Harnessing Heterogeneous Social Data to Explore, Monitor, and Visualize Urban Dynamics

Achilleas Psyllidis, Alessandro Bozzon, Stefano Bocconi and Christiaan Titos Bolivar

Abstract

Understanding the complexity of urban dynamics requires the combination of information from multiple city data sources. Besides traditional urban data, geo-localized social media provide human-generated content, which may reflect in (near) real time the activities people undertake in cities. The challenge is to devise methods and tools that enable the integration and analysis of such heterogeneous sources of information. Motivated by this, we developed *SocialGlass*, a novel web-based application framework to explore, monitor, and visualize urban dynamics. By deploying our platform in three real-world use cases, the paper elaborates on the benefits and limitations of integrating social media with related city datasets. It further shows how the inherent spatiotemporal, demographic, and contextual diversities of social data influence the interpretations of (dynamic) urban phenomena.

A. Psyllidis (Corresponding Author)
Chair of Hyperbody – Digitally-driven Architecture, Department of Architectural Engineering & Technology,
Faculty of Architecture and the Built Environment,
Delft University of Technology (TU Delft), 2628 BL, Delft, The Netherlands
Email: A.Psyllidis@tudelft.nl

A. Bozzon • S. Bocconi • C. Titos Bolivar
Web Information Systems (WIS), Department of Software and Computer Science,
Faculty of Electrical Engineering, Mathematics, and Computer Science
Delft University of Technology (TU Delft), 2628 CD, Delft, The Netherlands
Email: {A.Bozzon, S.Bocconi, C.TitosBolivar}@tudelft.nl

1. Introduction

Social data refer to data that are produced and/or processed by people [5]. Traditionally, such datasets, mainly in the form of census records, demographics, and spatial statistics, comprise one of the most substantial sources of information in urban analyses and planning. The interpretation of the respective datasets can provide insights into a variety of aspects pertinent to cities, such as income distribution, projections of population growth, socio-demographic inequalities, crime rate, and land-use allocation, among others. Traditional urban social data are essentially characterized by their relatively high quality, in terms of accuracy, completeness, validity, and general truthfulness of the information contained in the datasets. However, their timeliness and subsequent availability have significant limitations. These relate to their refresh rate, which spans from months to several years, and is largely based on costly and laborious surveys, as well as on-site observations. In this regard, they appear adequate for depicting static or semi-static urban phenomena, but fail to address the intricacies of city dynamics or the impact real-time activities have on diverse urban systems and citizens' daily life.

In recent years though, the term "social data" accommodates a wider range of resources with emerging importance for the urban context [6]. To this end, it now also includes human-generated content from smart devices and social media platforms. Contrary to traditional urban social data resources, these decentralized networks provide streams of information that can further be related to specific urban phenomena, such as pedestrian flows, human activities, and (near) real-time demographics. Therefore, speed differences between the conventional and emerging social data resources are quite apparent. Besides the additional temporal insights they bring, data from current Location-Based Social Networks (LBSNs) – such as Twitter, Instagram, and Foursquare – are further enriched with spatial attributes. The latter not only indicate the geographic properties of a particular space within a city, but also encompass semantics about the different activities that take place in it [13, 15]. However, these data sources also come with multiple limitations, due to the mismatch between their design purpose, and their application to urban analysis and/or planning. This often results in rather "noisy" data streams that obscure the extraction of meaningful information. Coupled with the diversities in technology penetration, they are further characterized by a multitude of personal, contextual, and cultural biases. How can we, thus, leverage the combined potential of both traditional and online urban social data sources? The challenge is to develop methods and tools that

allow such simultaneous combinations, so as to perform enriched urban analytics and, potentially, support city planning.

To address this challenge, the paper presents a novel web-based application framework – which we coined *SocialGlass*[1] – that integrates heterogeneous urban social data, enabling the exploration, monitoring, and visualization of city dynamics. By adopting a modular system-architecture in combination with semantic data description technologies, *SocialGlass* can simultaneously gather and accommodate data from multiple sources, as well as incorporate new ones with relatively low effort. *SocialGlass*, integrates (a) static and semi-static data, traditionally used by city authorities, pertinent to demographics (e.g. number of residents, gender and age distribution etc.), economic factors (e.g. income levels), and social statistics (e.g. crime rate); (b) real-time feeds from social media platforms, such as Twitter, Instagram, Flickr, Foursquare, and Sina Weibo; and (c) streaming data from sensors. Taking especially into account the contextual bias that characterizes the content retrieved from social media, *SocialGlass* further utilizes crowdsourcing and human computation techniques to perform reality verification, based on collective intelligence. With this, we additionally aim to address the challenge of actively engaging citizens in the collection, analysis, and sharing of urban data at the city scale, with relatively low costs.

