases of LBSM data from various perspectives, including sociodemographic, spatiotemporal, and semantic. This article examines LBSM data representativeness issues using empirical cases and discusses the impacts on smart city applications. The results provide insights for understanding the limitations of LBSM data for smart city applications and for developing mitigation approaches.

***Keywords*:** *Location-based social media*, *data quality*, *sampling biases*, *smart city*

Introduction

The White House launched *the Smart Cities Initiative* in 2015 to support cities, federal agencies, universities, and the private sector in developing new technologies that make cities more inhabitable and equitable (The White House 2016). The following years witnessed a number of federal agencies, private companies, as well as non-profits join the table and provide financial support for smart city development. There are several key technologies for smart city applications: 1) sensor-enabled physical infrastructure that provides real-time monitoring of urban resources; 2) communication infrastructure that connects the deployed sensors (e.g., the Internet of Things (IoT)); and 3) big data generated by the various sensors and the associated new theories and applications (Hancke, Silva, and Hancke 2013). These technologies are often inseparable. For instance, the massive data generated by sensors has contributed to the rise of big data (Batty 2013), which in turn expanded the definition of sensing technologies beyond just physical sensors (e.g., Bluetooth sensors). The increase of social networking sites (SNS) where people can share their social life has introduced new opportunities to monitor individuals' activities and the perception of their surroundings. Researchers have defined location-based social media (LBSM) as 'Social Network Sites that include location information'(Roick and Heuser 2013). LBSM has been widely used as potential resources to characterize social perceptions of places and to model human activities in various applications. Innovative concepts such as "human sensing" and "social sensing" were introduced into sensing technologies to refer to human observations of both physical and social geographies (Calabrese, Ferrari, and Blondel 2015; Liu et al. 2015).

However, like other types of big (geo) data, LBSM data have various data quality issues, such as accuracy, precision, completeness, and representativeness (Shi et al. 2018; Yuan, Wei, and Lu 2018). Different SNS tend to attract certain population groups and support the sharing of particular

content, making them limited in data representation (Golub and Jackson 2010). In other words, biased sampling (e.g., demographically, spatially, temporally, and semantically) naturally leads to data representativeness issues. If LBSM data are applied to decision-making in smart city services, understanding the sampling biases of such data is critical for implementing better policies or management practices. This study examines the representativeness issues of LBSM data caused by sampling biases from sociodemographic, spatiotemporal, and semantic perspectives. The terms "data representativeness issues" and "sampling biases" are used interchangeably in the rest of this paper. The main objective is providing a framework to examine LBSM-enabled smart city services and their limitations. We discuss the representativeness of LBSM data and their impacts on smart city applications by incorporating empirical analyses. The results provide valuable inputs for understanding how LBSM sampling biases may manifest themselves in smart city applications.

Challenges for Social Media Enabled Smart Cities

The implementation of a smart city requires the integration of three essential components: advanced information and communication technologies (ICTs), open governance, and resident-centered services; the third component is often overlooked in real-world smart city services. In other words, there is a tendency to over-emphasize the merit of technology in smart city services while the core purposes and functions of city operations are ignored (Kitchin 2015). The central goal of a city is to ensure the life quality of its residents through the management, preparation, and delivery of resources and services. To this end, ICTs form the technical backbone for smart cities, the service aspect serves as the ultimate goal, and the open governance aspect provides the means to achieve the goal. A smart city application or service that is built only on the merit of technologies without paying attention to people would risk disconnecting the service from its users. The

disconnection challenges go beyond identifying what services are needed; they include where, when, and by whom a service is needed.

In academia, the discussions about smart cities reflect a broad spectrum of views. Although a technocratic approach is not uncommon (e.g., Maeda 2012), researchers recognize the inherent comprehensive characteristics of smart cities (Perera et al. 2014). Harrison et al. (2010) emphasized that smart cities should successfully connect different infrastructures of a city – the physical infrastructure, the information-technology infrastructure, the social infrastructure, and the business infrastructure. Mohanty, Choppali, and Kougianos (2016) argued that a smart city is a system of systems where IoT and big data improve a city's operation and help it fulfill its objective of improving life quality. After discussing how big data technologies can support different smart city applications, Al Nuaimi et al. (2015) explored a number of open issues, including the role of social media for smart city applications and its ramifications, how differing levels of access to information affects an individual's power and political position, and the effectiveness and quality of smart city applications.

To ensure efficient services, smart city applications need to be built upon real-time measurements and massive data collection. LBSM data, as a complement to traditional sensors, are particularly useful due to their uniqueness in recording human experiences and behaviors at fine spatio-temporal resolutions (Doran et al. 2016). Taking advantage of LBSM data, urban studies have examined spatio-temporal dynamics of cities while seeking insights into the social, cultural and political aspects of urban life (Hochman and Manovich 2013; Licoppe 2016; Cabalquinto 2018).

However, social media data are not universally representative. Studies have examined demographic bias of social media data (e.g. Sloan et al. 2015 and Yuan et al. 2018). Social and political inequity was found to not only perpetuate the use of social media, but also feed back into

people's usage of urban space (Boy and Uitermark 2017). In addition to reflecting human experiences, social media may affect human experiences and opinions of space (Evans and Saker 2017). Hence, it is crucial to understand the limitations of LBSM data when applying them to smart city applications.

The Missing Parts from Social Media Data for Smart Cities

Researchers have identified 5*W*s and 1*H* ("Who", "Where", "When", "What", "Why", and "How") in social media studies (Khosrow-Pour 2018). "Who" refers to the challenges of identifying user groups on social media and evaluating the data quality associated with biased sampling (Longley and Adnan 2016). "Where" and "When" identify the spatio-temporal patterns from social media content, which are the most crucial aspects of LBSM data for geographic information science (GIScience) (Zhang et al. 2016). "What" focuses on mining the semantics of user-generated content for urban planning and e-governance (Hu 2018). "How" and "Why" focus on the underlying processes within the scheme of social media, such as "How does a social network form on SNS?" and "What are the motivations of SNS users?". Although "How" and "Why" questions help to understand the theoretical foundations of social media, most social media applications focus on the first 4*W*s ("Who", "Where", "When", "What") (Khosrow-Pour 2018). The rest of this paper focuses on the first 4*W*s to demonstrate the representativeness issues of LBSM for smart city applications.

