Diversity in Urban Social Media Analytics

Jie Yang(✉), Claudia Hauff, Geert-Jan Houben, and Christiaan Titos Bolivar

Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
{j.yang-3,c.hauff,g.j.p.m.houben,c.titosbolivar}@tudelft.nl

Abstract. Social media has emerged as one of the data backbones of urban analytics systems. Thanks to geo-located microposts (text-, image-, and video-based) created and shared through portals such as Twitter and Instagram, scientists and practitioners can capitalise on the availability of real-time and semantically rich data sources to perform studies related to cities and the people inhabiting them. Urban analytics systems usually consider the micro posts originating from within a city's boundary uniformly, without consideration for the demographic (e.g. gender, age), geographic, technological or contextual (e.g. role in the city) differences among a platform's users. It is well-known though, that the usage and adoption of social media profoundly differ across user segments, cities, as well as countries. We thus advocate for a better understanding of the intrinsic diversity of social media users and contents.

This paper presents an observational study of the geo-located activities of users across two social media platforms, performed over a period of three weeks in four European cities. We show how demographic, geographical, technological and contextual properties of social media (and their users) can provide very different reflections and interpretations of the reality of an urban environment.

Keywords: Social sensing · Urban analytics · User analysis

1 Introduction

A growing number of studies [3,12,20,21,29] have shown the potential of geo-located social media data (microblog posts, images, videos, etc.) as a relevant source to study the spatio-temporal dynamics of urban areas, e.g. a neighbourhood, a city, or an urbanised region. Platforms like Twitter, Instagram, and Sina Weibo can easily provide real-time access to a large volume of social data created by people from all walks of life. It therefore comes as no surprise that social media analysis is now a cornerstone of modern urban analytics solutions [17,30], and advocated as a fundamental component for decision making processes.

A common feature of prior work in this area is the belief that gathering large amount of data is sufficient to derive a detailed and *accurate* description (i.e. a description reflecting the reality) of the spatio-temporal properties of activities that occur within an area. All users contributing social media content are typically treated equally, independent of their origin (citizen or tourist), purpose in the city (e.g. resident or commuter), or demographic characteristics.

A. Bozzon et al. (Eds.): ICWE 2016, LNCS 9671, pp. 335–353, 2016.
DOI: 10.1007/978-3-319-38791-8_19

We argue that in order to draw correct conclusions (and subsequently base decision making on them), it is essential to investigate to what extent the *diversity* of the social media contributing population of users is an influencing factor. Diversity is an intrinsic property of social media platforms, and it is driven by complex socio-economic and socio-technical processes. As such, it represents both a challenge and an opportunity for urban analytics experiments and systems. When neglected, diversity hampers the generalisation and the validity of the obtained results, possibly leading to incorrect interpretations and subsequently erroneous courses of action.

Let us take as an example the case of a municipality monitoring social media to gather insights about the status of the city during a large street festival. The social media feeds may be overwhelmingly positive towards the event and thus the municipality builds a very favourable understanding of event as well. However, by neglecting the demographic distribution of users (or the lack of users from the target population), this understanding is not inclusive – it might miss the point of view of a significant share of the relevant population.

Diversity is not only a challenge though, it can also be a valuable source of information and should be exploited in order to understand, and leverage the intrinsic bias of social media data sources. Awareness of demographic distribution, for instance, helps in providing a sharper perspective, and an unbiased interpretation of an urban environment. Moreover, realising what (or whose point of view) is missing can lead to actions specifically devoted to bridging the existing knowledge gap, i.e. by means of targeted crowd-sourcing campaigns [7,8].

This paper aims to shed light on the entanglement that exists between social media platforms, their user populations, and the observations that can be obtained about an urban environment. We investigate four dimensions of diversity and seek an answer to the following overarching research question:

[RQ] How do *technological*, *geographical*, *demographic*, and *contextual* diversity impact the reflections of a city environment as perceived through social media?

We explore the influences of these factors in an experimental study performed by observing the social media activities in four different European cities (Amsterdam, London, Paris, and Rome), across two platforms (Twitter and Instagram), over a period of three weeks. We collected a dataset consisting of 1.87M of geo-located micro posts created by 198K of users. By employing state-of-the-art user modelling techniques offered by the `SocialGlass` platform [5,26], we were able to infer properties such as the gender, age, county and city of origin of social media users. We observed differences in population composition across cities and social media platforms, as well as diversity in the spatio-temporal properties of their online activities across roles (i.e. residents vs. tourists), genders, and ages.

Previous studies addressed differences in the composition of social media user populations across platforms and countries as general trends [23,25], or analysed the behaviour of social media users to identify spatio-temporal regularities in urban environments [21,22]. To the best of our knowledge, our work departs from previous efforts by being the first offering a principled analysis of the diverse spatio-temporal characterisations of user activities in urban environments that

include demographic and contextual aspects. We are not seeking to validate social media data, our goal is to highlight the diversities. Our findings prove the need for user modelling techniques in urban analytics, as a fundamental component in real-time social data processing pipelines designed for awareness, control, and prediction purposes.

The remainder of this paper is organised as follows. Section 2 presents an overview of the related work. Section 3 describes the applied experimental methodology and includes a set of specific research questions and hypotheses derived from our guiding question introduced in this section. We present our findings in Sect. 4 before turning towards an outlook into the future in Sect. 5.

2 Related Work

Urban Computing [2,3,16,17,30] is a consolidated area of investigation. More recently, the increasing pervasiveness of social media has led to a wealth of research works devoted to the creation of scalable solutions for exploring varied urban dynamics.

Several works have studied the behaviour of citizens by measuring spatio-temporal regularities in geo-located social media traffic for the purpose of event-detection [18], urban area characterisation [12,28], live-tracking and venue recommendation for city-scale events [3], city-scale [9] and global-scale [15] mobility patterns, and community detection [29]. All these works aimed to show that social media sources can be a good approximation to real human behaviour in cities, by measuring user-dependent indicators such as the number of tweets and/or users showing activities in a region of interest, in a specific period of time. [14] exploits geo-referenced data from Facebook and Foursquare to perform venue classification in a city, based on users interest profiling. [12] focuses on discovering the diverse social compositions and dynamics within the city of Pittsburgh, PA, through the analysis of social media data.