The scope of this work is, first, to investigate the potential complementary value that (near) real-time social data can have in understanding the complex dynamics of cities. Further, it aims to empirically explore and critically compare the different insights provided by traditional urban data sources and contemporary social media, besides analyzing the intrinsic diversities of the latter, as regards the users and the context within which these data are generated. Finally, it discusses how the various biases may impact data interpretations by different stakeholders, such as planners, policy- and decision-makers. In this regard, the paper contributes:

1. A novel web-based set of tools and methods that allow the simultaneous integration and exploration of static, semi-static, and (near) real-time social and crowd-sourced data, with application to urban analytics.

2. Interactive web interfaces for monitoring and visualizing diverse city dynamics, such as pedestrian flows and real-time social activities (filtered

[1] http://social-glass.org

by user types, gender, age ranges, venue categories, ethnicities, and nationalities, in various cities across different countries).

3. A set of real-world use cases and experimental studies of the presented tools for empirically investigating different diversity levels of social media data in varied contexts, their sample biases, and their abilities to provide complementary insights to governmental stakeholders, planning authorities, and citizens.

The remaining of the paper is organized as follows: In Section 2 we outline the related work. Section 3 introduces the proposed method. Section 4 presents the architecture of the web-based tools and interfaces, comprising *SocialGlass*. Section 5 describes the application of *SocialGlass* instances to three real-world case studies, for understanding urban dynamics within different contexts. In Section 6 we discuss the results, identify benefits and limitations of the introduced methods, and elaborate on the potential usefulness of the platform for different city stakeholders. Ultimately, Section 7 summarizes the conclusions and indicates future lines of research.

2. Related Work

Motivated by the increasing pervasiveness of contemporary social media, there is a wealth of research works – placed within the emerging field of Urban Informatics [9], also often referred to as Urban Computing [10, 11, 16] – that investigate the potential added value of LBSN data and Call Detail Records (CDRs) in understanding various aspects of the urban environment. In the majority of these cases, researchers aim at studying such aspects by means of more scalable methods, as compared to conventional practices of urban data gathering (e.g. surveys, empirical studies through field observations etc.).

In their work, Quercia & Sáez-Trumper [15] examine the possibility of using data from Foursquare as an alternative source of information for understanding urban deprivation in different neighborhoods of the city of London. Based on results from relevant research studies, which identify a relationship between the provided facilities within a neighborhood and its corresponding deprivation index, the authors aimed at investigating whether this relationship is also detectable in data stemming from social media resources. Contrary to more conventional studies that perform analyses based on official land-use records, the authors conduct venue mapping by means of

Foursquare posts and, further, compare them with spatial statistics on deprivation. The article concludes that streaming social media feeds can more easily detect physical changes in the urban fabric, as compared to traditional land-use records, and thus assist in monitoring the fluctuation of a neighborhood's deprivation levels. In addition, they can provide fine-grained temporal insights into future predictions of urban deprivation.

Closely related to our paper's scope, the approach of Cranshaw et al. [7] focuses on discovering the diverse social compositions and dynamics within the city of Pittsburgh, PA, through the analysis of social media data. This analysis led to the identification of internal social sub-clusters within neighborhoods that would be difficult – if not impossible – to detect, solely through official census records, or would otherwise require rather costly and labor-intensive on-site observations. In a similar manner, Del Bimbo et al. [8] utilize geo-referenced data from Facebook and Foursquare to perform venue classification in a city, based on users' interest profiling.

The monitoring and analysis of human mobility through geo-located data streams have also recently been the focus of numerous relevant studies. Noulas et al. [12] used Foursquare data at a worldwide scale to explore citizens' mobility patterns in numerous metropolitan areas. In another study [13] they analyzed the distribution of human activities and identified urban sub-communities within cities, by means of place categories derived from a large-scale geo-referenced Twitter dataset (with links to Foursquare venues). Following a different approach, Amini et al. [2] gathered CDRs from mobile phones to explore and compare the influence social segregation has on human mobility patterns between developing and developed countries. Alhasoun et al. [1] also utilized CDRs to understand human commuting patterns at the city scale.

What the aforementioned studies share in common is the potential of easily accessible and geo-localized social media data in providing scalable solutions for exploring varied city dynamics. In addition, they all argue for the limitations traditional urban data gathering methods have in providing relevant insights at very large scales. However, these studies focus entirely on content stemming from a single data source. On the contrary, in our study we combine insights from several – municipal, sensor- and web-based – sources. We further consider the inherent diversities in social (media) data, by taking into account the various geographic, demographic, and contextual biases, instead of following a rather uniform approach.

3. Method

The emerging potential of social (web) data in understanding the urban environment, constitutes the driving force in our work. To this end, we focus our investigation on the following intrinsic dimensions of social data: (a) sample coverage, (b) timeliness, (c) integration of multiple sources, (d) veracity, and (e) visualization variability. Hereby, we investigate whether human-generated data can provide content-rich and trustworthy insights – thus, adequate for use by different stakeholders – into different aspects of cities.

With the objective of exploring, monitoring, and visualizing complex urban dynamics, we developed *SocialGlass* taking the aforementioned dimensions into account. The following paragraphs further elaborate on the approach adopted for each one of these parameters.