Who are reflected by LBSM data?

LBSM users are not a random sample in terms of their social, economic and demographic background (Golub and Jackson 2010). Pinterest, for instance, particularly attracts women between the ages of 25 and 34 with average household incomes of $100,000. One crucial challenge

in quantifying these biases is obtaining accurate data on social media users because many SNS do not require users to provide personal information. Salganik (2018) discussed common characteristics of big data, such as the lack of demographic information and the representativeness of the data. Previous research either conducted user surveys or harvested user profiles or posts to infer their demographics (Longley and Adnan 2016). A survey-based study by Zickuhr (2013) found that LBSM use in the U.S. is not equal across age, gender, and race and that young people, women, and ethnic minorities have a greater LBSM presence. Recent studies in computational social science reported similar findings about the biases of digital trace data and the importance of combining such data with traditional survey data (e.g., Foster 2017) .

Naturally, such demographic biases may impact the reliability of applying LBSM to urban services. For example, Rizwan et al. (2018) found that check-ins from female users were more spread out in the city, whereas check-ins from male users showed a more clustered pattern in centralized districts. Zhong et al. (2015) identified the connection between user demographic factors (e.g., age, gender, and income) and their points of interest (POIs) check-ins. They constructed a model to effectively predict the demographics of Weibo users based on their check-in time and location.

Our case study shows how the senior population (age 65+) in China is systematically under-represented on Weibo (Figure 1). Using a random sample of 230,000 Weibo users who checked-in their locations at least once between March – November 2015, Figure 1 displays their distribution using an index of under-representativeness, $I_{UR}(i)$.

$$I_{UR}(i) = \log\left(\frac{P_C(i)}{P_W(i)}\right) \quad (1)$$

where $P_C(i)$, $P_W(i)$ represents the percentage of demographic group i (e.g. seniors) in the census data and Weibo data, respectively. A positive value indicates that a demographic group is underrepresented in Weibo. The lower the value, the better this demographic group is represented.

[Figure 1 near here]

This underrepresentation of seniors on Weibo demonstrates a strong regional pattern. For instance, the provinces in northeast China show a significantly clustered pattern of senior underrepresentation (Moran's I z-value=2.20, p-value $<$ 0.05). Another cluster of underrepresentation is in southwest China (e.g., Sichuan, Chongqing, Guizhou). A deeper understanding of the underlying mechanisms for this pattern requires a comprehensive analysis of the factors affecting senior citizen usage of LBSM, such as cultural backgrounds, economic development, resources allocated to support seniors, etc., which is beyond the scope of this paper. Although this example focused on provincial-level patterns, understanding this unbalanced population representation in social media data is essential for LBSM-enabled smart city applications to effectively serve residents. Future studies should conduct a city or sub-city level analysis to explore how spatial scales and the modifiable areal unit problem (MAUP) may impact the results.

Despite the challenges in mitigating LBSM demographic biases, it still provides a valuable data source for smart city applications. Salganik (2018) pointed out that even though social scientists are more used to probabilistic random samples from a well-defined population, nonrepresentative data can still provide valuable insights, especially in the exploratory stage of outlier patterns and causations. Therefore, if city officials were to rely on nonrepresentative social media data to engage a broader audience in urban planning and infrastructure renovation (Borje et al., 2009), it would be important to identify suitable research questions. For example, it is feasible

to answer questions like "Are there abnormal spatial clusters in the city during a musical festival?" based on LBSM data; however, questions like "What is the average number of people impacted by Hurricane Harvey in each county?" requires more representative data and cannot be answered solely based on social media data.

In addition, it is possible to adjust the sampling process through methods like adopting stratified sampling or combining LBSM samples with other public survey data, which would help mitigate the influence of LBSM user sampling biases. Previous studies have applied machine learning algorithms to estimate LBSM users' demographic information, such as age and gender, based on their profile information and the semantics of their posts (Longley and Adnan 2016). Another option is to generate stratified samples across space based on census data (e.g., income).

***Where* and *When* are things happening on LBSM?**

The public sector has utilized spatio-temporal information from LBSM data to model human activities in various smart city applications. For example, the Livehoods Project aims to better understand the dynamics of urban dwellers and re-image cities using LBSM data (Cranshaw et al. 2012). However, studies have demonstrated that check-in data tend to cluster in certain areas, causing an biased profiling of activities across space and through time (Sloan et al. 2015). For example, Bawa-Cavia (2011) identified social hubs (i.e., where social media users are more likely to generate a high density of activities) in London, New York, and Paris. Sun and Gonzalez Paule (2017) also identified that highly rated restaurants are more likely to cluster spatially and receive more ratings on Yelp. Hence, there are inherent biases and representativeness issues in the spatio-temporal data acquired from LBSM.

Austin, Texas (ATX) is among the U.S. cities that actively pursue smart city development (City of Austin 2017). Using a 4-month Twitter dataset from January to April of 2016, we calculated the

number of geotagged tweets by census tracts and correlated that with census data to evaluate social media usage in different urban areas (P_{LBSM}). Here we use the residential population data from the 2010-2014 American Community Survey conducted by the United States Census Bureau (2015).

$$P_{LBSM} = F_{LBSM} / \text{Population} \quad (2)$$

where F_{LBSM} is the frequency of geotagged tweets in an area.

[Figure 2 near here]

As shown in Figure 2, P_{LBSM} is not spatially uniform across the city. Towards the city center (the red polygon), the amount of LBSM check-ins is disproportionally high. This is potentially due to geographic distribution of POIs in Austin, such as the bars and restaurants on 6th Street, the convention center, and the Texas State Capitol. Another cluster of check-ins is at the Austin Airport (marked by an arrow). It is clear that certain districts in central Austin attract more people to check in, therefore these locations tend to be over-sampled in LBSM data. Therefore, it is essential to quantify such spatial biases and properly adjust the representativeness of LBSM data when developing smart city applications, such as emergency response or transportation, to represent human mobility (Liu et al. 2014).