Recent studies also analysed the geographical and temporal variability of social media data when used for urban analytics purposes. [21] studied citizens' mobility patterns in Houston, San Francisco and Singapore by exploiting Foursquare data. An earlier study [22] used a large-scale geo-referenced Twitter dataset (with links to Foursquare venues) to identify urban sub-communities within cities. [20] studied the regional variability of categories of point of interests from the point of view of the temporal signature of social media activities.

Despite the abundance of previous works, research on the technical and contextual variability of spatio-temporal social media activities is currently lacking. Previous studies addressed the issue of social media population composition from an ethnographic [23], human computer interaction [25], or marketing perspective. We join the ongoing debate about the issues of big (social) data [4,11,13] by addressing the question of technical, geographical, demographic, and contextual diversity, and by providing evidence on intrinsic demographic and contextual bias in spatio-temporal social media activities.

3 Methodology

The goal of this paper is to investigate the impact of *geographic* (city), *demographic* (age, gender), *technological* (social media platform) and *contextual* (user role – visitor vs. citizen) factors on the perception of urban environments through social media. We are guided by the following three research questions:

RQ1: How does the choice of social media platform affect the spatio-temporal characterisation of an urban environment as observed through social media activities?

We hypothesise that significant differences between the amount and nature of social media activities across cities exist. We postulate an interplay between a social media platform of choice and a targeted urban environment. To test our hypothesis, we target several cities in the same period of time.

RQ2: How does the relationship of social media users with an urban environment affect the spatio-temporal characterisation of their social media activities?

The active population of a city is composed of people having different *roles* with respect to the urban environment. In addition to *residents*, cities temporarily host *local visitors* (i.e. people residing in a different city, but in the same country), and *tourists* (i.e. people residing in a different country). We hypothesise the amount and nature of social media activities to significantly differ across user roles in the targeted urban environment. To test our hypothesis, we infer the roles of our social media users with respect to the four investigated cities and explore the impact of this user partition in urban analytics.

RQ3: How do demographic properties of social media users affect the spatio-temporal characterisation of an urban environment as observed through social media activities?

We hypothesise that user attributes such as gender and age impact the spatio-temporal characterisation of a city. Analogous to **RQ2**, we explore this premise by partitioning our social media users according to these (automatically inferred) demographic properties.

To answer these questions, we crawled geo-located social media activities produced in **four cities** (Amsterdam, London, Paris, and Rome) over a period of **three weeks** (February 20th to March 12th 2014)[1]. For data gathering and exploration, we employ the `SocialGlass` platform, which will be introduced next.

3.1 Data Gathering and Pre-processing

Our study relies on data collected from Twitter and Instagram. We focus on geo-located content, i.e. micro posts augmented with explicit geographic coordinates either as measured by the localisation service of the user device, or inferred by the social network according to the IP address of the user.

[1] A motivation for this selection of cities and dates will be provided in Sect. 3.2.

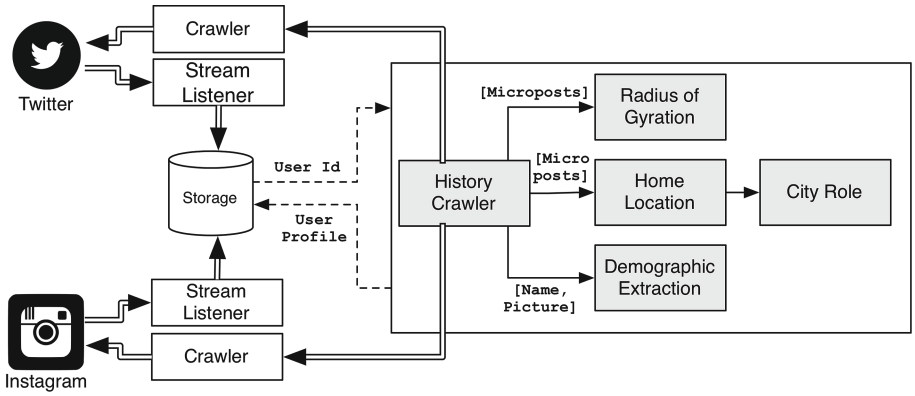

Fig. 1. Data gathering and pre-processing pipeline.

Figure 1 depicts the data gathering and pre-processing pipeline adopted in our study. The streaming APIs of each service push the content and metadata of each geo-located micro post produced in a given area of choice. Two listeners (one for each social network) monitor a stream each and store the content in a centralised repository. To support user-related analyses, the **History Crawler** retrieves, for each new user in the repository, all accessible historic content published by the user. Twitter's API allows the retrieval of the most recent 3,000 tweets of a user, while Instagram allows the retrieval of the entire post history. By taking into account the post history of a user, the **Home Location** module estimates the most likely home area, information which is often not provided in social network user profiles. We use a variation of the method described in [9] where, instead of a custom grid size we make use of geohashes. Once the coordinates of the estimated home location are identified, we use a reverse geocoding service (`Geo-names`) to determine the user's *city* and *country* of origin.

Functional to our study is the characterisation of the *role* a user plays in a city. The **City Role** module classifies users according to the relation of their (estimated) home city and the currently analysed urban area. We assign a user to one of the following three classes:

- *Resident*, if the user's home city is the same as the city under study;
- *Commuter*, if the user's home city is different from the city under study while both are in the same country; and
- *Foreign Tourist*, if the user's home location is in a different country from the city under study.

