City Usage Analysis using Social Media

Master's Thesis

Christiaan Titos Bolivar

City Usage Analysis using Social Media

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK INFORMATION ARCHITECTURE

by

Christiaan Titos Bolivar
born in Schiedam, The Netherlands

Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

© 2014 Christiaan Titos Bolivar.

City Usage Analysis using Social Media

Author: Christiaan Titos Bolivar
Student id: 4009010
Email: c.titosbolivar@gmail.com

Abstract

Over the last few years, social media have become part of the daily life of many people, leading scientists to study their users and the data they produce in numerous contexts. For instance, geo-enabled social media provide us with the means to study the dynamics and features of large geographical areas. In this thesis, our goal is to leverage social media to study cities, and their usage by people of different origin (e.g. citizens vs. tourists) and demographics.

We design and implement a system that uses Twitter and Instagram as data sources, defining and extracting several features about the city and its users, such as finding points of interests, paths, differentiating users in gender, age, and their role in the city. We also build a proof of concept visualization tool that allows non-scientific users to analyze a city using our extracted data.

The system is used for an in-depth analysis, where we compare the usage, as observed through the lens of social media, of cities like Amsterdam, London, Paris, and Rome over a three week period on both Twitter and Instagram. We show that, through social media, it is possible to observe differences in usage patterns, both in the temporal sense, but also in regards to the places that are visited in the city.

Thesis Committee:

Chair:	Prof. dr. ir. G.J.P.M. Houben, Faculty EEMCS, TUDelft
University supervisor:	Dr. ir. A. Bozzon, Faculty EEMCS, TUDelft
Committee Member:	Dr. ir. M.B. van Riemsdijk, Faculty EEMCS, TUDelft

Preface

In front of you is the result of my master's thesis, which I have worked on over the past 10 months, here at the Delft University of Technology. After three years doing my Bachelor, and then my Master for two years, it comes down to this.

During my life I always spent much of my time on the internet, which could only lead to me doing this project at the end of my student career at the Web Information Systems department. Over the years I became more and more interested in the "hidden" data on the world wide web, and the gold mine that is social media. Thus I am glad I chose this topic for my thesis, and I have gained many valuable insights.

I would like to take this opportunity to thank my supervisor Alessandro Bozzon, whose knowledge, critical thinking, feedback, and sheer enthusiasm for the project have kept me going, and kept me motivated throughout. I also thank Stefano Bocconi, who helped me countless times on the technical side of things (darn virtual machines). The meetings with Stefano en Alessandro were always very pleasant. I would go in with a few questions, and leave with many more ideas.

The bi-weekly Omnicron meetings were also very valuable, and thus I would like to thank Claudia Hauff and my fellow students for their much appreciated feedback. Also thanks go to Geert-Jan Houben and Birna van Riemsdijk for being on my thesis committee.

Christiaan Titos Bolivar
Delft, the Netherlands
August 20, 2014

Contents

List of Figures

Chapter 1

Introduction

Urban analysis has been a popular research area for many years. Learning about the dynamics and demographics of a city is of vital importance for numerous fields, such as transport and infrastructure, politics and policy-makers, but also marketing. Typically, the data used for this kind of analysis comes primarily from surveys, and thus is more difficult to gather.

Our aim is instead to provide the means to understand to which extent social media can be used as a data source. Location based social networks and geo-enabled social media are more popular then ever. Users can now not only share a text message or picture, but their location as well. This publicly available data can provide us with many insights in how, why and when people move and what they visit. While the percentage of people who use social media is relatively low, and geo-tagged content of those posts also is but a small percentage, it is shown in previous research that geo-enabled social media are in fact a good representation of actual human behavior.

Previous work regarding city analysis has been done with geo-enabled online services such as Foursquare, Flickr and Twitter. Over the last few years another social platform has gained in popularity, namely Instagram. Instagram is a mobile app where users can take pictures or short videos, easily edit them (adding filters for example) and then share them online. Research using Instagram has been quite limited so far, though it is shown Instagram has a great potential in the field of urban analysis [24].

In this work we will use Twitter and Instagram as data sources, as they are two of the most popular micro-blogging tools, and the data is publicly available. Both tools are also used in "normal life", meaning that as opposed to a service such as Flickr, people post on these platforms all day long and the nature of the content is varied. This makes these tools naturally suitable to use in a urban analysis context.

1.1 Research Questions

We define several research questions to guide our research. The main research question of this project is defined as:

> How can geo-enabled social media be used to characterize the usage of a city during some time period?

This question boils down to "who goes where and when". We are interested in movement, characterizing users, and characterizing areas. In order to answer this question, we define several sub-research questions

1. How can social media be used to create (live) demographics of people and places and their relationships?
2. In what ways do different social media differ in regards to location data?
3. Can we find differences in city usage between cities using social media data?

1.2 Contributions

The main contribution of this work is two-fold. We design and implement an extensible system, which has an offline part that collects social media posts and users, and calculates various demographics and mobility features. The online part of the system visualizes the collected data, which allows end users to analyze and research the city. This tool will provide decision makers with several different views of their city, which enables them to help answer research questions of their own.

Previous research in the field of urban analysis using social media, is heavily based on Twitter and Foursquare, and more geared toward social-graph characteristics, such as the number of friends user have and the relation between that and city usage. In regards to users, we focus more on demographics and activities (what do they visit). We also use Twitter and Instagram as data sources, and only use Foursquare to discover points of interests in cities.

We also perform an extensive analysis where we compare the usage of four cities, Amsterdam, London, Paris, and Rome. For this analysis we collected posts from Twitter and Instagram over a 3 week period. We aim to use our system to discover differences between the four cities in regards to user activity, venue activity and paths, as well as differences between Twitter and Instagram

1.3 Outline

This thesis is structured as follows. First we will give an overview of related work in the field of urban analysis in Chapter 2, looking at work using social media as a data source, but also work using other sources such as GPS data. In Chapter 3 we describe how we characterize a city, by defining various attributes of the city itself, its users, and the mobility of both. Chapter 4 will describe the design and implementation of the system, and in Chapter 5 we perform our evaluation and analysis on the data, where we investigate differences between cities, in regards to user demographics and paths, and differences in the usages between social media platforms.