To analyze the temporal patterns of LBSM, we aggregated that same Twitter dataset from Austin by recording the number of geotagged tweets for each hour. To validate the human mobility pattern, we used Bluetooth data collected by ATX along major streets and freeways and aggregated them for each hour during the same period. This dataset captures the presence of Bluetooth-enabled devices when they passed by a receiver. Although Bluetooth data does not measure physical movement of a well-sampled population perfectly, it captures the "naturally occurring" physical movement better than LBSM does. Due to the lack of ground truth data for human mobility patterns,

we used Bluetooth data as a proxy for physical movement in this study. Both datasets are normalized to the range [0,1] and can be considered an indication of hourly human activity reflected by LBSM and Bluetooth, respectively. There is a clear time lag where the peak activity for Twitter data is a few hours after that of Bluetooth data (Figure 3). This demonstrates the biases of LBSM data in reflecting the temporal patterns of human activities, indicating that there is a discrepancy between LBSM posts and when the activity referenced by the posts took place. It is likely that LBSM users did not post on LBSM when they were rushing to work, which potentially caused the time lag between when the events occurred and the posting on LBSM of those events. This type of temporal bias has crucial implications for smart city services that respond to real-time mobility patterns of urban dwellers, such as transportation and event planning.

[Figure 3 near here]

The "where" and "when" challenges are beyond simple sampling biases. Locations and time stamps from LBSM data may have different levels of accuracy; space-based geotagged posts with precise x- and y-coordinates are more accurate than place-based posts using a descriptive of or reference to a loosely-defined location. For example, "Houston" can refer to its downtown area, the centroid of the city, or anywhere within the city limit as determined by that social media platform. In the context of multimedia, whether it's a picture or video, the area of interest (AOI) can be captured either on-scene (i.e., where it was posted) or off-scene (i.e., a certain distance away from the post location). The resulting on- or off-scene capture can be attributed to the time-lag between data capture and posting, which is not uncommon when the multimedia file is large and mobile bandwidth is limited, or when the user is moving quickly (e.g., in a vehicle). Moreover, social media platforms like Foursquare allow users to check in to receive points when they are within the vicinity of a certain location. These practices may introduce spatio-temporal uncertainties into LBSM data.

As people share critical information on social media during a disaster (Smith et al. 2017), first responders and GIScientists harvest LBSM to identify vulnerable populations, conduct damage assessments, and allocate appropriate resources for disaster relief and response (Fohringer et al. 2015). Thus, location accuracy is vitally important for effective emergency response and disaster management.

In a case study that harnessed tweets and crowdsourced data containing water-depth information during Hurricane Harvey in 2017, more than 95 percent of relevant posts contained multimedia (i.e., text-only posts account for ~5 percent). Among all collected tweets, 244 geotagged tweets had a valid location - either represented by a latitude/longitude pair or by selecting a place with predefined coordinates (e.g., University of Houston, Downtown Houston, etc.) The geographic features and landmarks captured in the pictures/videos were compared with those found in Google Maps' Street View and/or 3D views. There were 107 (~44 percent) and 137 (~56 percent) off- and on-scene posts, respectively. In this study, off-scene posts were defined as those located 500 meters away from the AOI reference location. By comparing the location of the AOI and the corresponding post location shared on SNS, the distance offset between them was calculated as spatial error (Figure 4). During Hurricane Harvey, the mean spatial errors of geotagged posts related to water-depth were 47.7 m (or 198.1 m, excluding 104 posts that had accurate GPS-derived coordinates) and 5,219 m for on- and off-scene posts, respectively (Table 1). This finding supports the cautious use of LBSM and the importance of reporting spatial accuracy for emergency planning and disaster management.

[Figure 4 near here]

[Table 1 near here]

What do people talk about on LBSM?

In addition to biases in the "who," "where," and "when" aspects, semantic biases from LBSM are also worth noting. The content of social media is closely related to the functionalities and characteristics of each SNS platform (Morstatter et al. 2013). Inevitably, various biases exist when conducting sentiment analysis, public opinion collection, and topic extraction from such datasets. Instead of expressing opinions on public matters such as traffic, politics, or urban planning, social media users are more willing to publicly discuss topics related to their personal life (e.g., leisure activities) (Lansley and Longley 2016). Although most people, LBSM users included, spend most of their time around a few key locations, it remains difficult to identify the associated land use at these frequently visited locations just by examining the semantics of social media posts (Soliman et al. (2017). Previous findings reported a sentimental bias in which people are more likely to post on social media under the influence of positive emotions (Mitchell et al. 2013). Hence, if policy makers aim to collect opinions on city services, it is crucial to understand the nature, popularity, and associated sentiment of various topics on social media.

[Figure 5 near here]

In English, verb-noun phrases (e.g., "attend a wedding") are often used to describe human behavior and activities, but complete verb-noun phrases are not common on the Chinese language website, Weibo. In a study analyzing the geotagged Weibo posts in Beijing from January 2016 to January 2017, we used verbs instead of verb-noun phrases to examine activity space. The top 10 verbs extracted from Weibo (translated to English) are: eat [吃] (195,623), sleep [睡] (45,923), buy [买] (38,449), encourage [加油] (27,252), take pictures [拍(照)] (26,943), have fun [玩] (23,311), work [上班] (19,634), study [学习] (19,323), deliver [送(货)] (16,053) and stroll [逛] (12,888). These

verbs correspond to different types of activities, such as employment, leisure, education, etc. We used fuzzy c-means to cluster the verbs into the following categories: *work*, *study*, *daily life*, *leisure* and *others*. The fuzzy c-means algorithm generates a membership degree representing how well each verb fits into a certain category. The final results in each category include verbs with a membership degree larger than 0.8. Although there may be outliers due to the fuzziness of the algorithm, this method has proven to be effective in categorizing words based on their semantics (Cao, Song, and Bruza 2004). Several verbs in the "*others*" category, such as "Prevent [防治]" or "Bribe [贿赂]" can potentially be related to public issues; however, these verbs make up less than 1 percent of the entire sample set. Word clouds of two types of activities ("work" and "daily life") are shown in Figure 6. The non-verbs in the word clouds are due to the difference between Chinese and English.

[Figure 6 near here]

We found that certain types of activities, such as industrial activities (e.g., manufacturing) or political discussions, are rarely discussed on Weibo because social media users prefer to share topics related to their daily life. Another example of thematic bias can be illustrated by a search for keywords like "traffic accident" or "car accident" from Twitter data collected between 05/2015-04/2016 in Austin. The results only yielded 388 records, and 383 of them were from a verified public account, "*Total Traffic Austin*," which is owned by a private company in the business of traffic and weather broadcasting. This demonstrates the lack of discussion of certain issues on SNS, which brings challenges to the application of LBSM data to smart city services that requires topic extractions.