Lastly, the **Demographic Extraction** module estimates users' *gender* by means of a multi-modal decision tree classifier: starting from the profile picture and the name of a user, we combine the output of a state-of-the-art face detection and analysis component `Face++`[2] with the output of a dictionary-based

[2] `Face++`, http://www.faceplusplus.com.

gender recognition module[3], which consumes the home location of a user to disambiguate country-dependent names (e.g. "Andrea" can both be a male and female name, depending on the country of origin). We use `Face++`'s age classifier as-is and classify users into three age groups:

- *Young*: users between the ages of 15 and 30;
- *Middle-aged*: users between the ages of 31 and 45;
- *Older*: users above 45 years of age

We note that this age-based user grouping relies on younger ages than the literature of social and physiological science (e.g. [1,19]) as the use of social media is more familiar to the younger generations [10]. Indeed, in the *Older* age group, very few users are currently active on Twitter/Instagram compared to users of lower ages, yielding sparse data sources.

Whereas home location estimation based on textual evidence is an established research area with known high accuracies, this is yet the case in the inference of demographics from natural images. Thus, in order to determine the quality of the **Demographic Extraction** module we evaluate the face detection output as well as the age and gender classifiers on a manually labelled corpus of 628 culturally diverse Twitter user profiles. We find face recognition to be very precise in the identification of faces when present ($Precision = 98.5\,\%$). The moderate recall we observe ($Recall = 65\,\%$) is caused by profile images that portrait the user in a non-standard manner (partial visibility of the face, "artistic" image filters, etc.). Our gender classifier combines face analysis with name analysis and subsequently reaches a higher recall level, with a moderate drop in precision (85 %). Age detection has been performed only on profile pictures for which a face could be detected; there, we observed an age detection accuracy of 88 %.

3.2 Dataset

Our study focuses on four European cities: Amsterdam, London, Paris, and Rome. Their selection is motivated by their commonalities: they are (1) capitals of their respective countries; (2) popular touristic destinations, while being, at the same time, (3) characterised by a very vibrant business ecosystem, and by (4) the presence of a consistent and multi-cultural resident population. It is worth noting that the selection also reflects our intent of excluding profound cultural and economical differences, which might affect technological penetration and usage of social media. The four cities do exhibit differences which we find desirable for our analyses with respect to climate, geography (area and morphology) as well as size of the resident population.

Our window of data collection (February 20, 2014–March 12, 2014), which falls outside of national holidays or large-scale events in the target cities, was chosen to minimise the chances of spurious observations due to exceptional city usage anomalies. Both Stream Listener (cf. Fig. 1) were parameterised with

[3] `Genderize`, https://genderize.io.

the bounding boxes associated with each of the cities in GeoNames[4], thus ensuring that only micro posts from within the desired city boundaries are added to our storage repository.

We gathered 1.87 million micro posts from 198 thousand users across the three weeks of data collection. Table 1 summarises the amount of geo-located micro posts retrieved for each considered city in the indicated time window. The total number of Twitter users varies between 5,600 (Amsterdam) and 49,200 (London). The total number of micro posts varies between 53,100 (tweets in Amsterdam) and 498,700 (tweets in London). Although the short crawling time window does not allow us to gather millions of micro posts per city, we believe that our data set is robust (and diverse) enough to make our exploratory analysis generalizable across social networks, cities and users.

3.3 Metrics

To compare the frequency and nature of social media activities across cities and diversity factors, we adopt five common measures:

- the absolute number of social media activity performed in the time span of crawling (**#Posts**);
- the unique number of social media users that performed geo-located activities in the observation period (**#Users**);
- the temporal distribution of social media activities in a city, averaged on a 24 h span (**Time**);
- the *temporal diversity* calculated using the *Gini Coefficient* over the temporal distribution of social media activities (**#Gini.Temporal**); and
- the *spatial diversity* calculated using the *Gini Coefficient* over the geographical distribution of social media activities (**#Gini.Spatial**).

The *Gini Coefficient* is a measure of statistical dispersion, commonly used to measure inequality. It is beginning to be used as an important metric in urban analytics as well, e.g. [27]. A coefficient of zero indicates a uniform distribution, while a coefficient of one indicates maximal inequality among values. Intuitively, values at the extreme of the range are very unlikely when analysing social media activities.

4 Findings

We now discuss our findings, and focus on each of the following sections on one particular diversity aspect, while keeping the remaining variables fixed.

[4] For instance, the bounding box of the Amsterdam area is available at http://www.geonames.org/2759794/amsterdam.html.

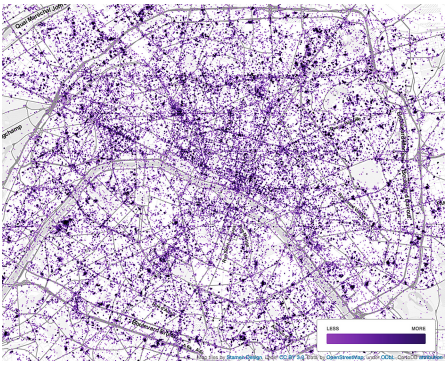

Fig. 2. User activities in Paris through the lens of Twitter.

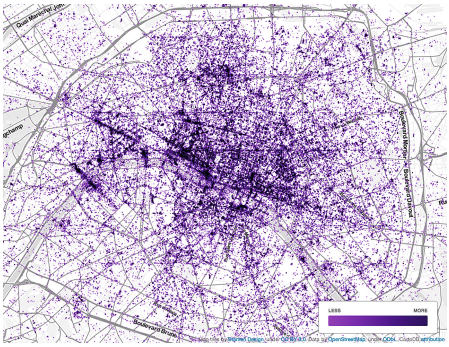

Fig. 3. User activities in Paris through the lens of Instagram.

4.1 Social Media Diversity

Let us first explore the influence the social media type has on city sensing (**RQ1**). Table 1 contains an overview of the number of users and posts gathered across the four cities on the two platforms. Not surprisingly, a city's size has a significant effect on the absolute number of users and postings made from within the city boundaries: Amsterdam has less than a million inhabitants while London has more than 8 million - this difference in scale can also be found in the total number of users across the two platforms: 17,222 users posted from within Amsterdam while 109,280 users posted from within the city boundaries of London.