Chapter 2

Related Research

The subject of this work is urban analysis using social media. As mentioned earlier, the aspects of urban analysis are *who*, *where*, *when*. These aspects can be summed up as user mobility.

Studying user mobility is not a new area of research. Going back at least a decade there has been much research done in this topic. Before the rise of social media researches made use of other sources of mobility data, such as mobile phone GPS data. Nowadays most research is based on data collected from services such as Twitter and Foursquare.

In this chapter we fill first discuss work that is primarily based on user mobility. Following that we discuss work with a slightly different goal, location recommendation, and we will see which user characteristics, if any, are used here. Next we will focus on large scale geo-temporal analysis of user crowds. Finally, we summarize analysis that has been done in regards to location-based social networks.

2.1 User Mobility Analysis and Prediction

We will now discuss several papers that have tackled the problem of characterizing user mobility in the past, both in the context of analysis and in the context of recommendation.

WhereNext [17] is a method that aims to predict the next location of a moving object. It uses the Trajectory Pattern Mining algorithm defined by Gianotti [11] using GPS and GSM data.

Bayir et. al. propose a Mobility Profiler framework that profiles cellphone users based on the paths they traveled [6]. They use a Sequential Apriori Algorithm to discover patterns in the user paths. Each pattern, aside from the visited locations, also contains time contextual data, namely the distribution over days of the week, and the distribution over time slices.

The problem with the pre-social media approaches to pattern mining is that it deals with very detailed GPS data, with many location points for each user (one every second), and this is not the case for social media posts. This means concepts as travel/-transition time, time at one location, and speed are much more difficult to define, if not completely absent. However, social media approaches have the benefit of being richer in data, and the data is freely available online.

Yin et al. [28] use geo-tagged Flickr photos to extract trajectory patterns, and to subsequently rank and diversify the patterns based on the relationships between users, locations and trajectories. The diversifying is done in order to present the user (a tourist) with several interesting routes he can take while visiting a city. The Flickr photos are mapped to a location using their tags. The user trajectories consist of all the locations visited in one day. The trajectory patterns are mined with the PrefixSpan algorithm.

Noulas et al. [18] mine several user mobility features in order to predict the next place a user visits. They analyze a Foursquare data set consisting of 35 million check-ins gathered in a period of five months. The researchers define the "Next Check-in Problem" as the exact place a user will visit next considering his historical data and current location. Three sets of features are defined: user mobility features (historical visits at target venue, categorical preferences, social filtering), global mobility features (venue popularity, distance, activity transitions, place transitions), and temporal features (checkins per day and hour). Used separately, the features categorical preference, geographic distance, and venue hour, all give good prediction results. All the features are used in a supervised learning framework using M5 Decision, which performs better than using the features separately. Though the paper claims it predicts the next point in a path, the actual method does not take into account the path the user actually takes at the moment, instead is uses their entire history.

LearNext does use the users current trail to predict the next point for a user [5]. The common user patterns of movement are extracted from a Flickr dataset. The prediction task is modeled as a learning to rank problem, in contrast to Noulas et al. who defined it as a binary classification problem. The researchers define two sets of features, one set describing the current user trail, and another set describing candidate PoIs. Two machine learning techniques, Gradient Boosted Regression Trees and Ranking SVM, are evaluated and compared to WhereNext and a Random Walk [16] approach. It is shown that these techniques are consistently outperformed by GBRT and Ranking SVM.

2.2 Location Recommendation

Location recommendation is a slightly different topic than next point prediction that was described before. The goal of location recommendation is to simply recommend a relevant location for the user to visit, (i.e. "I am here, what can I visit here?").

Ye et al. [27] develop a *friend-based collaborative filtering* (FCF) approach for location recommendation. They also propose a variant *Geo-Measured FCF* (GM-FCF) technique which utilizes heuristics observed from geospatial characteristics in their Foursquare dataset. FCF looks at the social friends of a user, and uses similarity measures (the locations visited) between friends to recommend a location. GM-FCF expands upon this by also taking into account the distance between friends in calculating similarity.

Bao et al. [4] recommend locations based on personal preference, and social opinions which are mined from local experts. The personal preference of users is extracted from the categories of the venues the user has visited. The categories are organized in a graph, where a lower layer of the graph indicates a *subcategory*. Each node has a

value of the number of visits, and additionally TF-IDF is calculated for each node. The graph has the advantages of reducing concern about different data scales of different users, handling data sparsity (by looking at categories instead of venues), and enabling computation of similarity between users who live in different cities (and thus do not share any venues).

Noulas et al. [19] analyze several recommendation algorithms used to recommend new (Foursquare) venues and the assumptions these are based on. These assumptions are:

- Users will check in at the most popular venues (`popularity`)
- User preferences can be captured in a succinct group of categories (`activity`)
- Users will exclusively visit the places visited by their friends (`socialnet`)
- Capturing the locality that users tend to frequently visit will increase the likelihood of finding new venues (`distance` from home)
- historically like-minded users will continue to have shared preferences in the future (`kNN`, `placenet`, `matrix`)

The researchers also propose a random walk (with restart, `RWR`) approach to recommending a venue. In their evaluation, using a Foursquare dataset, is is shown that `popularity`, `activity` and `RWR` perform best. It is also shown that the results of different cities agree with each other

Balduini et al. [1] demonstrate a Continuous Predictive Social Media Analytics system (CP-SMA) which operates on social media streams in order to recommend venues to visitors of city scale events. The interesting module of this system in relation to our work is the *Visitor Modeler* component, which creates *historical* profiles and *event* profiles. These profiles are based on the users' online conversations, demographics, trends, online presence, and influence, by using semantic extraction tools. Venues are linked to tweets by textually comparing the content of the tweet to the name of a venue.