Discussion and Conclusion

LBSM provides rich data sources for modeling human activities and capturing citizens' perceptions in the age of instant access. However, this emerging data source also brings challenges for smart city services. Understanding these challenges is crucial for developing meaningful LBSM-enabled smart city services. This research takes an initial step of formalizing these challenges into a framework of four critical aspects ("Who", "Where", "When", and "What"). Nevertheless, there are other challenges that are not fully addressed in this paper, including but not limited to:

- Other data quality issues of LBSM: Although we briefly touched upon accuracy and precision, our research focuses on the representativeness of LBSM. In addition to the 4*W*s, other data quality issues can also affect smart city services. First, most LBSM application program interfaces (APIs), including Twitter and Weibo APIs, are only able to obtain around 1 percent of all geo-located posts, raising questions about data completeness. Second, fake check-ins and bots are inevitable issues on LBSM. Without a valid method to address this issue, LBSM-enabled smart city applications may be misguided. Third, the demographic profiles of LBSM users are mostly estimated by algorithms or from self-reported data, and it is a challenge to validate the credibility of such information. Fourth, LBSM data are biased towards overrepresenting "central users" from the perspective of the communication network, where a small group of users generate a disproportionate amount of data. Questions like, "Do we have overrepresentative data from central users in a social network," are essential to assess LBSM data quality and the scientific rigor of experimental design. Finally, social media platforms are driven by technological advancements and fast-changing culture so it is important to consider how LBSM data biases evolve. For example, the once-popular image and video hosting service, Flickr, lost many active users after changing ownership several times, which raises concerns regarding the representativeness of Flickr data.

- Distinctions between smart city services: Each smart city service has its own objectives and functionalities, which naturally leads to different data needs. For example, a service that aims to collect public opinions on urban infrastructure will be more sensitive to semantic biases, whereas a service that is designed to respond to certain urban events may be affected more by the spatio-temporal bias embedded in LBSM data. Policy makers should carefully investigate whether and to what degree a smart city service may be influenced by LBSM data biases based on the 4*W*s discussed in this paper. Table 2 lists several sample city services that are particularly well-suited to rely on LBSM data. Note that the reasons listed in Table 2 are only hypotheses, which should be carefully tested when developing smart city services.

[Table 2 near here]

It is important to note that the 4*W*s in this study are often inseparable. Specifically, the "*Who*" aspect (i.e., user group biases) contributes to both spatio-temporal biases ("*Where*" and "*When*") and semantic biases ("*What*"). For example, SNS tend to attract young people, who have their own preferred check-in locations and topics to discuss on social media. Figure 7 expands the social sensing framework discussed in Liu et al. (2015) and illustrates how these 4*W*s interact with each other:

- Suppose there are n demographic groups, and the number of people in each demographic group is $[U_1, U_2, \ldots U_n]$. The participation rate of each demographic group on LBSM is $[p_1, p_2, \ldots p_n]$. The total LBSM sample S can be calculated as:

$$S = \sum_{i=1}^{n} U_i * p_i \quad (3)$$

  This corresponds to the "*Who*" component.

- Assume that users in this sample S are conducting m types of activities in real life (e.g., work, leisure, study, etc.), where the number of activities in each type is denoted by $[T_1, T_2, \ldots T_m]$, and the probability of each activity type getting posted on social media is $[q_1, q_2, \ldots q_m]$. As a result, the number of activities reflected on social media can be noted as:

$$D = \sum_{j=1}^{m} T_j * q_j \quad (4)$$

  where q_j is highly dependent on the nature of each activity (i.e., the "*What*" component), and T_j is determined by sample S from the previous step.

- From a spatio-temporal perspective, activities conducted by users in S are unevenly distributed across space and time. The probability density of type j activity happening in a spatio-temporal unit (x,y,t) can be represented as $v(x,y,t,j)$, where x, y, t represent latitude, longitude, and time, respectively. The probability of a type j event happening at (x,y,t) appearing social media is proportional to $v(x,y,t,j)* q_j$. This demonstrates the various factors that may affect LBSM data quality, including the users who are posting in this unit (i.e., the sample S), the specific location/time of this unit (x,y,t), and the type of activities being conducted in this unit (q_j).

[Figure 7 near here]

To sum up, user sampling bias is the foundation of social media biases. In the meantime, activities conducted by these users distribute unevenly across space and time. Furthermore, these unevenly distributed activities also have different likelihoods of being posted to social media. Therefore, we should consider the 4*W*s in a synergistic way when developing smart city services.

Despite a lack of solutions to fully address or quantify these deficiencies of LBSM data, there are several ways to mitigate the potential problem. First, LBSM data can always be supplemented or corroborated by other data sources, such as census data and survey data, to improve the representativeness of LBSM samples. Second, it is important to identify target user groups from SNS data. Although user sample biases are inevitable, researchers can still extract the most representative groups on different SNS sites and design their research objectives according to the user groups available. Third, due to the low spatio-temporal sampling resolution of LBSM data, it is necessary to reevaluate the validity of classic mobility models, measurements, and algorithms when applied to such datasets.

The contribution of this research is two-fold. First, we provided a research framework to better understand the relationship between LBSM data and smart city applications through its limitation in reflecting human activities. The results also lay the groundwork for future efforts of applying LBSM data to various smart city services, such as real-time traffic management, early warning systems, and emergency responses. Second, the case studies provided empirical support to quantify the biases of two widely-used LBSM datasets (Weibo from China and Twitter from the United States) from four perspectives ("*who*", "*where*", "*when*", and "*what*"). The results of this research also provide a reference for policy makers and aid their efforts in applying LBSM data to city services. Future research can focus on extending this framework and identifying the distinctions of biases for different LBSM platforms.

Acknowledgements

Our deep and sincere thanks to Dr. Ling Bian and Ms. Jennifer Cassidento for the tremendous amount of work they put into organizing this special issue. We thank the anonymous reviewers

for their constructive comments, which greatly improved the content and clarity of this paper. Mr. Lei Zhang helped improving the grammar and style of this work.

References

Al Nuaimi, E., H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi. 2015. Applications of big data to smart cities. *Journal of Internet Services and Applications* 6 (1):25.