Table 1. Overview of the data collected across two platforms and four cities during a three week period. #Users and #Posts are expressed in thousands.

	Amsterdam		Rome		Paris		London	
	Twit.	Inst.	Twit.	Inst.	Twit.	Inst.	Twit.	Inst.
#Users	6.6	10.6	5.7	15.9	17.8	32.4	49.2	60.0
#Posts	53.2	67.4	73.3	108.7	369.4	261.6	498.7	434.9
Gini.Spatial	.630	.772	.538	.615	.224	.499	.439	.588
Gini.Temporal	.255	.330	.313	.341	.310	.327	.286	.321
Inhabitants (in millions)	0.8		2.6		2.2		8.5	

Across the four cities, Instagram is slightly more popular, drawing a larger number of users than Twitter. This is natural, considering that Instagram is a platform focused on images more than on text, and the fact that all four cities are major tourist destinations. As will be covered in more detail later, based on

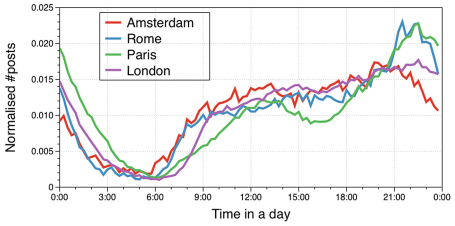

Fig. 4. User activities over time through the lens of Twitter.

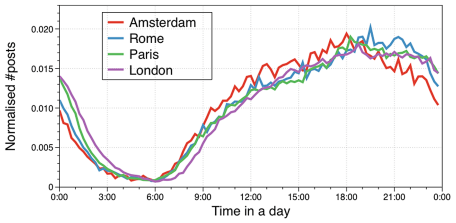

Fig. 5. User activities over time through the lens of Instagram.

Table 2. #Users for which demographic and role information are available. #Users and #Posts are expressed in thousands.

		Gender		Age			Role			All
		Male	Female	Young	Mid-aged	Older	Resi.	Loca.	Fore.	
Twitter	Amsterdam	3.0	2.1	1.9	1.3	0.4	2.2	2.0	2.0	6.6
	Rome	1.2	1.0	0.9	0.6	0.2	1.7	0.6	0.9	5.7
	Paris	6.2	6.0	5.3	2.2	0.4	8.2	4.6	3.7	17.8
	London	21.7	17.7	15.4	8.7	1.8	22.7	16.0	7.0	49.2
Instagram	Amsterdam	2.6	3.4	2.4	0.9	0.1	4.0	1.0	4.4	10.6
	Rome	1.1	1.8	1.4	0.5	0.1	3.1	0.9	2.0	15.9
	Paris	7.1	11.0	6.7	2.2	0.3	13.1	3.6	11.6	32.4
	London	14.3	21.0	13.0	4.4	0.5	28.7	10.5	12.8	60.0

the results in Table 4 it is also evident that Instagram has both more tourists and a higher ratio between tourists and residents than Twitter.

Gini.Spatial provides us with an interesting difference between the two platforms in terms of spatial distribution: Instagram posts are clustered spatially much closer together than Twitter postings; most strikingly in Paris. Figures 2 and 3 visualise the spatial distributions of user activities in Paris on Twitter and Instagram, respectively. While Twitter users do not show an obvious preference in posting locations, Instagram users are more in favor of tourism locations such as the Eiffel Tower and Les Champs-Elysees. In addition to the preference of Instagram users towards specific locations, it can also be observed that these users show a higher activity intensity at places close to public transportation hubs. While not shown, similar observations hold for the other cities.

Analogous to *Gini.Spatial*, *Gini.Temporal* shows that Twitter users post more evenly over time than Instagram users. To further inspect the difference of posting time on Twitter and Instagram, Figs. 4 and 5 depict the distributions of #posts over time for all 4 cities on Twitter and in Instagram, respectively. Overall, the distributions based on Instagram data are very similar to each other across the four cities, which is not the case for the Twitter-based distributions. Moreover, the Instagram-based distributions have a single mode, i.e. the time point with locally maximal activities, around 1800; while the Twitter

Table 3. Paris: top words within Instagram and Twitter posts at the most active hours.

Twitter@1200	`cest, paris, jai, fait, plus, trop, im, va, bien, faire, tout, bon, quand, vais, comme, ya.`
Twitter@2100	`cest, jai, paris, trop, plus, fait, bien, tout, va, quand, ya, mdr, demain, faire, bon, comme.`
Instagram@1800	`paris, love, france, fashion, beautiful, les, show, day, parisfashionweek, art, sunset, louvre, happy, sun, amazing.`

distributions have multiple modes, occurring at around 1200 and 2100. To further understand the cause of this difference, Table 3 reports the top words in social media posts at the these hours in Paris. In counting word occurrence, we remove the stop words in French and English, and transform all words into lowercase. While the top words in Twitter are all French words, the top words in Instagram, on the contrary, are in English. This clearly indicates that there are more tourist users using Instagram in Paris, which is also reflected in the semantics of their posts, which include terms such as `fashion, art, louvre, sunset` etc.

4.2 User Role Diversity

We hypothesised that the role of users w.r.t the city which accommodates their activities is important to be considered in urban analytics (**RQ2**). To investigate this premise, we report the statistics of users' social media activities partitioned by user role in Table 4 – note that we only consider those users in this analysis, for whom we were able to attribute gender, age and user role.

Table 4. Comparative statistics of social media activities across social media and user roles. #Users and #Posts are expressed in thousands.