2.3 Crowd based geo-temporal analysis in Social Networks

The closest to our work is the study of large crowds of users in order to characterize areas or cities.

Jiang et al. [13] attempt to discover urban spatial-temporal structures by studying human activity patterns that are constructed from travel surveys. The travel surveys contain data such as what type of activity a person does when and where. Similar to our own project, this paper wants to analyze differences between users in how they use urban spaces. K-means clustering via PCA (principle component analysis) is used to cluster daily activity patterns of different user groups, as well as clustering individual daily traces of people.

Lee et al. [15] measure geographical regularities of crowd behaviors in order to develop a geo-social event detection system. Regularities are measured by three indicators; the number of tweets in a region of interest (RoI) in specific period of time, the

number of users in a RoI within a specific time period, and the number of users moving in and out of the RoI. Days are divided into sections (morning, evening etc), and for each time period a box plot is created. By using the box plot one can then determine if the activity in a RoI is irregular or not.

In [25] this same approach is used to characterize urban areas. The box-plots are now used to extract crowd behaviors for urban areas. The behavioral patterns are then analyzed and empirically labeled (for example, *bedroom towns*, *office towns*, etc)

One of the few examples of research that uses both Foursquare and Twitter data is the research of Kling and Pozdnoukhov [14]. However, their area of interest is topic modeling. Every checkin is considered a "word" in a "document" that is then used with LDA (Latent Dirichlet Allocation) to learn topics.

Balduini et al. [3] use Twitter and Instagram to analyze the activity on social media during a large city scale event (in this case, Fuorisalone 2013, previous work focused on London [2]). A feed of Twitter and Instagram post is processed using C-SPARQL to extract hashtags and sentiment. The results are aggregated in a web tool that allows users to easily get an overview of the (live) activities of people during the event.

Livehoods [9] is a very similar application compared to our research goal. The goal of the researchers is to study the composition of a city on a large scale using social media. Using a spectral clustering model, the application can group nearby venues into clusters (referred to as *livehoods*). The model takes into account the spatial proximity of venues, as well as the *social proximity* based on the users that have checked in to venues. The interactive website[1] provides tools to analyze clusters, see which venue categories are popular in that cluster, the temporal activity, and which related clusters users have also visited. Our research definitely shares some of the goals of this paper, however we intend to focus more on the different types of users and which paths they take. Furthermore we will use Twitter and Instagram posts in addition to Foursquare checkins.

A concrete goal of crowd based analysis is community detection. Wang et al. [26] use location based social networks (i.e. Foursquare) to achieve this task. Several features are defined in order to cluster different users together. These are *user-venue similarity*; where each user is represented as a vector of visited venue categories, and similarity between two users is calculated with cosine similarity. *Venue-user similarity*, where a venue is represented as a vector by treating users as its features. *User-social similarity* to characterize the social relationships, here similarity is calculated with Jaccard similarity. *User geo-span similarity*, uses the *the radius of gyration* (which indicates how far and how often a user moves). The last feature is *venue temporal similarity*, where a week is divided into time slots of 1 hour thus creating a weekly temporal band for each venue category.

2.4 Analysis of Location based Social Networks

A large scale study of user behavior in Foursquare is done in [20]. A dataset of 700 thousand users collected (via Twitter, which at the time was approximately 20% to 25% of the complete Foursquare user base) over a period of a 100 days is analyzed by the researchers. About 20% of the users have just one checkin, 40% above 10

[1] `http://livehoods.org/`

checkins, and about 10% of the users has more than 70,000 checkins. Overal activity during weekdays has three peaks, in the morning when people go to work, during lunch time, and between 6pm and 8pm during the commute home. During the weekend the activity is much smoother. Around 10% of consecutive checkins are made within 10 minutes, which rises to about 30% within 100 minutes. The analysis also shows that the data can provide valuable insights in how activities of mobile users success each other.

Cheng et. al. [8] also analyze a large dataset consisting of 22 million checkins and 220,000 users, gathered via Twitter. Before analyzing the dataset, the set is cleaned by removing all checkins that imply a speed faster than 1000 miles per hour. The home location is also calculated for each user. Analysis on *where* the checkins are shows that the most popular checkin venues are restaurants, coffee shops, stores, airports and other venues that are part of daily activities. Daily temporal patterns show (global) peaks at 9am, 12pm and 6pm. By comparing three cities (Amsterdam, Los Angeles, and New York) it is shown that the daily checkin pattern can reflect on the "heartbeat" of a city, for example, Amsterdam has a higher activity around 9am, LA during the afternoon and New York has the most activity compared to the other cities during the night. The weekly temporal patterns show that during weekdays there are distinct peaks during lunch time and dinner time, while in the weekends these peaks fade and activity is much more constant.

The researchers also study mobility patterns by looking at three characteristics, user displacement, radius of gyration, and returning probability. For user displacement it is shown that human motion modeled with checkins follows a Lévy Flight [23], which is consistent with previous analysis on human mobility. Around 34.5% of all users have a radius of gyration less than 10 miles, while only 14.6% have a radius over 500 miles. Users in coastal cities have on average a higher radius of gyration compared to users in inland cities, but people in central states also have a high radius because of the long distances. The return probability is defined as the probability a user returns to a venue within x hours since his first visit. The analysis shows there are peaks at 24 and 168 hours (signifying daily and weekly return patterns), and that the probability decreases over time.

These works have been the main inspiration of this thesis. They show that social media sources are a good approximation to real human behavior in cities. The main source however here is Foursquare, and we aim to build upon this analysis with using Twitter and Instagram as sources, and also include user analysis based on features and attributes described by the works in the previous three sections.