Batty, M. 2013. Big data, smart cities and city planning. *Dialogues in Human Geography* 3 (3):274-279.

Bawa-Cavia, A. 2011. Sensing the urban: Using location-based social network data in urban analysis. Paper presented at the First Workshop on Pervasive Urban Applications (PURBA), San Francisco, CA, June 12.

Boy, J. D., and J. Uitermark. 2017. Reassembling the city through Instagram. *Transactions of the Institute of British Geographers* 42 (4):612-624.

Cabalquinto, E. C. 2018. Smartphones as locative media. *Information Communication & Society* 21 (12):1751-1754.

Calabrese, F., L. Ferrari, and V. D. Blondel. 2015. Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys* 47 (2).

Cao, G., D. Song, and P. Bruza. 2004. Fuzzy k-means clustering on a High dimensional semantic space. Paper presented at Advanced Web Technologies and Applications, Berlin, Heidelberg, April 14-17.

City of Austin. 2017. Smart cities strategic roadmap update. http://www.austintexas.gov/edims/document.cfm?id=287471 (accessed November 20, 2018).

Cranshaw, J., R. Schwarts, J. Hong, and N. Sadeh. 2012. The Livehoods Project: Utilizing social media to understand the dynamics of a city. Paper presented at the Sixth International AAAI Conference on Webpages and Social Media, Dublin, Ireland, June 4–7.

Doran, D., K. Severin, S. Gokhale, and A. Dagnino. 2016. Social media enabled human sensing for smart cities. *AI Communications* 29 (1):57-75.

Evans, L., and M. Saker. 2017. *Location-based social media: Space, time and identity*. Cham, Switzerland: Palgrave MacMillan.

Fohringer, J., D. Dransch, H. Kreibich, and K. Schroter. 2015. Social media as an information source for rapid flood inundation mapping. *Natural Hazards and Earth System Sciences* 15 (12):2725-2738.

Foster, I. 2017. *Big data and social science : A practical guide to methods and tools*. Boca Raton, FL: CRC Press Taylor & Francis Group.

Golub, B., and M. O. Jackson. 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal-Microeconomics* 2 (1):112-149.

Hancke, G., B. Silva, and J. Hancke, Gerhard. 2013. The role of advanced sensing in smart cities. *Sensors* 13 (1):393.

Harrison, C., B. Eckman, R. Hamilton, P. Hartswick, J. Kalagnanam, J. Paraszczak, and P. Williams. 2010. Foundations for smarter cities. *IBM Journal of Research and Development* 54 (4).

Hochman, N., and L. Manovich. 2013. Zooming into an Instagram city: Reading the local through social media. https://firstmonday.org/ojs/index.php/fm/rt/printerFriendly/4711/3698 (accessed October 10, 2018).

Hu, Y. 2018. Geo-text data and data-driven geospatial semantics. *Geography Compass*:e12404.

Khosrow-Pour, M. 2018. *Encyclopedia of information science and technology*. Fourth edition. ed. Hershey, PA: IGI Global / Engineering Science Reference.

Kitchin, R. 2015. Making sense of smart cities: Addressing present shortcomings. *Cambridge Journal of Regions, Economy and Society* 8 (1):131-136.

Lansley, G., and P. A. Longley. 2016. The geography of Twitter topics in London. *Computers Environment and Urban Systems* 58:85-96.

Licoppe, C. 2016. Mobilities and urban encounters in public places in the age of locative media: Seams, folds, and encounters with 'pseudonymous strangers'. *Mobilities* 11 (1):99-116.

Liu, Y., X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, and L. Shi. 2015. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers* 105 (3):512-530.

Liu, Y., Z. W. Sui, C. G. Kang, and Y. Gao. 2014. Uncovering patterns of inter-urban trip and spatial interaction from social Media check-in data. *PLoS ONE* 9 (1):e86026.

Longley, P. A., and M. Adnan. 2016. Geo-temporal Twitter demographics. *International Journal of Geographical Information Science* 30 (2):369-389.

Maeda, A. 2012. Technology innovations for smart cities. Paper presented at 2012 Symposium on VLSI Circuits (VLSIC), Honolulu, HI, 13-15 June 2012.

Mitchell, L., M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. 2013. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE* 8 (5):e64417.

Mohanty, S. P., U. Choppali, and E. Kougianos. 2016. Everything you wanted to know about smart cities: The Internet of things is the backbone. *IEEE Consumer Electronics Magazine* 5 (3):60-70.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. Paper presented at Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM-13), Cambridge, Massachusetts, USA, July 8–11.

Perera, C., A. Zaslavsky, P. Christen, and D. Georgakopoulos. 2014. Sensing as a service model for smart cities supported by Internet of Things. *Transactions on Emerging Telecommunications Technologies* 25 (1):81-93.

Rizwan, M., W. Wan, O. Cervantes, and L. Gwiazdzinski. 2018. Using location-based social media data to observe check-in behavior and gender gifference: Bringing Weibo data into play. *ISPRS International Journal of Geo-Information* 7 (5):196.

Roick, O., and S. Heuser. 2013. Location based social networks - definition, current state of the art and research agenda. *Transactions in GIS* 17 (5):763-784.

Salganik, M. J. 2018. *Bit by bit : Social research in the digital age*. Princeton: Princeton University Press.

Shi, W., A. Zhang, X. Zhou, and M. Zhang. 2018. Challenges and prospects of uncertainties in spatial big data analytics. *Annals of the American Association of Geographers* 108 (6):1513-1520.

Sloan, L., J. Morgan, P. Burnap, and M. Williams. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS ONE* 10 (3):e0115545.

Smith, L., Q. Liang, P. James, and W. Lin. 2017. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management* 10 (3):370-380.

Soliman, A., K. Soltani, J. Yin, A. Padmanabhan, and S. Wang. 2017. Social sensing of urban land use based on analysis of Twitter users' mobility patterns. *PLOS ONE* 12 (7):e0181657.

Sun, Y., and J. Gonzalez Paule. 2017. *Spatial analysis of users-generated ratings of Yelp venues*

The White House. 2016. FACT SHEET: Announcing over $80 million in new federal investment and a doubling of participating communities in the White House smart cities initiative. https://obamawhitehouse.archives.gov/the-press-office/2016/09/26/fact-sheet-announcing-over-80-million-new-federal-investment-and (accessed November 11, 2018).