Sample coverage. Social media platforms are still used by a limited – yet perpetually increasing – amount of people. As such, the data derived from them cover, by definition, specific population groups (e.g. mainly younger populations, people with access to broadband services, platform popularity differs per cultural group etc.). To minimize the subsequent limitations and, thereby, attain broader sample representation, *SocialGlass* incorporates data from Twitter, Instagram, Foursquare, Flickr, and Sina Weibo. Besides, based on its modular architecture (described in Sect. 4), additional media platforms can be accommodated.

Timeliness. Different data refresh rates uncover equally diverse aspects of the urban environment and, further, cater for varied tasks performed by the multitude of city stakeholders. For instance, urban planners generally utilize a mixture of semi-static (e.g. demographics, real-estate records, geo-spatial statistics etc.) and near real-time datasets (e.g. energy consumption data). On the other hand, traffic managers are mostly interested in real-time data feeds from traffic monitoring resources. In this regard, *SocialGlass* encompasses: (a) static and semi-static datasets (e.g. gender and age demographics, income levels, crime rates etc.); (b) (near) real-time streams from social web platforms, discussed in the previous paragraph; and (c) real-time sensor feeds, distributed throughout the urban fabric.

Integration of multiple sources. The ingestion of various and rather heterogeneous social data sources subsequently raises interoperability issues. To this end, *SocialGlass* implements a semantic enrichment and integration component, backed by a developed domain ontology (aligned to a multitude

of external semantic models and structured vocabularies) [14]. Hereby, it formally describes the different data sets in a machine-processable way, for the purposes of content enrichment and data fusion.

Veracity. Social web data are generally characterized by information noise and high levels of untruthfulness. This often results in biased conclusions that may obscure decision-making processes. *SocialGlass* addresses this challenge by operating in the following two modes: (a) Social Sensing, which performs on-demand service requests to social media users, for data cleansing and linkage, employing crowdsourcing techniques, and (b) Human Computation, for active crowd engagement via dedicated platforms (in particular, CrowdFlower[2] and Amazon Mechanical Turk[3]).

Visualization variability. To serve the purposes of both data exploration and monitoring, *SocialGlass* respectively offers a map-based web platform, as well as a real-time city dashboard for interactive urban analytics. These interfaces allow stakeholders to make custom overlaps (in a layered fashion) across different visualization types that correspond to diverse information sources. The visualization types include dynamic point clusters, intensity heat maps, user paths, choropleth maps, and spatiotemporal graphs.

In the following Section we describe how the above-described dimensions were incorporated in the system architecture of *SocialGlass*. Their implementation and adaptation in real-world use cases are further discussed in Sect. 5.

4. System Architecture

The multiple heterogeneities of social data, in terms of accuracy, resolution, timeliness, scale, and speed pose several architectural challenges, towards the implementation of a scalable software system. To address these challenges, we employed a modular architecture, optimized for loose coupling and scalability. *SocialGlass* is structured around four main components, with each one comprising multiple modules that perform different functionalities (Fig. 1). The components respectively cater for (a) data ingestion and analysis, (b) semantic enrichment and integration, (c) exploration and visualization, and (d)

[2] http://www.crowdflower.com

[3] https://www.mturk.com/

real-time monitoring. In turn, each module is driven by the general functionality of the cluster it belongs to. The communication among them embarks on message queues. Thereby, the output of a specific module can be received, in the form of a message post, by any other module of the cluster. In this way, new data sources can be ingested in the system with relatively low effort, thus catering for issues pertinent to sample coverage.

In [5, 14] we provided detailed descriptions of the components, sub-systems, and modules of *SocialGlass*, along with their technical characteristics and individual functionalities. In this paper, we briefly outline the major features – by also linking them to the dimensions described in Sect. 3 – and introduce the real-time monitoring instance of the framework.

Data Ingestion and Analysis component. This particular component encompasses the functions of data acquisition, cleansing, enrichment, polishing, geo-localization, and spam/bot removal. It further comprises modules for Point of Interest (POI) mapping, demographic profiling (i.e. gender and age estimation of social media users), user role identification (e.g. residents, commuters, and foreign tourists), path extraction, and sensor data analysis. Current data sources include publicly available governmental records, CitySDK[4] open data, diverse social media platforms (Twitter, Instagram, Foursquare, Flickr, Sina Weibo), as well as various sensor data. In addition, it incorporates modules for crowdsourcing and human computation. Thereby, the data ingestion and analysis component addresses the dimensions of sample coverage, timeliness, and veracity.

[4] http://www.citysdk.eu

Fig. 1. System architecture of the *SocialGlass* platform

Semantic Enrichment and Integration component. This component simplifies the interconnections among the different data sources and providers. It specifically caters for tackling the several syntactic and semantic (spatial, temporal, and thematic) heterogeneities of the ingested data. To this end, it adopts Semantic Web technologies and Linked Data principles, by

means of a developed ontology and online tools for visualization of Resource Description Framework (RDF) data. Subsequently, the semantic enrichment component addresses the integration of multiple data sources.