Chapter 3

Design: Analyzing City Usage

In order to analyze a city we need to define people, mobility and the city itself. In this chapter we describe what features and attributes can be extracted from location-based social media, and our design of a pipeline that can extract, enrich and analyze these attributes.

Starting with the city itself, we define it as a simple coordinate bounding box. The city can be further divided into points of interest, which are places such as restaurants, town squares, shopping malls, museums etc.

People (henceforth referred to as users) can be characterized in a number of ways. We focus on the demographics of a user, such as his home, his age and gender. We are also interested into what role the user plays in a city.

We describe the mobility of a user with his radius of gyration, but also the paths the user takes in the city. By aggregating the paths all the users traverse in a city, and finding the common path patterns, we have a way of characterizing mobility of a city as a whole.

3.1 User Paths and Patterns

In this section we will discuss how we extract paths from location data, and how we discover frequent patterns.

3.1.1 Points of Interest

Definition 1. *Post* = $\langle$ *id, timestamp, latitude, longitude, text, user* $\rangle$

Each post posts provide us with a set of coordinates and a timestamp.

In order to create meaningful paths we need to map these coordinates to *points of interests* (PoI). In previous research several methods of PoI extraction have been proposed, such as using tags of the post to map the location to Wikipedia articles [28, 5]. Other research focuses solely on Foursquare checkin data [4, 19], that provide *venues* and *categories*. Using Foursquare venues as PoIs gives added information (categories, location, popularity) supplied by Foursquare, in addition to being a straightforward and simple way of extracting PoIs. We thus use the Foursquare venue model as our definition of a point of interest:

Category	Examples
Arts & Entertainment	Museums, Music Venues, Theaters, Stadiums
College & University	College and University buildings, Fraternity Houses
Event	Conferences, Conventions, Festivals
Food	Restaurants, Cafeteria, Cafes, Coffee Shops
Nightlife Spot	Bars, Nightclubs, Pubs
Outdoors & Recreation	Sport Fields, Beaches, Parks, States & Municipalities
Professional & Other Places	Convention Centers, Libraries, Schools
Residence	Residential Building, Homes
Shop & Service	Shops, Banks, Gyms
Travel & Transport	Airports, Public Transport, Hotels

Table 3.1: Venue Categories

Definition 2. $PoI = \langle$ *id, latitude, longitude, city, country, name, category, root_category, checkin_count* $\rangle$

Each venue has a category. The category list defined by Foursquare contains over a 100 categories and is hierarchical[1], therefore we add an extra *root_category* field to a venue, so we can keep our analysis focused on the top level categories. The different categories are listed in Table 3.1.

We then need to link a PoI to a post. We define this as a *visit*, meaning that if a user has a post nearby a PoI, he has visited that PoI.

Definition 3. $visit = \langle post, poi \rangle | poi = M(post)$

Where M is a mapping function that maps a post to a PoI. The implementation of this mapping function is described in Chapter 4.

3.1.2 Paths

As described in the related work, there exist a number of techniques to determine user paths. However, the data-source those techniques use is GPS or mobile phone data. This data is very detailed, and thus the extracted paths can be very detailed, taking into account how long users stay in one place, how fast they move etc. Social media data however is very sparse, in the sense that many users do not post more than a handful of times per day. This means paths will not be very detailed, and we do not have information such as how long a user stays at a certain location. This means the trajectory pattern approach [11] is infeasible. Instead we define a path as follows:

Definition 4. $path = \{id, \langle poi_i, timestamp_i \rangle\} \mid poi_i \neq poi_{i+1} \wedge timestamp_i < timestamp_{i+i}$

In other words, a path is a list of PoI's sorted on time, where no two subsequent PoI's are the same.

[1] `https://developer.foursquare.com/categorytree`

3.1.3 Path Patterns

To find common sub-sequences of paths we refer to previous work done in the field of pattern mining. Yin et al. [28] apply the PrefixSpan algorithm [21] in a similar context as we do, to good results. We apply this algorithm to find patterns with a support greater than 5.

Definition 5. $path\_pattern = \langle id, \{poi_i\}, \{paths\}\rangle \mid \{poi_i\} \in \{paths\}$

A path pattern is then modeled as a ordered set of PoIs, and a set of the paths this pattern is a part of.

3.2 User Attributes

One of the most important aspects of our research goal is the "who" question. Knowing why people take a certain route, or visit certain places we feel gives incredible insights into city usage.

To solve this problem we can extract a number of attributes, or features, of users, that describe their demographic. We first define our user model, and then explain each attribute:

Definition 6. *user* = ⟨ *id, name, profile_picture, home, city_role, gyration, gender, age* ⟩

3.2.1 Home location

Users of social media are often not required to register their home location. However, by taking into account the full post history of a user, we can use all their posts to approximate their home location.

We use a similar method as described by Cheng et al. [8] to determine a users home location. They explain that you cannot simply take the average location of all the users posts, as then you could end up in the middle of nowhere. For example, if a user posts 30% of the time in Amsterdam, and 70% of the time in Rotterdam, the average location would end up in between the two cities, but the home location would most likely be in Rotterdam. This is why we need to find the actual place where the user posted most often, and use that as a approximation of the users home location.

The method is a recursive grid search. Cheng et al. use a custom grid size, we instead use geohashes[2], which already represents coordinates as a grid. The algorithm is defined in Algorithm 1. We start by mapping all the posts of a user to a geohash of length 2. Then in the next step we select the geohash with the highest number of posts and its adjacent geohashes. We then map all the posts made in those geohashes to geohashes with a length incremented by one. We continue doing this until we have geohashes with length 8 (which is about 20 meters) and we select the center of the geohash with the highest number of posts as the users home.