United States Census Bureau. 2015. American community survey. https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2014/5-year.html (accessed September 25, 2018).

Yuan, Y., G. Wei, and Y. Lu. 2018. Evaluating gender representativeness of location-based social media: A case study of Weibo. *Annals of GIS* 24 (3):163-176.

Zhang, W., B. Derudder, J. Wang, W. Shen, and F. Witlox. 2016. Using location-based social media to chart the patterns of people moving between cities: The case of Weibo-users in the Yangtze river delta. *Journal of Urban Technology* 23 (3):91-111.

Zhong, Y., N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. 2015. You are where you go: Inferring demographic attributes from location check-ins. Paper presented at Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, February 2-6.

Zickuhr, K. 2013. Location-based services. Pew Research Center's Internet & American Life Project, Pew Research Center, Washington, D.C.

Author information

YIHONG YUAN is an Assistant Professor in the Department of Geography, Texas State University, San Marcos, TX 78666. Email: y_y18@txstate.edu. Her research interests include spatio-temporal data mining, big geo-data analytics, and human mobility modeling.

YONGMEI LU is a Professor and Chair of the Department of Geography, Texas State University, San Marcos, TX 78666. Email: yl10@txstate.edu. Her research interests include GIScience and the application of GIS, spatial analysis, and modeling in health research and crime studies.

EDWIN T. CHOW is an Associate Professor in the Department of Geography, Texas State University, San Marcos, TX 78666. E-mail: chow@txstate.edu. His research interests include geocomputation, human dynamics, hazards, and environmental modeling.

CHAO YE is a M.S. student in the Institute of Remote Sensing and Geographic Information Systems, School of Earth and Space Sciences, Peking University, Beijing 100871. E-mail: gisyc@pku.edu.cn. His research interests include place semantics and natural language processing.

ABDULLATIF ALYAQOUT is a Ph.D. student in the Department of Geography, Texas State University, San Marcos, TX 78666. Email: afa21@txstate.edu. His research interests include cartography, geovisualization, and applications of GIScience in volunteered geographic information for natural hazards assessment and citizen-government interaction studies.

YU LIU is a Professor of GIScience in the Institute of Remote Sensing and Geographic Information Systems, School of Earth and Space Sciences, Peking University, Beijing 100871. E-mail: liuyu@urban.pku.edu.cn. His research concentration is in GIScience and big geo-data.

Figure captions

Figure 1. I_{UR} (65+) by province (excluding the South China Sea islands).

Figure 2. The distribution of P_{LBSM}.

Figure 3. Twitter and Bluetooth hourly activity comparison.

Figure 4. Example of an off-scene post with the inferred original and relocated post locations.

Figure 5. Spatial distribution of Weibo posts.

Figure 6. Word Clouds (A) work; (B) daily life.

Figure 7. Interactions among the 4*W*s.

Table 1. Statistical summary (in meters) of on-scene and off-scene spatial error. The numbers in parenthesis of on-scene posts exclude the 104 posts with an accurate coordinate from a GPS-enabled smart phone. See text for more details.

Post Location	***Count of posts***	***Min***	***Max***	***Mean***	***Standard Deviation***
On-scene	137 (33)	0 (33.3)	499.5	47.7 (198.1)	106.8 (132.4)
Off-scene	107	512.8	34,175.7	5,218.9	6,618.7

Table 2. Example smart city services/applications based on LBSM data.

Application	***Reasoning based on the 4Ws***
Identify hotspots of city night life	a. Social media particularly attracts young people ("Who"). b. It is more likely for users to post during leisure activities ("What").
Analyze tourist behaviors at transportation hubs (e.g., an airport)	a. Users are more likely to check-in at certain locations, such as arriving at an airport or a train station ("Where"). b. Users are more likely to post to social media when they travel ("What").
Model user behaviors during national holidays (e.g., the spring festival in China)	a. Users are more likely to post on social media during holidays ("When"). b. Users are more likely to travel to new destinations during holidays ("Where"). c. Users are more likely to conduct leisure activities during holidays ("What").

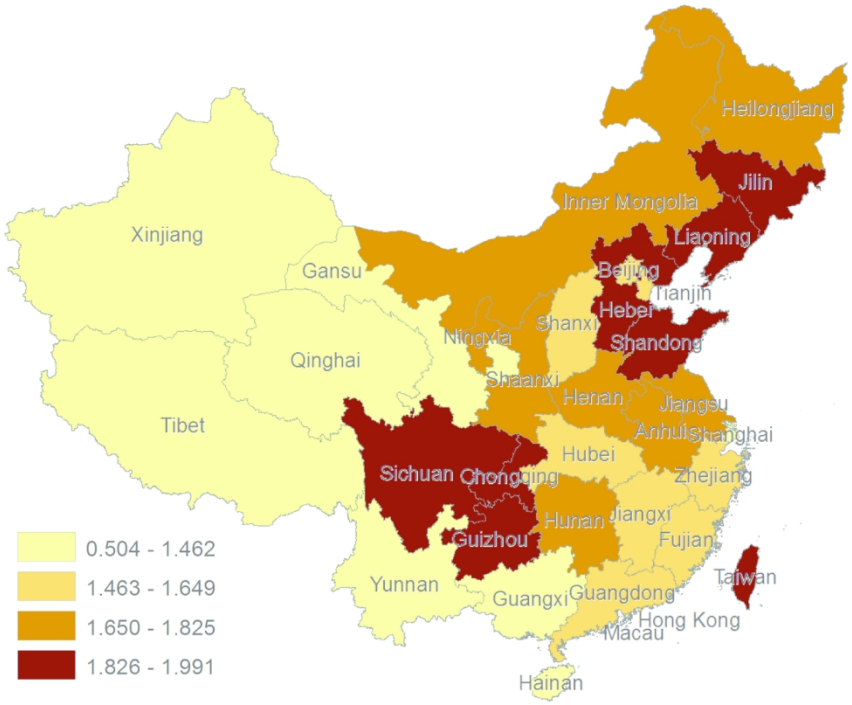

Figure 1. I_{UR} (65+) by province (excluding the South China Sea islands).