Data Exploration and Visualization component. Users are given the possibility to perform interactive data exploration, along the multiple ingested sources, by means of an online, map-based environment. Information is visualized in a customizable-layered fashion. Layers can be initiated, deactivated, and organized according to the user needs. Further, they can be filtered down by data source (e.g. Twitter, Instagram, Flickr, CitySDK etc.), visualization type (e.g. choropleth maps, path routes, activity heat maps, dynamic point clusters etc.), and preferences (e.g. time span, user role, gender or age rage, ethnicity, venue category etc.). Visualization and analysis results can be exported and saved for further sharing or reuse. Thus, the data exploration component caters for visualization variability.

Real-Time Monitoring component. Complementary to the data exploration and visualization component, it offers (near) real-time dashboard tools for monitoring, analyzing, and assessing dynamic changes in the urban environment (e.g. city-scale events, traffic etc.). Through real-time "push" backend filtering and (near) real-time "pull" frontend data update, it offers a multitude of dynamic visualizations, based on real-time sensor and social media feed retrieval. This particular component addresses the dimensions of timeliness, veracity, and visualization variability.

5. Use Cases

In an attempt to explore the potential of combined social data in understanding urban dynamics, as well as for testing the flexibility of *SocialGlass* hereof, we deployed it in three real-world use cases. Each one of these studies refers to different urban contexts – in terms of both scale and morphology of the city fabric – diverse time periods, and serves different purposes. The first two cases (Milano Design Week 2014, and Amsterdam Light Festival 2015) aim to assess the impact of city-scale events on the patterns of human mobility. The third case comprises a comparative study of urban dynamics across 4 cities (Amsterdam, London, Paris, Rome) through the lens of heterogeneous social data. Our goal is to evaluate the insights provided by the approaches described in this paper and their potential added value for different city stakeholders (e.g. decision/policy-makers, urban

planners, citizens etc.). We further focus on the dimensions outlined in Sect. 3 and discuss the results in Sect. 6.

5.1 Milano Design Week 2014 (MDW 2014)

The Milano Design Week comprises an annually repeated event, accommodated in various venues throughout the city of Milan for the period of one week. The event has a twofold impact on the city's fabric. First, new temporary functions are allocated in a large number of existing venues (more than 500), which normally have a different land-use (e.g. parking lots, bars, and other public spaces temporarily turn into exhibition spaces). Second, the extensive amount of visitors – mainly arriving from abroad – differentiates the standard demographic composition of the city. Thereby, within a short time period, Milan's urban fabric undergoes several spatial and demographic changes. In this case, traditional urban data sources, such as land-use diagrams and census records are of no use, as they cannot indicate the various alterations that take place in the city. Also, on-site observations and surveys for analyzing the extent and impact of these changes would require substantial human and financial resources. Previous records pertinent to the event were also limited, as the amount of visitors, venues, and exhibition sites varies per year. Motivated by these challenges, we explored the potential of social media data in providing scalable insights into such dynamic urban changes.

We deployed *SocialGlass* during the entire event period and monitored real-time geo-referenced feeds from Twitter, Instagram, and Foursquare relating to it. Our purpose was twofold: first, to distill demographic information and user profiles of attendees, in order to enable a venue recommendation system; and second, to activate social media users for crowd sensing and data cleansing. In this context, the system collected 1.879.187 geo-referenced tweets and 1.090.237 Instagram posts, stemming in total from more than 30.000 unique users in the proximity of the event venues.

By analyzing the obtained data feed, we were able to spot significant variations in the demographic composition of the city, during the event [3]. Thereby, we may assert that social media can be beneficial in detecting such sudden alterations in the urban environment. In addition, media streams were integrated with data provided by the organizers (e.g. the program of the event). A semantic analysis of the social network activities contributed to the creation of links between the visitors and the venues, as well as to the development of event-related topical profiles of the attendees [4]. The latter

gave us the opportunity to provide meaningful recommendation to people, based on their individual interests. Though, in this case, we focused solely on event recommendation, results indicate potential future applications, for instance, in the domain of route recommendation systems.

5.2 Amsterdam Light Festival 2015 (ALF 2015)

The Amsterdam Light Festival refers to an art-related city-scale event with a longer duration, compared to that of MDW. In particular, its 2015 edition took place from November 27, 2014 to January 18, 2015. The event creates subtle interventions in the urban fabric (for instance, no changes in buildings' functions occur, as is the case in the MDW), in the form of art installations, scattered – yet also linked to one another through a network of planned routes – around the city. Our goal in this study was to assess the impact of the event on the city of Amsterdam across different user types, as well as to analyze the subsequent changes in human mobility patterns and visiting behaviors. How can we, thus, monitor and measure the dynamic alterations in these patterns in a scalable way? And how can we identify which ones are caused by the event in particular, and are not pertinent to other incidents in the city?