[2] `http://en.wikipedia.org/wiki/Geohash`

Algorithm 1 Determining the home location of a user

```
1: function HOME(history)
2:     grid ← {}
3:     length ← 2
4:     for all posts p in history do
5:         h ← geohash(p_latitude, p_longitude, length)
6:         grid[h] ← grid[h] ∪ {p}
7:     end for
8:     home ← FindHome(grid, length + 1)
9: end function

10: function FINDHOME(grid, length)
11:     m ← getMaxCell(grid)    ▷ getMaxCell() returns the geohash of the grid with the most posts
12:     if precision > 8 then
13:         return center(m)    ▷ center() returns the center coordinates of the given geohash
14:     else
15:         newgrid ← {}
16:         for all posts p in center and adjacent cells do
17:             h ← geohash(p_latitude, p_longitude, length)
18:             newgrid[h] ← newgrid[h] ∪ {p}
19:         end for
20:         return FindHome(newgrid, length + 1)
21:     end if
22: end function
```

3.2.2 City Role

One important aspect of characterizing the user, is understanding the role he plays in a city. We choose to classify users in relation to the city, by looking at the home location of the user. We define three simple classes: Resident, Local Tourist, Foreign Tourist.

Resident
If the city of the users home location is the same as the city under study.

Local Tourist
If the city of the users home location is different than the city under study, but it is still the same country.

Foreign Tourist
If the users home location is in a different country.

These classes are easy to determine, but can provide great insight in how different types of people use a city. The psuedocode for determining the different roles is shown in Algorithm 2.

Algorithm 2 Determining the user city role

```
1:  function USERROLE(user, city)
2:      home ← Home(history(user))
3:      if home.city = city then
4:          role ← RESIDENT
5:      else if home.country = city.country then
6:          role ← LOCAL_TOURIST
7:      else
8:          role ← FOREIGN_TOURIST
9:      end if
10:     return role
11: end function
```

	User 1	User 2	User 3	User 4	User 5
$r_1 - r_{cm}$	2	30	100	2	30
$r_2 - r_{cm}$	2	50	1	1	40
$r_3 - r_{cm}$	2	1	1		70
$r_4 - r_{cm}$	2	2	2		2
$r_5 - r_{cm}$	7	1	3		25
r_g	1.73	4.10	4.63	1.22	5.78

Table 3.2: Example of radius of gyration

3.2.3 Radius of Gyration

The radius of gyration is a useful measurement when interested in mobility of users. The radius of gyration is an indication of how far and how often a user travels. Radius of gyration is defined as:

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cm})^2} \tag{3.1}$$

Where n is the number of posts of the user, and $r_i - r_{cm}$ is the distance between post r_i and the center of mass of all the posts (i.e. the average location) r_{cm}.

The radius of gyration can indicate how mobile a user is. Table 3.2 shows a small example of how it works. User 1 travels the shortest distances, thus he has a small radius of gyration. User 5 travels the longest distances, and has the highest radius. Users 2 and 3 have traveled far twice and once respectively, with a comparable total distance traveled and thus they have a comparable radius of gyration. User 4 only traveled short distances twice, and as a result has the lowest radius. This simple example clearly shows that radius of gyration is a good measurement for mobility.

3.2.4 Gender and age

There exist several techniques for determining user properties such as gender and age, based on their social media profile. A general model of a social media profile usually contains a *screen name*, most platforms also require a *full name*, the profile also has a *profile picture*, and some platforms also have a short *profile description*.

Burger et al. [7] performed a study to see what features of a user perform best in determining the gender of a user. They used Twitter as a data source, and conclude that the *full name* is the most informative feature, having an accuracy of 89.1

Users also have a profile picture. If these contain a picture of the user himself, facial recognition techniques can be applied to extract certain properties of the person in the picture, such as age and gender.

Both approaches are not foolproof, as not all social media platforms require real full names, and not every profile picture contains the face of the user. A combination of these two approaches will be the most effective.

Chapter 4

A System for City Usage Analysis

The implementation of a system for city usage analysis has two main components, an offline calculation component, and an online visualization tool. This division follows from the purpose of this system, which is twofold. The main purpose is to enable us to perform an analysis based on the properties we described in the previous chapter, to gain insight in city usage.

The second purpose is to also show the benefit of this kind of city analysis in a real world use case. To this end we developed the visualization tool that can demonstrate, to a a non-scientific crowd, the added value of our analysis approach.

We choose two social media sources to provide our system with data, Twitter and Instagram. In theory any geo-enabled social media source can be integrated in our system in a later stage.

4.1 Requirements

The complete system needed to be flexible in the sense that it could be easily extended with more possible data extracted from social media posts. By keeping the application modular it can also be fitted into existing architecture.

In regards to the front end, the visualization tool needed to be straightforward in design, but still give many data visualization options. The tool should be a proof of concept for an application that non-scientific people could use. Key requirements thus were the ability to present different views, add and combine different filters, and include the functionality of a timeline in order to see temporal patterns. Like the back-end, the visualization tool also should be easy to extend.

4.2 Architecture & Implementation

Figure 4.1 shows the architecture of the complete system, both the online and offline part. The gray boxes in the figure represent components of the system not made for specifically for this thesis, but they are used. The boxes with dashed lines are components that in a later point in time can be integrated.