		Amsterdam			**Rome**			**Paris**			**London**		
		Resi.	Comm.	Fore.	Resi.	Comm.	Fore.	Resi.	Comm.	Fore.	Resi.	Comm.	Fore.
Twitter	#Users	2.2	2.0	2.0	1.7	0.6	0.9	8.2	4.6	3.7	22.7	16.0	7.0
	#Posts	26.9	15.2	10.3	52.2	4.8	9.3	293.4	43.8	26.9	319.1	74.0	59.3
	Gini.Spa	.721	.435	.765	.502	.591	.745	.233	.233	.473	.394	.513	.610
	Gini.Tem	.242	.295	.266	.332	.304	.298	.327	.279	.290	.295	.330	.270
Instagram	#Users	4.0	1.0	4.4	3.1	0.9	2.0	13.1	3.6	11.6	28.7	10.4	12.8
	#Posts	30.3	6.4	26.8	45.4	7.5	21.3	123.7	19.3	93.6	252.3	45.3	92.8
	Gini.Spa	.753	.658	.837	.556	.707	.804	.465	.470	.573	.553	.633	.689
	Gini.Tem	.333	.368	.324	.335	.329	.382	.337	.339	.317	.321	.331	.315

Residents vs. Foreign Tourists: The distribution of Twitter users shows a clear pattern across all four cities: there are more Residents than Foreign Tourists; we find the largest observed difference in London (23K Residents and 7K Foreign Tourists) and the smallest in Amsterdam. With the exception of Amsterdam, the same observation can be made about our Instagram users.

We believe the reason for Amsterdam to behave differently is due to tourists' average length of stay in each city: this metric is shortest in Amsterdam[5].

Commuters vs. Foreign Tourists: the trends we observe are similar to those just described for Residents vs. Foreign Tourists, though their magnitude is lower. Instagram is more popular with Foreign Tourists than Twitter, which in general is more often used by Commuters – in terms of #Users and #Posts. The exception here is Rome. We hypothesise the difference to be due to different demographic distribution of the social media population. However, we lack sufficient information to support our hypothesis, for which a validation will be sought in future work.

Residents vs. Commuters: Interesting to note for these two user groups is the ratio of Commuters and Residents. This ratio is largest in Amsterdam and London, implying that those two cities attract more commuters than Paris or Rome.

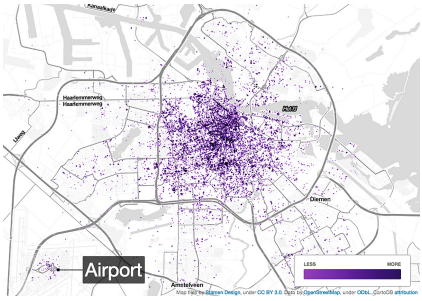

Fig. 6. Resident activities in Amsterdam through the lens of Instagram.

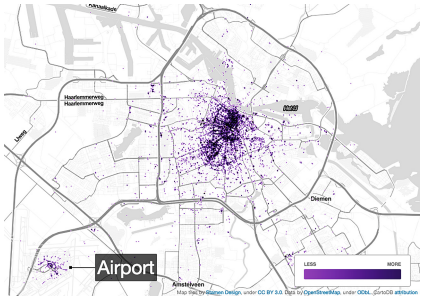

Fig. 7. Foreign Tourist activities in Amsterdam through the lens of Instagram.

Analysis of Gini.Spatial: Based on the *Gini.Spatial* measure computed over the different user roles, we notice that Foreign Tourists are more clustered in specific areas than the other user roles across all cities.

Figures 6, 7, 8 and 9 show the spatial distribution of activities for the three user roles and the example cities of London and Amsterdam. We hypothesise that Foreign Tourists are also active at airports and thus include Amsterdam's

[5] E.g. https://www.rolandberger.com/media/pdf/Roland_Berger_European_Capital_City_Tourism_20120127.pdf.

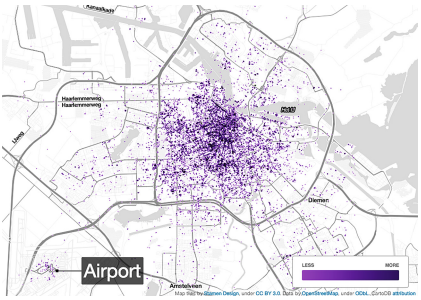

Fig. 8. Resident activities in London through the lens of Instagram.

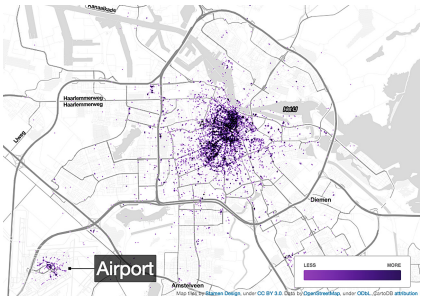

Fig. 9. Commuter activties in London through the lens of Instagram.

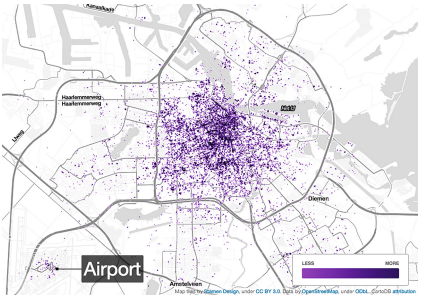

Fig. 10. Temporal distribution of *Resident* activities (Twitter).

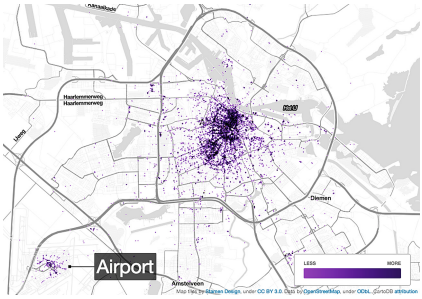

Fig. 11. Temporal distribution of *Commuter* activities (Twitter).

Schiphol Airport in the maps. The clusters of Foreign Tourist activities can be clearly distinguished, they are grouped around the city center and the major transportation hub (the airport), as compared in Figs. 6 and 7. Comparing the spatial distribution of Residents and Commuters we find in general the activity areas of Residents to be more balanced than those of Commuters, which can be shown in Figs. 8 and 9.