To address these challenges, we correlated data from several social media platforms – specifically Twitter, Instagram, and Foursquare – with publicly available municipal records that represent the standard demographic statistics of Amsterdam (e.g. gender distribution, age range, income rates etc.). The monitoring period – using social web sources – started two weeks before the beginning of ALF, continued throughout the entire period of the event, and finished two weeks after its end. The covered area included the wider region of Amsterdam, as well as the International Airport (Schiphol). In this way, we aimed to gain a better understanding of the event's influence on the city and the dynamics of visiting patterns. To this end, the system collected a total of 26,740,669 geo-referenced tweets (linked to Foursquare venues) and 15,959,566 Instagram posts.

In our analysis, we defined "visitors" as persons whose social media activity is detected in the proximity of the various ALF installations. We further defined "visits" as the corresponding activities of, for instance, creating or forwarding a post. Bearing these definitions in mind, we detected 15,345 visitors with 28,110 visits through Instagram, and 2,344 visitors with 5,876 visits through Twitter. This activity differentiation is probably pertinent to the strong visual character of the installations, which attract more photos

(Instagram) than microblog posts (Twitter). This overall count was further filtered down to address the total number of visitors and visit for each one of the 47 installations.

Yet, besides popularity it is important to gain insights into the diversity of visitors and their correlation to their visiting behaviors. Based on modules included in *SocialGlass* (e.g. demographic profiler, home locator, user role identifier etc.) we were able to extract real-time demographic information, with regard to user roles, as well as gender and age distribution. We defined three categories of users: *Residents* (home location within the Amsterdam area), *Local Tourists* (home location within the Netherlands, yet not in Amsterdam), and *Foreign Tourists* (home location in a country other than the Netherlands). Our analysis showed that *Residents* and *Foreign Tourists* were the ones who mostly visited the ALF event. In addition, *Foreign Tourists* tended to gather at popular venues, while *Residents* showed a more balanced distribution of visits across the sites. Visiting patterns also presented temporal diversities per user category. For instance, *Residents* had a more balanced activity throughout the event, while *Foreign Tourists* had much more fluctuating patterns (Fig. 2).

As regards demographics, both female and male visitors showed similar temporal patterns of visits. The same was observed for the different age groups. Yet, there were many fluctuations relating to individual installations (e.g. some artworks were much more appreciated by visitors within the age range of 31–45 years). Also, with the help of the Path Extractor module, we figured that the paths followed by *Foreign Tourists* were less diverse, compared to those of *Residents*.

Thus, our analyses showed that specific types of visitors demonstrate considerably distinctive behaviors, based on their social media activity. The additional comparison of social media information to municipal statistics also revealed some interesting insights into the spatiotemporal demographic changes, as a result of ALF. For instance, the administrative areas of Jordaan, Grachtengordel-West, and De Weteringschans are mainly characterized by a predominant male residential population. In addition, the largest age group is between 25 and 44 years old. Prior to ALF these were also the most predominant groups, in terms of social media activity. However, during ALF, female visitors were twice the amount of male ones, while the largest age group was between 15 and 24.

[a] ALF – Platform: Instagram | User type: Resident | Time Period: 18–21h

[b] ALF – Platform: Instagram | User type: Foreign Tourist | Time Period: 18–21h

Fig. 2. Activity heat maps of the ALF, illustrating the average spatial distribution of (a) *Residents*, and (b) *Foreign Tourists*, from 6pm till 9pm during the entire event period, based on their Instagram posts. In the Timeline we observe that *Residents*' activity is rather balanced, while *Foreign Tourists* present significant fluctuations, especially around the holiday season. In addition, *Foreign Tourists* appear to gather at popular venues, while *Residents*' activity is more spatially dispersed.

5.3 Comparing Urban Dynamics Across Four Cities Through Heterogeneous Social Media

We performed an experimental study using *SocialGlass* for a period of three weeks, across four European cities (namely Amsterdam, London, Paris, and Rome) combining together data collected from three different social media platforms (namely Twitter, Instagram, and Foursquare). Our goal was to investigate the potential of social media in demonstrating spatiotemporal activity patterns across different urban contexts. Moreover, we explored the influence of the intrinsic geographic, demographic, and contextual diversities that characterize social media on the conclusions drawn by the analyses. In this way, we aimed to assess the degree to which social media analysis may constitute a complementary approach to traditional urban analytics.

The selected cities vary significantly in terms of scale, total population, and morphology of the urban fabric. For that reason, they comprise challenging cases for testing whether these variations in physical space affect the social media activity, taking place in them. Geo-located data feeds were collected from February 20 to March 12, 2014. Neither city-scale events nor any national holiday occurred in this particular period, at least to our knowledge. Thereby, we aimed to lessen the extent of anomalies in activity patterns that would mislead the analysis interpretations. The final dataset comprises 933,272 Twitter posts (generated by 79,298 unique users), and 826,806 Instagram posts (generated by 118,514 unique users) in all four cities. The total 1,760,078 posts were mapped to Foursquare venues, using the corresponding API[5].