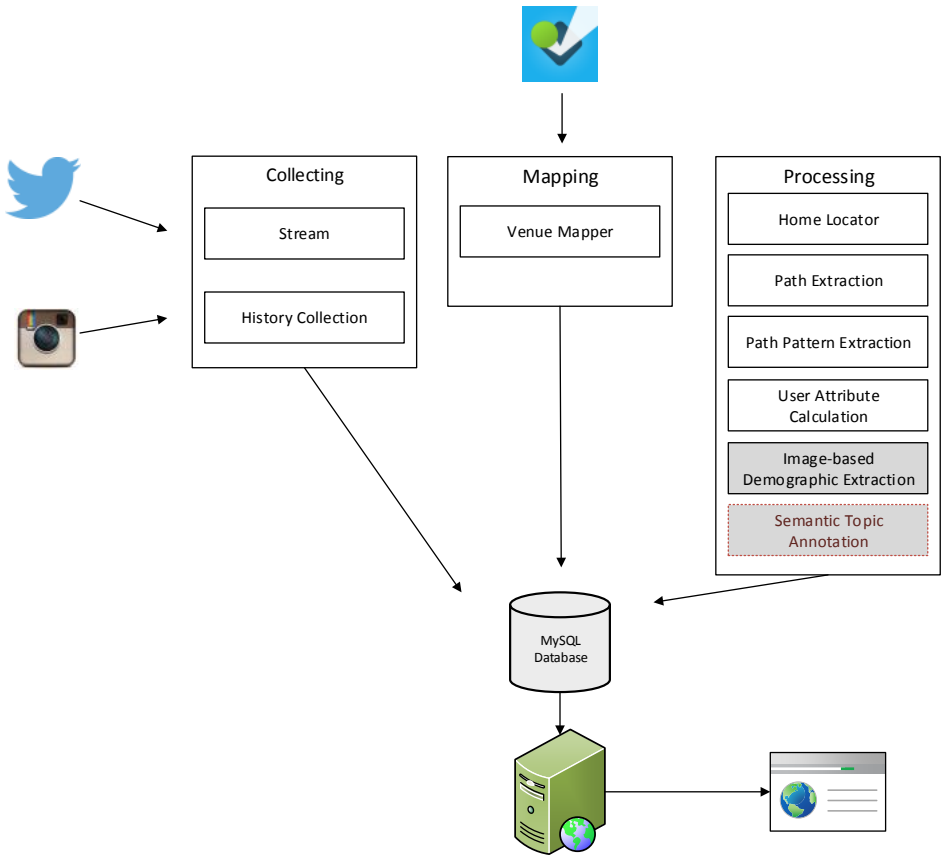


Figure 4.1: High level overview of the system architecture

4.2.1 Pipeline

The offline part of the system is modular in design, and therefore can be seen as a pipeline. The application implemented in Java.

It is assumed the system is supplied with a stream or a pre-existing crawl of tweets and Instagram posts within a certain area A. The pipeline can then be used as such:

Stream

The stream module provides the system with an ongoing supply of tweets and Instagram posts. The Twitter stream is implemented using the Twitter4J library, and for Instagram we implemented our own listener.

User History Collecting

When the system discovers a new user via the provided stream, this component collects their post history. Twitter can retrieve upto 3000 tweets of the user, and with Instagram it is possible (through a bit of a hack) to collect the entire post history of the user.

Venue mapping

Once the posts of the user are collected, we map them to PoI's. The mapping function $M(post)$ described in Section 3.1.1, is implemented as a call to the Foursquare API.

The Foursquare API has three different methods of mapping coordinates to a venue[1] (referred to as *intents*), a *checkin* intent which returns the venues a typical user is most likely to checkin to given his current location, a *browse* intent which returns all venues in an area, and finally a *match* intent which is used to give near exact matches. The match intent is not usable in our use case, as the API requires a name to search for, and we do not have that data. In theory one could use all words in a tweet or Instagram description as the *name* parameter, however another limitation of the API is the fact that it does not support multi-word queries.

The checkin and browse intents do not differ significantly from each other, however Foursquare recommends to use the checkin intent, as it emulates the action of a Foursquare user checkin the best. This means that the mapping will be more akin to real life, and will prefer popular venues over obscure ones if they are both nearby.

The API returns a list of venues and we select the first venue within a 40 meter radius.

The module can be configured to map all the posts, or just the posts in the region and/or timespan of interest.

User home locator

When we have all the posts of the user, we can run the algorithm to determine the home location of the user.

Once the home coordinates are determined, we then also create a "fake" Foursquare venue for this users home, with the category *Home*. Each post of the user within a 10 meter radius of his home location is mapped to this venue.

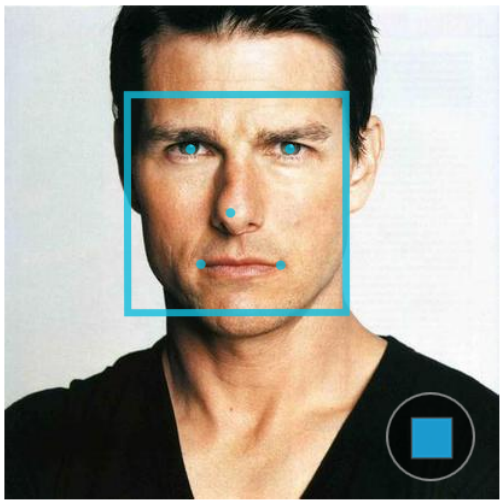
As each venue needs a city and country, we reverse geocode the home coordinates to find the address. We use the geonames.org API to find the country. Geonames.org has a very detailed dataset of all kinds of adresses, and also has a high API rate limit, making it a good fit for our needs. However, geonames does not return a simple city name. For example for locations in London, you either get the boroughs such as Westminster, Chelsea, etc, or even more lower level administrative districts. We are only interested in the simple city name *London*. For this reason we use a second reverse geocoding service, Google Maps, to find the simple city name for the result returned by geonames. Google has a much lower API rate limit, which is why we only use it as a secondary geocoding source, and we also cache the calls made to Google.

This module can be run concurrently with the venue mapping, as it does not depend on the venue mappings.

User attributes calculation

This module calculates various user attributes we are interested in. Once the home location is determined, we can classify the user as one of the roles we defined earlier, and we can calculate the radius of gyration, and extract gender and age.