Analysis of Gini.Temporal: Based on the derived coefficients we find Commuters in Amsterdam and London to have the strongest preference of a posting time, compared to both Residents and Foreign Tourists. An explanation for this finding may be based on the premise that commuters post mostly *during* the commute. Rome and Paris, on the other hand, accommodate most of their workers inside the cities, resulting in a higher *Gini.Temporal* coefficient of Residents Figs. 10, 11 and 12 depict the temporal distributions of activities among the three user roles. Comparing the variation between cities in each of the graph, we conclude that Residents exhibit the largest variations among the four cities: Amsterdam Residents are the most (relatively) active during the day while least active at night; Residents in Paris and Rome, on the other hand, show a much higher number of activities at night. Though with less variation, similar patterns emerge for

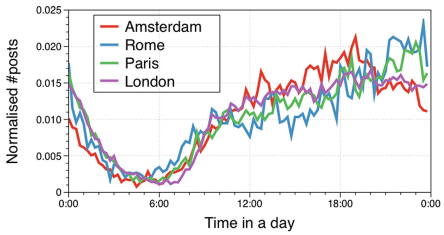

Fig. 12. Temporal distribution of *Foreign Tourist* activities (Twitter).

Foreign Tourists, indicating that each city possesses its individual 'gene', which can influence the temporal preference of Foreign Tourists' social media activities.

We also make the (sensible) observation that Commuters' activity frequencies decay earlier than those of Residents and Foreign Tourists alike — after 1800 most Commuters are likely to move towards their residential places outside the city boundaries. Interestingly, Commuters in Amsterdam have a dramatically higher (relative) number of activities than Commuters in other cities.

Table 5. Top words in Instagram posts of users with different user roles in Amsterdam.

Resident	`amsterdam, spring, day, happy, like, morning, fun, today, good, sunday, friends, food, one, holland.`
Commuter	`amsterdam, would, much, life, ever, person, win, blessed, chance, thx, amstelveen, goed, centraal.`
Foreign Tourist	`amsterdam, love, holland, netherlands, travel, canal, europe, happy, amazing, city, good, morning, museum.`

Table 5 lists the most frequent words among users in Amsterdam according to their designated roles (and on the Instagram platform). Clear differences emerge: Residents' posts concern their daily lives (`today, fun, friends` etc.), Commuters' posts are more high-level (`life, ever`) and refer to locations (`amstelveen, centraal`), while Foreign Tourists' posts – as expected – revolve around traveling (`holland, travel, europe`), and tourist attractions (`canal, museum`).

4.3 User Demographic Diversity

Lastly we turn to an exploration of **RQ3**, that is, the effect of user demographics on their social media activities in the city.

Table 6. Comparative statistics of social media activities across user genders. #Users and #Posts are expressed in thousands.

		Amsterdam		Rome		Paris		London	
		Male	Female	Male	Female	Male	Female	Male	Female
Twitter	#Users	3.0	2.1	1.2	1.0	6.2	6.0	21.7	17.7
	#Posts	18.0	13.5	18.1	16.7	89.7	98.0	187.2	157.4
	Gini.Spatial	.624	.615	.501	.564	.233	.221	.444	.419
	Gini.Temporal	.274	.305	.308	.329	.327	.297	.291	.291
Instagram	#Users	2.6	3.4	1.1	1.8	7.1	11.0	14.3	21.0
	#Posts	13.6	18.5	12.6	20.3	50.8	78.1	88.7	135.5
	Gini.Spatial	.779	.770	.659	.634	.500	.499	.601	.577
	Gini.Temporal	.335	.348	.338	.349	.319	.343	.319	.331

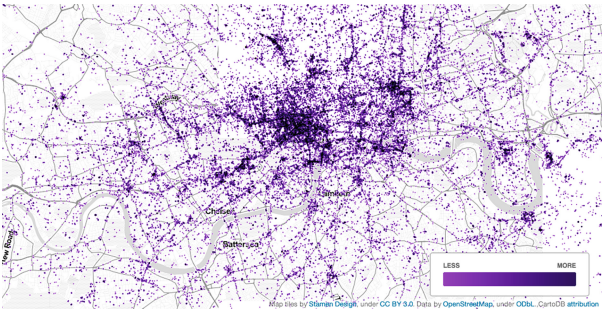

Fig. 13. *Male* user activities in London (Instagram).

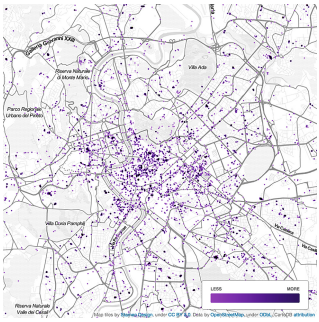

Fig. 14. *Male* user activities in Rome (Twitter).

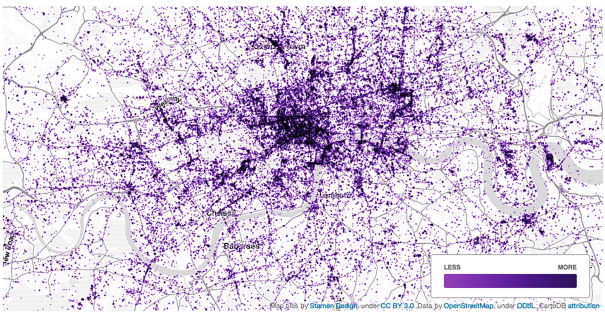

Fig. 15. *Female* user activities in London (Instagram).

Fig. 16. *Female* user activities in Rome (Twitter).

Diversity of User Gender. In Table 6 we report the by now familiar metrics separately for Male and Female users. Across all cities, there are more Male than Female users on Twitter, and more Female than Male users on Instagram,

indicating a clear and divergent preference of social media portals. In terms of #Posts this trend is especially strong on Instagram.