Through our analysis, we identified several differences in the usage patterns among the studied social media platforms, as well as in the location data each platform indicates. However, here we report on the findings pertinent to demographic and human mobility information extracted from the collected data feeds, as being more important for urban analysis (Fig. 3). Similarly to the ALF case (in Sect. 5.2), we classified users in three categories: *Residents*, *Commuters*, and *Foreign Tourists*. According to social media demographics, the predominant group of people in the city of Amsterdam is *Commuters*. Yet, the same city features the largest amount of *Residents*, in comparison to the remaining three. As expected, *Residents* in all four cities show larger spatial distribution of activities than *Foreign Tourists*,

[5] https://developer.foursquare.com/docs/venues/search

who tend to aggregate around city centers (Fig. 3). In addition, *Residents* of Amsterdam present a high mobility peak between 6pm and 9pm (average of the three-week monitoring period). After 10pm their activity density is significantly lower and start rising again between 6am to 9am. On the contrary, in London, Paris, and Rome the highest activity of *Residents* is observed after 8pm and further presents a constant rise up until midnight. Yet, more dense patterns are detected again after 8–9am. In this regard, citizens of Amsterdam can be considered as early-risers, while the inhabitants of the remaining three cities appear to go later to bed, but also wake up later. As regards *Foreign Tourists*, in London, Paris, and Rome they present denser mobility patterns after 9pm, while in Amsterdam their activity footprint is notably lower after 8pm.

In reference to the venues these activities relate to, both *Commuters* and *Foreign Tourists* appear more active in *Arts & Entertainment* than *Residents* in all four cities. *Foreign Tourists* are also consistently more active in the *Travel and Transport* land-use category. Mainly comprising venues, such as train stations, airports, and hotels it constitutes a rather expected result. Yet, seen from a different perspective, it also indicates that social media information has potential to accurately describe an anticipated reality. Traditional surveys would also reach the same conclusion, though in a less-scalable and rather costly way. We also observed that Twitter feeds in all cities contain much less information with regard to Points of Interest (POIs), as compared to Instagram posts. Subsequently location-based data stemming from Instagram are mainly focused on city attractions and landmarks (especially those appealing to *Foreign Tourists*), while those stemming from Twitter largely contain information about the kind of activities taking place in a venue. Subsequently, Instagram information is more important to stakeholders who perform analysis on tourist mobility (as we also observed that in all four cities, *Foreign Tourists* generate 10–30% more Instagram posts than Twitter ones). On the contrary, Twitter data provide more indications on everyday life in the city (most of the posts are generated by *Residents*). Finally, we detected that in all four cities, users between the ages of 16 and 30 present more dispersed activity patterns than the younger or older age groups, who appear more spatially concentrated (according to their social media footprint).

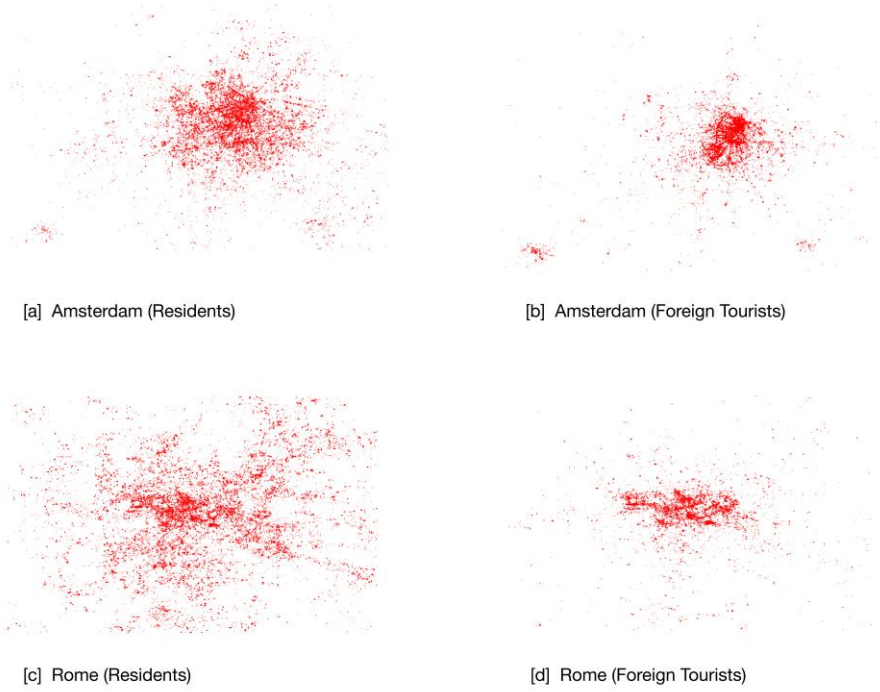

Fig. 3. Scatterplots illustrating the spatial distribution of *Residents* and *Foreign Tourists* in Amsterdam and Rome, based on their Twitter, Instagram, and Foursquare posts.

6. Discussion

Having empirically validated several instances of the *SocialGlass* platform in real-world use cases, we can critically discuss the benefits and limitations of our approaches to understanding urban dynamics. We paid particular attention to social web data and examined their potential contribution to city analytics. In the following paragraphs we elaborate on the results obtained from the application of *SocialGlass*, focusing on the dimensions outlined in Sect. 3.