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



(a) The image containing a face


```
{
  "face": [
    {
      "attribute": {
        "age": {
          "range": 5,
          "value": 23
        },
        "gender": {
          "confidence": 99.9948,
          "value": "Male"
        },
        ...
        "race": {
          "confidence": 99.9637,
          "value": "White"
        },
        "smiling": {
          "value": 1.24896
        }
      }
    }
  ]
}
```


(b) Face++ API response

Figure 4.2: Determining gender and age using facial recognition

For the age and gender extraction, we use the Face++[2] service for facial recognition. Face++ is one of the leading online services in the field of face detection, recognition and analysis, both in terms of accuracy and popularity. They provide a REST API which accepts URLs and returns a list of properties, including age and gender, together with confidence values for each property. Figure 4.2 shows an example input/output.

To test the performance of the Face++ service, we created a ground truth, by manually checking the profile pictures of Twitter users. Table 4.1 shows how successful Face++ was in detecting faces in profile pictures. It shows that if there is a face present in the picture, 410 cases out of 628 it could recognize it. There are also very little false positives (cases where there is no face, but Face++ does detect a face). Table 4.2 shows how the tool performs in actually determining age and gender. Both properties have about 80% accuracy.

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

	Face Detected	**No Face Detected**	**Total**
Face present	410	218	628
No Face present	6	133	139
Total	416	351	767

Table 4.1: Performance of Face++ API for detecting faces

	Age	%	**Gender**	%
Correct	329	80.24%	361	88.05%
Not Correct	69	16.83%	44	10.73%
Not Sure	12	2.93%	5	1.22%

Table 4.2: Performance of Face++ API for determining gender and age

Path extraction

The path extraction module collects all posts of a user in a given city, during a specific time frame. Paths are then constructed by collecting all tweets or posts of a user in a single day. We first only look at the coordinates, if two subsequent coordinates are the same, we only keep the first instance. We then have a list of candidate posts for a path.

Then we map those candidate posts of the path to PoI's (if they are not already mapped) and again filter out all subsequent identical PoI's. If the path length is equal or greater than 2, we save it to our database.

Pattern extraction

We use the implementation of PrefixSpan provided by the SPMF Java library [10]. As the patterns are not user specific, this can be run after certain intervals, when enough new paths have been extracted.

4.2.2 Visualization tool

A real world use case of our research is exemplified in the visualization tool. The tool is a web application developed in PHP and Javascript. The map and the different visualizations are made using Leaflet with a number of plugins.

The system provides many different properties and combinations of properties that can be visualized, however to keep the application as streamlined as possible, we chose to focus on the key aspects; *what* (PoI's), *who* (user role), and *when* (temporal dimension).

The tool provides several visualization types which allows the user to study the key aspects on different levels. Figure 4.3 shows how the user can select a visualization and add it to the map. For high level views of a city, the *Heatmap* visualization type is best suited, showing the user at a glance popular areas, and changes over time. For a more detailed view of a city, we provide a *Choropleth* visualization, that can visualize the key aspects for each district of the city. And finally for the most low-level view of the data, for example to inspect individual points of interest, the tool has a *Point* and *Path* visualization.

The different visualization types also give different functionality:

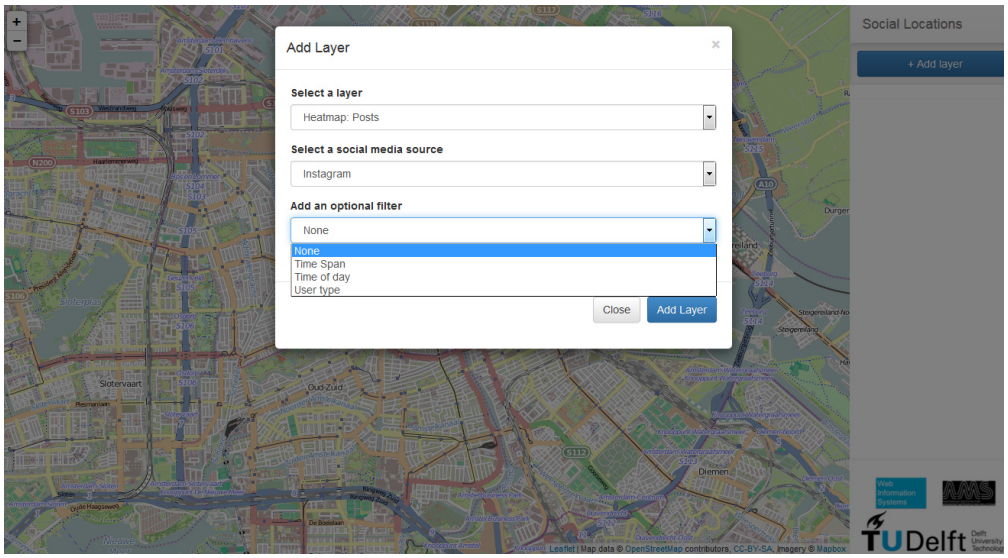


Figure 4.3: Layer selection pop-up. Here the user chooses which layer, which social media source, and possible filters.

Points

We provide two Point visualizations, a Post Point layer, and a Venue Point layer. In these layers the points are the tweets/instagram posts and the mapped venues respectively. Clicking on a point gives you extra information about it. The radius of venue points is coupled with the popularity of the venue. We use the Leaflet markercluster plugin to also cluster points when zoomed out, to not overwhelm the browser. In Figure 4.4 the Post Point layer is shown, with Instagram as the data source, and a filter added to only show posts made between 6PM and 8PM.

Paths

The extracted paths are shown as edges between each venue on the path. The more times a direct line between two venues appear, the thicker the line gets. The radius of a venue point is determined by the number of times it appears on a path. Figure 4.5 shows this visualization.

Clicking on an edge gives you additional data about the path, such as the start and end venue, and also the number of paths this edge is a part of.

Heatmap

In addition to the Point layer, one can also choose to display posts as a heatmap. This can be filtered on user role, and platform type (Twitter, Instagram).

Figure 4.6 shows a regular heatmap. In Figure 4.7 the heatmap is shown with the timeline function enabled. This feature enables the user to see changes in activity over time.