155x119mm (300 x 300 DPI)

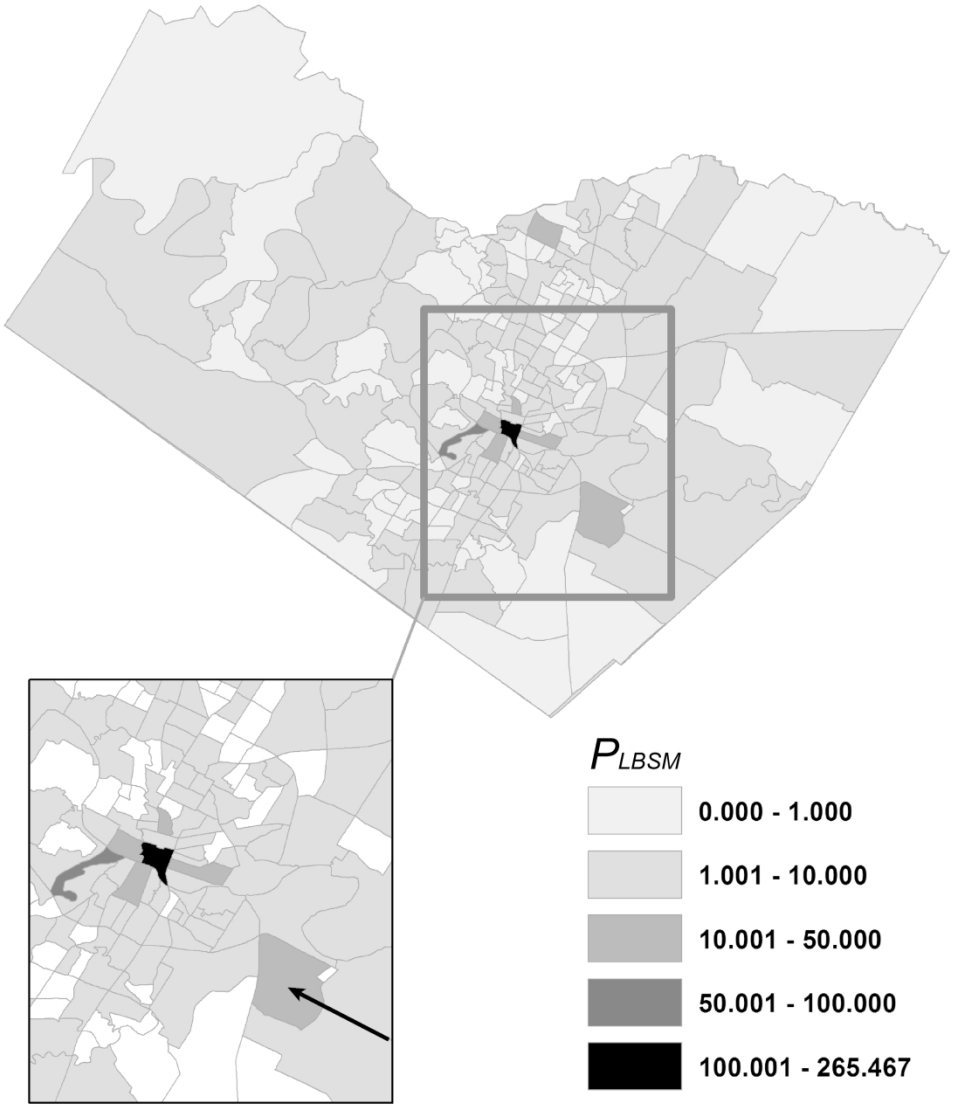

Figure 2. The distribution of P_{LBSM}.

246x279mm (300 x 300 DPI)

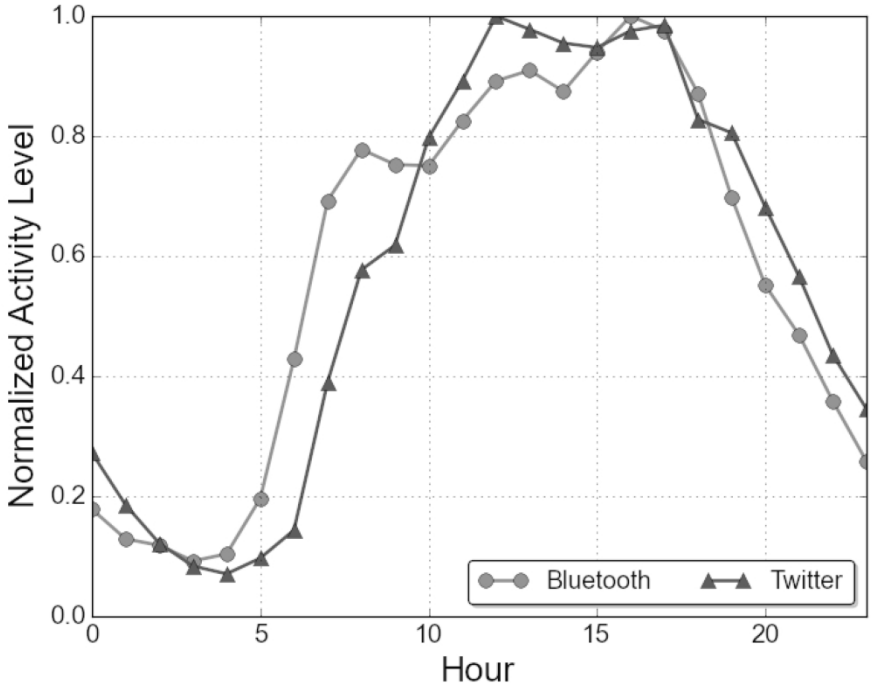

Figure 3. Twitter and Bluetooth hourly activity comparison.