Gender also plays a role in the spatial dimension of social media activities. Looking at the Gini.Spatial of user activities in Instagram, one could observe that the social media activities of Male users are in general more geographically clustered. This holds in all four cities, most evidently in London and Rome. Similar phenomenon can also be found in Twitter, with an exception in Rome, where it can be found that the activities of Female users are more clustered than that of Male users. Figures 13, 14, 15 and 16 visualise the different cases in London through Instagram, and Rome through Twitter.

Looking at *Gini.Temporal*, we observe no obvious pattern within our Twitter users, while the Instagram data shows that the temporal activities of Male users are more evenly distributed than that of Female users. Overall, the above observations suggest that Male and Female users are distinct in their social media activities in cities, which calls for a careful separation of user genders in relevant research.

Diversity of User Age. Lastly, we analyze the diversity of social media activities of users in different age groups. As stated before, we classify users into three categories: *Young*, *Mid-adged* and *Older*. As expected, we find from Table 7 that Young users use social media more than Mid-aged users, who in turn use it more than Older users across all cities and both platforms.

Table 7. Comparative statistics of social media activities across user ages (Young, Mid-aged, and Older). #Users and #Posts are expressed in thousands.

		Amsterdam			Rome			Paris			London		
		Y	M	O	Y	M	O	Y	M	O	Y	M	O
Twitter	#Users	1.9	1.3	0.4	0.9	0.6	0.1	5.2	2.2	0.4	15.4	8.7	1.8
	#Posts	12.6	9.6	2.7	17.0	9.4	2.8	99.6	30.0	5.1	149.0	77.4	17.0
	Gini.Spa	.614	.611	.654	.514	.533	.590	.234	.254	.308	.415	.479	.458
	Gini.Tem	.265	.312	0.333	.322	.316	.357	.311	.323	.318	.290	.288	.307
Instagram	#Users	2.4	0.9	0.1	1.4	0.5	0.1	6.7	2.2	0.3	13.0	4.4	0.5
	#Posts	13.8	5.4	0.9	15.9	6.5	9,7	52,3	18,1	2,8	87,9	29,9	3,6
	Gini.Spa	.790	.778	.813	.661	.679	.729	.502	.514	.474	.589	.605	.606
	Gini.Tem	.334	.353	.449	.351	.364	.363	.332	.334	.340	.324	.332	.350

Gini.Spatial shows for three of the four cities and both platforms that Older users have a more clustered area of social activities. An example is presented in Figs. 17, 18 and 19, where we have visualised the Instagram user activities of the three age groups for the city of Amsterdam. The tendencies for Mid-aged users to be more clustered than Young users exists as well, however, the contrast is less stark. Turning to *Gini.Temporal* we find that Young users post tend to post slightly more evenly over time than Older users who have a work-life schedule to adhere to.

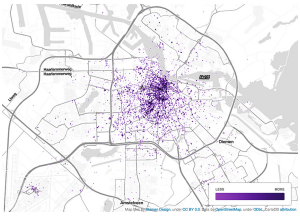

Fig. 17. *Young* user activities in Amsterdam (Instagram).

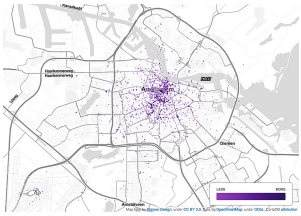

Fig. 18. *Mid-aged* user activities in Amsterdam (Twitter).

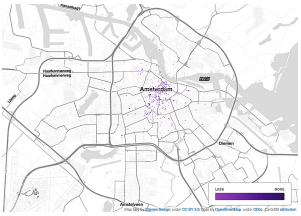

Fig. 19. *Older* user activities in Amsterdam (Instagram).

4.4 Threats to Validity

We have taken a set of users and their social media micro posts and partitioned them along several dimensions to determine the effect (if any) on the urban environment as perceived through social media.

In our exploratory work, we have considered four cities across a period of three weeks, leading to a dataset of 200K users and nearly 2 million posts. These four cities were carefully selected to be different in some dimensions but not others, leading us to believe that the differences we observed will only increase with a wider choice of urban areas.

The main limitation (and threat to validity) of our work is the dataset size: millions of posts are generated on popular social medial portals within a single day. However, on the one hand the 3 week continuous time period is selected in data crawling to exclude large-scale events and national holidays, thus minimises the potential bias; one the other hand, as became evident in this paper, an in-depth and thorough analysis of the various dimensions of interest (contextual, demographic, technological and geographical) can only be conducted for a small number of cities within the scope of a paper.

A second threat that needs to be acknowledged is the exclusive use of geo-located micro posts, which form a small minority among all created micro posts on the social media platforms today. We cannot guarantee that our findings also hold in exactly the same manner for the set of non-geo-located posts. There may be (small) deviations.

Finally, we acknowledge the study to rely on user modelling techniques (i.e. to infer gender, age, and user role) that can be applied only when user profile information is available, and have limitations in terms of accuracy [5,6,24]. While an analysis of the performance of such techniques is outside the scope of this paper, we stress our reliance on state-of-the art and state-of-the-practice solutions.

5 Conclusions

With the increasing value of social media data in urban analytics and decision making, the need for an accurate reflection and representation of the urban environment through the lens of social media is becoming stronger every day.

While past works have commonly treated all social media users in an urban area as one and the same, we have shown that by focusing on different user segments, different reflections of the urban environment surface.

This diversity of users and their impact on urban analytics should not be treated as a problem however. On the contrary, diversity can play a fundamental role analyzing and understanding urban phenomenons. In this paper we have analysed the influence of multiple diversity factors anchored in technology, geography, demographics and context. In future work we plan to expand our analyses across a larger set of users, a larger set of cities and greater urbanised regions across continents as well as across a wider range of diversity factors.