Analyses, based on human-generated data from social media, are inevitably geographically, demographically, and contextually biased. Subsequently, *sample coverage* is strongly affected. Geographic biases can be overcome by the larger penetration of media technologies. Taking into account the increasing amounts of smartphone owners, as well as the growing percentages of social media users worldwide, we believe that in the near

future this particular limitation will drastically be decreased. Having tested the platform in cities belonging to developed countries in all our studies, we did not encounter significant obstacles in terms of technology penetration. However, we did observe anomalies in the representation rates of specific nationalities (such as Germans and Chinese). This is mainly due to media platform usage preference (e.g. Chinese people use Sina Weibo in place of Twitter), rather than technology penetration issues. On the other hand, demographic biases relate to the unequal representation of certain social groups in the data collected by web platforms. For instance, Twitter, Instagram, Foursquare, among others, are mainly used by younger populations, who are also more familiar with technological advances. Yet, each one of these platforms also presents intrinsic demographic diversities (e.g. Twitter is mainly used by young males, while Instagram is mostly preferred by females). It was also obvious in our studies that older populations were poorly represented. Stakeholders who may utilize such data (e.g. policy/decision-makers, urban planners etc.) in their tasks have to bear these limitations in mind. Contextual biases may pose an additional impediment to the interpretation of conclusions. We saw that city-scale events – such as MDW (Sect. 5.1) and ALF (Sect. 5.2) – have a strong influence over the dynamics of human mobility across the urban fabric. When aiming at studying everyday usage patterns in cities, monitoring has to exclude periods of planned events or related activities (as also discussed in Sect. 5.3) to avoid data anomalies.

Notwithstanding the aforementioned biases, *SocialGlass* provides more scalable methods to explore and monitor activity patterns – even in real-time (Sect. 5.1) – as compared to traditional surveys. It is capable of covering a large sample of city dwellers, within a single or multiple urban contexts (Sect. 5.3) simultaneously, with relatively low effort and costs. Though census records generally cover several social groups in a more balanced way than social media, they fail to address how these groups actually experience the city. Our platform, however, can provide enriched information about their day-to-day mobility habits.

Timeliness is another important factor, especially when exploring the dynamics of cities. In the cases of the MDW (Sect. 5.1) and ALF (Sect. 5.2) – where existing, up-to-date, and relevant data were completely missing – (near) real-time social media feeds constituted valuable sources in coping with the multiple alterations in the urban fabric (in terms of both temporary land-use allocations and visitor flows). We believe that such sources will also be highly

beneficial to traffic managers, by cross-validating camera and sensor observations with human-generated content.

However, cross-validation requires that we simultaneously integrate data stemming from multiple sources. In the paper, we indicated the several limitations of each social data source in the analysis of urban dynamics. Diversities in scale, speed, and semantics correspondingly illustrate different, yet partial, views of the city. We further identified, through our use cases, the constraints of single social media sources. Therefore, we argue that the *integration of multiple data sources* is essential towards more comprehensive urban analyses. The modular architecture of *SocialGlass*, in combination with its semantic integration tools, allow for such data fusion. In the presented case studies, we merged insights from different social media platforms together with municipal and/or governmental records, where available (the ALF case in Sect. 5.2 is an example of this). In addition, we aim to expand the sample coverage by integrating supplementary media platforms (such as Flickr and Sina Weibo) in future use cases. Stakeholders making use of *SocialGlass* are also given the opportunity to easily ingest datasets they frequently work with and combine them with the collected social media streams.

When working with human-generated data, *veracity* issues may further obstruct city authorities from making the right decisions. Though strongly associated with the biases mentioned previously, the untrustworthiness of social web datasets is also the result of their general "noisiness". We believe that the application of crowdsourcing and human computation techniques can be highly helpful, in this regard. However, crowd engagement still poses critical challenges to the performance of successful data cleansing and interpretation by people. Yet, crowdsourcing is a promising direction for actively involving citizens in the information loop process (i.e. data creation, polishing, and verification).

Finally, we ponder on the importance of *visualizing* the analyzed results in various different ways. As is the case with data integration, single modes of information visualization have very limited potential to address the needs of multiple stakeholders. Our platform accommodates multiple visualization possibilities that can further be overlapped, in a layered fashion. We believe that, in this way, *SocialGlass* will enable different city stakeholders to perform meaningful analyses, by overlapping data they frequently use (e.g. real-estate records, land-use, economic, energy data etc.), with insights from social media.