203x152mm (300 x 300 DPI)

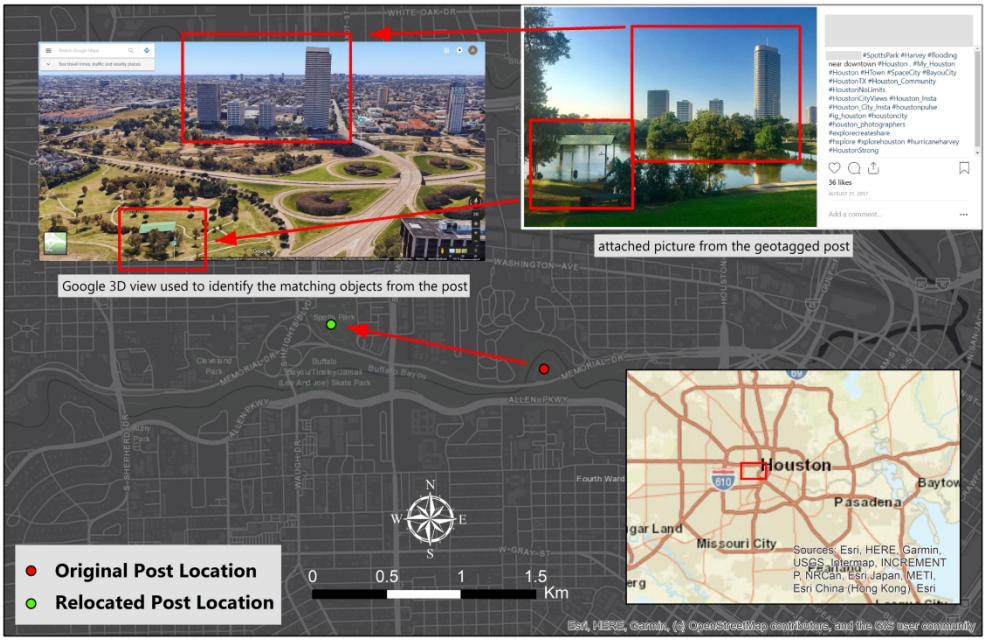

Figure 4. Example of an off-scene post with the inferred original and relocated post locations.

254x165mm (300 x 300 DPI)

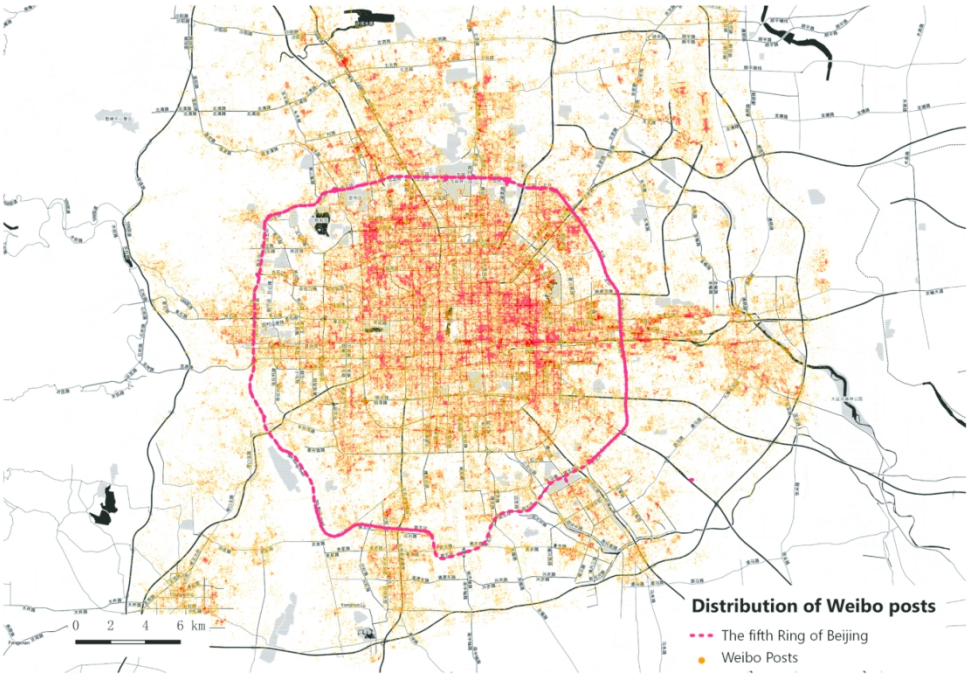

Figure 5. Spatial distribution of Weibo posts.

237x167mm (300 x 300 DPI)

Figure 6. Word Clouds (A) work; (B) daily life.

260x79mm (300 x 300 DPI)

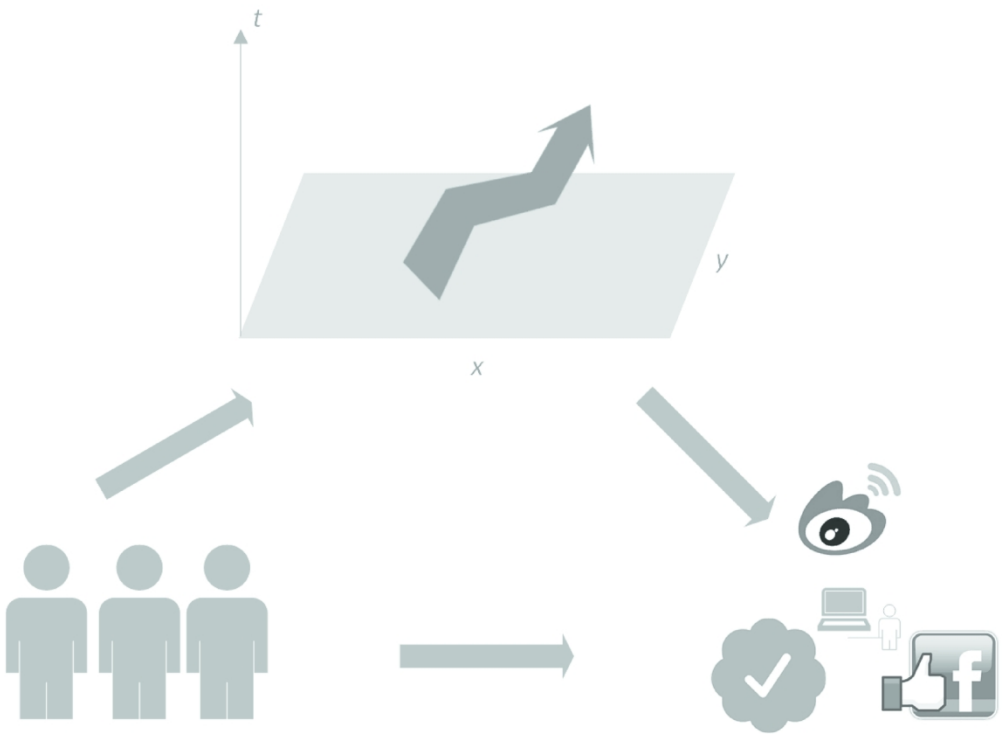

Figure 7. Interactions among the 4*W*s.

144x107mm (300 x 300 DPI)