Acknowledgements. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

References

1. Al-Zahrani, M.S., Bissada, N.F., Borawski, E.A.: Obesity and periodontal disease in young, middle-aged, and older adults. J. Periodontol. **74**(5), 610–615 (2003)
2. Balduini, M., Bocconi, S., Bozzon, A., Valle, E.D., Huang, Y., Oosterman, J., Palpanas, T., Tsytsarau, M.: A case study of active, continuous and predictive social media analytics for smart city. In: Proceedings of the Fifth International Conference on Semantics for Smarter Cities, S4SC 2014, vol. 1280, pp. 31–46, Aachen, Germany (2014). CEUR-WS.org
3. Balduini, M., Bozzon, A., Valle, E.D., Huang, Y., Houben, G.: Recommending venues using continuous predictive social media analytics. IEEE Internet Comput. **18**(5), 28–35 (2014)
4. Bernaschina, C., Catallo, I., Ciceri, E., Fedorov, R., Fraternali, P.: Towards an unbiased approach for the evaluation of social data geolocation. In: Proceedings of the 9th Workshop on Geographic Information Retrieval, GIR 2015, pp. 10:1–10:2. ACM, New York, NY, USA (2015)
5. Bocconi, S., Bozzon, A., Psyllidis, A., Titos Bolivar, C., Houben, G.-J.: Social glass: a platform for urban analytics and decision-making through heterogeneous social data. In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015 Companion, pp. 175–178 (2015)
6. Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S., Ratti, C.: Choosing the right home location definition method for the given dataset. In: Liu, T.-Y., et al. (eds.) SocInfo 2015. LNCS, vol. 9471, pp. 194–208. Springer, Heidelberg (2015). doi:10.1007/978-3-319-27433-1_14
7. Bozzon, A., Brambilla, M., Ceri, S., Mauri, A., Volonterio, R.: Pattern-based specification of crowdsourcing applications. In: Casteleyn, S., Rossi, G., Winckler, M. (eds.) ICWE 2014. LNCS, vol. 8541, pp. 218–235. Springer, Heidelberg (2014)
8. Bozzon, A., Fraternali, P., Galli, L., Karam, R.: Modeling crowdsourcing scenarios in socially-enabled human computation applications. J. Data Seman. **3**(3), 169–188 (2013)
9. Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring millions of footprints in location sharing services. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 81–88. AAAI (2011)

10. Correa, T., Hinsley, A.W., De Zuniga, H.G.: Who interacts on the web?: the intersection of users personality and social media use. Comput. Hum. Behav. **26**(2), 247–253 (2010)
11. Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W., Zook, M.: Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. Cartography Geogr. Inf. Sci. **40**(2), 130–139 (2013)
12. Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N.M.: The livehoods project: utilizing social media to understand the dynamics of a city. In: Breslin, J.G., Ellison, N.B., Shanahan, J.G., Tufekci, Z. (eds.) Proceedings of the 6th International AAAI Conference on Weblogs and Social Media. The AAAI Press (2012)
13. Boyd, D., Crawford, K.: Critical questions for big data. Inf. Commun. Soc. **15**(5), 662–679 (2012)
14. Del Bimbo, A., Ferracani, A., Pezzatini, D., D'Amato, F., Sereni, M.: Livecities: revealing the pulse of cities by location-based social networks venues and users analysis. In: Proceedings of the 23rd International Conference on World Wide Web, WWW 2014 Companion, Republic and Canton of Geneva, Switzerland, International World Wide Web Conferences Steering Committee, pp. 163–166 (2014)
15. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C.: Geolocated Twitter as proxy for global mobility patterns. Cartography Geogr. Inf. Sci. **41**(3), 260–271 (2014)
16. Kindberg, T., Chalmers, M., Paulos, E.: Guest editors' introduction: urban computing. IEEE Pervasive Comput. **6**(3), 18–20 (2007)
17. Kostakos, V., O'Neill, E.: 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. Information Science Reference, Hershey, Philadelphia, USA (2008)
18. Lee, R., Sumiya, K.: Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN 2010, pp. 1–10. ACM, New York, NY, USA (2010)
19. Malatesta, C.Z., Izard, C.E., Culver, C., Nicolich, M.: Emotion communication skills in young, middle-aged, and older women. Psychol. Aging **2**(2), 193 (1987)
20. McKenzie, G., Janowicz, K., Gao, S., Gong, L.: How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. Comput. Environ. Urban Syst. **54**, 336–346 (2015)
21. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C.: A tale of many cities: universal patterns in human urban mobility. PLoS ONE **7**(5), e37027 (2012)
22. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In: The Social Mobile Web, Papers from the ICWSM Workshop, Barcelona, Catalonia, Spain, 21 July 2011
23. Palfrey, J., Gasser, U.: Born Digital: Understanding the First Generation of Digital Natives. Basic Books Inc., New York (2008)
24. Paraskevopoulos, P., Palpanas, T.: Fine-grained geolocalisation of non-geotagged tweets. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 105–112. ACM (2015)
25. Pater, J.A., Miller, A.D., Mynatt, E.D.: This digital life: a neighborhood-based study of adolescents' lives online. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 2305–2314. ACM, New York, NY, USA (2015)

26. Psyllidis, A., Bozzon, A., Bocconi, S., Titos Bolivar, C.: A platform for urban analytics and semantic data integration in city planning. In: Celani, G., Sperling, D.M., Franco, J.M.S. (eds.) Computer-Aided Architectural Design Futures. The Next City - New Technologies and the Future of the Built Environment. CCIS, vol. 527, pp. 21–36. Springer, Heidelberg (2015)
27. Tranos, E., Nijkamp, P.: Mobile phone usage in complex urban systems: a space-time, aggregated human activity study. J. Geogr. Syst. **17**(2), 157–185 (2015)
28. Wakamiya, S., Lee, R., Sumiya, K.: Crowd-based urban characterization: Extracting crowd behavioral patterns in urban areas from Twitter. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, pp. 77–84. ACM, New York, NY, USA (2011)
29. Wang, Z., Zhou, X., Zhang, D., Yang, D., Yu, Z.: Cross-domain community detection in heterogeneous social networks. Pers. Ubiquit. Comput. **18**(2), 369–383 (2014)
30. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. **5**(3), 38:1–38:55 (2014)