7. Conclusions and Future Work

This paper explored the potential of harnessing heterogeneous social data for understanding the dynamics of cities. It further introduced *SocialGlass*, a novel web-based application framework that provides methods and tools to integrate, analyze, and monitor multiple urban data sources (e.g. municipal records, social media feeds from different platforms, sensor data etc.). Based on its modular architecture, the system can easily incorporate custom datasets and additional components. The rationale behind the platform was structured around five dimensions of social data; namely, *sample coverage*, *timeliness*, *integration of multiple sources*, *veracity*, and *visualization variability*. Through the application of *SocialGlass* to three real-world use cases referring to different urban contexts, we empirically validated its capacities. We observed that *SocialGlass* sufficiently addresses the limitations of the aforementioned dimensions. Further, we showed that social media analyses better highlight the dynamic changes in cities (in both the built environment and the citizen flows), but can also complement existing urban data sources.

As part of future work, we aim to conduct studies with a focus on urban planning challenges (e.g. social segregation, livability of cities, cultural inequalities etc.), so as to better support planners and make further contributions to the field of urban informatics.

Acknowledgments. This research is funded by the Greek State Scholarships Foundation I.K.Y. (by the resources of the Educational Program “Education and Lifelong Learning”, the European Social Fund (ESF), and the EU National Strategic Reference Framework (NSRF) of 2007-2013, and by a scholarship of the Alexander S. Onassis Foundation. It is also financially supported by the Foundation for Education and European Culture (IPEP) and the A. G. Leventis Foundation. The EIT ICT Labs, and the Amsterdam Institute for Advanced Metropolitan Solutions (AMS) additionally provide partial funding. Last but not least, we would like to thank Jie Yang for his valuable contribution to the data analysis process.

References

Alhasoun F, Almaatouq A, Greco K, Campari R, Alfaris A, Ratti C (2014). The City Browser: Utilizing Massive Call Data to Infer City Mobility Dynamics. In: 3rd International Workshop on Urban Computing (UrbComp 2014). UrbComp: New York, NY

Amini A, Kung K, Kang C, Sobolevsky S, Ratti C (2014) The Impact of Social Segregation on Human Mobility in Developing and Industrialized Regions. EP J Data Science 3(6):1-20

Balduini M, Bocconi S, Bozzon A, Della Valle E, Huang Y, Oosterman J, Palpanas T, Tsytsarau M (2014) A Case Study of Active, Continuous and Predictive Social Media Analytics for Smart City. In: 5th Workshop on Semantics for Smarter Cities at the 13th International Semantic Web Conference (S4SC@ISWC 2014). Springer, Berlin–Heidelberg, pp 31–46

Balduini M, Bozzon A, Della Valle E, Huang Y, Houben GJ (2014) Recommending Venues Using Continuous Predictive Social Media Analytics. IEEE J Internet Computing 18(5):28 - 35

Bocconi S, Bozzon A, Psyllidis A, Bolivar CT, Houben GJ (2015) Social Glass: A Platform for Urban Analytics and Decision-making Through Heterogeneous Social Data. In: Gangemi A, Leonardi S, Panconesi A (eds) 24th International World Wide Web Conference (WWW 2015). ACM, New York, NY, pp 175–178

Ciuccarelli P, Lupi G, Simeone L (2014) Visualizing the Data City: Social Media as a Source of Knowledge for Urban Planning and Management. Springer, Berlin–Heidelberg

Cranshaw J, Schwartz R, Hong JI, Sadeh N (2012) The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In: Sixth International AAAI Conference on Weblogs and Social Media. AAAI, pp 58–65

Del Bimbo A, Ferracani A, Pezzatini D, D'Amato F, Sereni M (2014) LiveCities: Revealing the Pulse of Cities by Location-Based Social Networks Venues and Users Analysis. In: 23rd international conference on World Wide Web (WWW 2014). ACM, New York, NY, pp 163–166

Foth M, Choi JHJ, Satchell C (2011) Urban Informatics. In: Conference on Computer Supported Cooperative Work (CSCW 2011). ACM, New York, NY, pp 1–8

Kindberg T, Chalmers M, Paulos E (eds) (2007) Urban Computing [Special Issue]. J Pervasive Computing 6(3)

Kostakos V, O'Neill E (2009) Cityware: Urban computing to bridge online and real-world social networks. In: Foth M (ed) Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City. Hershey, New York, pp 196–205

Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A Tale of Many Cities: Universal Patterns in Human Urban Mobility. J PLoS One 7(5):e37027

Noulas A, Scellato S, Mascolo C, Pontil M (2011) Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In: Workshop on the Social Mobile Web at ICWSM 2011. AAAI, pp 570–573

Psyllidis A, Bozzon A, Bocconi S, Bolivar CT (2015, in press) A Platform for Urban Analytics and Semantic Data Integration in City Planning. In: G. Celani (ed) Computer-Aided Architectural Design Futures – New Technologies and the Future of the Built Environment. CCIS 527. Springer–Verlag, Berlin–Heidelberg, pp 1–16

Quercia D, Sáez-Trumper D (2014) Mining Urban Deprivation from Foursquare: Implicit Crowdsourcing of City Land Use. IEEE J Pervasive Computing 13(2):30-36.

Zheng Y, Carpa L, Wolfson O, Yang H (2014) Urban Computing: Concepts, Methodologies and Applications. ACM Transaction J Intelligent Systems and Technology (ACM TIST) 5(3):1-55