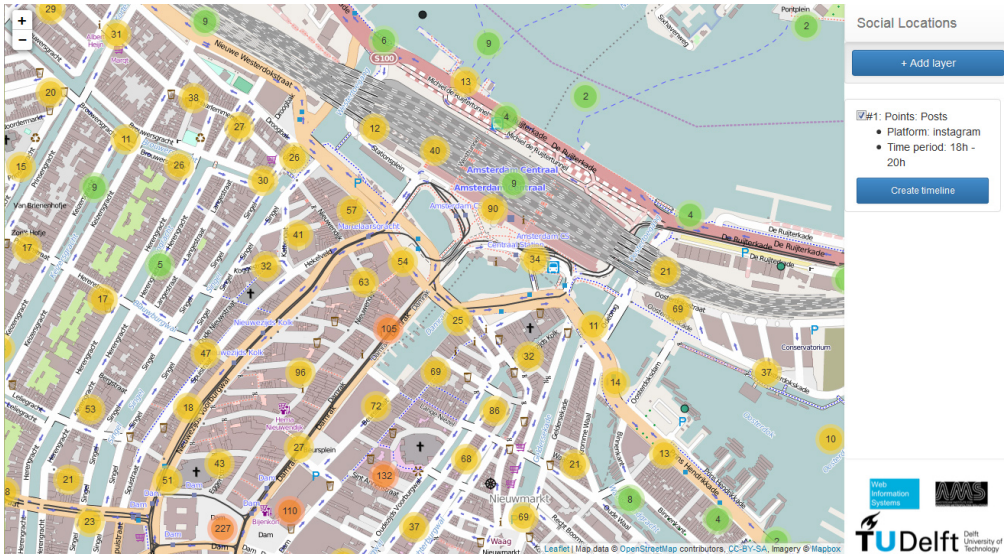


Figure 4.4: Points visualization. The circles represent an automatically generated cluster of points. The color and the number of the circle indicate how many posts are made in this cluster.

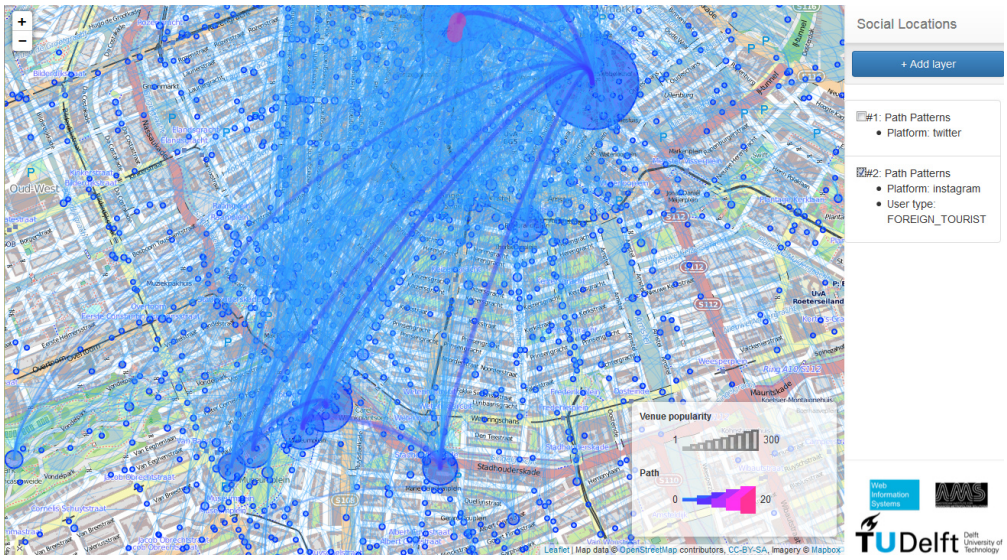
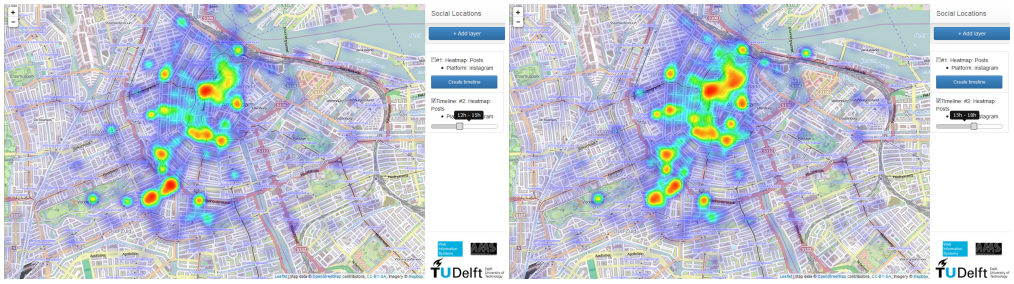


Figure 4.5: Path visualization. Here we see all the paths traversed by foreign tourists in Amsterdam. The large circles are popular venues (museums). The thick lines between them indicate these venues are often on the same path.

Figure 4.6: Heatmap visualization.



(a) 12-15h (b) 15-18h

Figure 4.7: Heatmap visualization using the timeline functionality. By sliding the handle we can observe changes over time.

Choropleth

The choropleth layers provide the most functionality. For each city we have collected a GeoJSON file of the administrative districts in the city. This districts are drawn on the map, allowing the user to inspect the city per district. The tool provides a Venue Category layer which shows the most popular category in that district, shown in Figure 4.9, a Posts layer which shows the number of posts made in each district (essentially a more advanced heatmap) shown in Figure 4.8, and a User role layer which shows the most common user role (Resident, Local Tourist, Foreign Tourist) in the district. Clicking on a district opens a pop-up showing more detailed statistics of that district.

These layers combined with a number of filters (time period, user role, venue category) provides the most insightful views of a city.

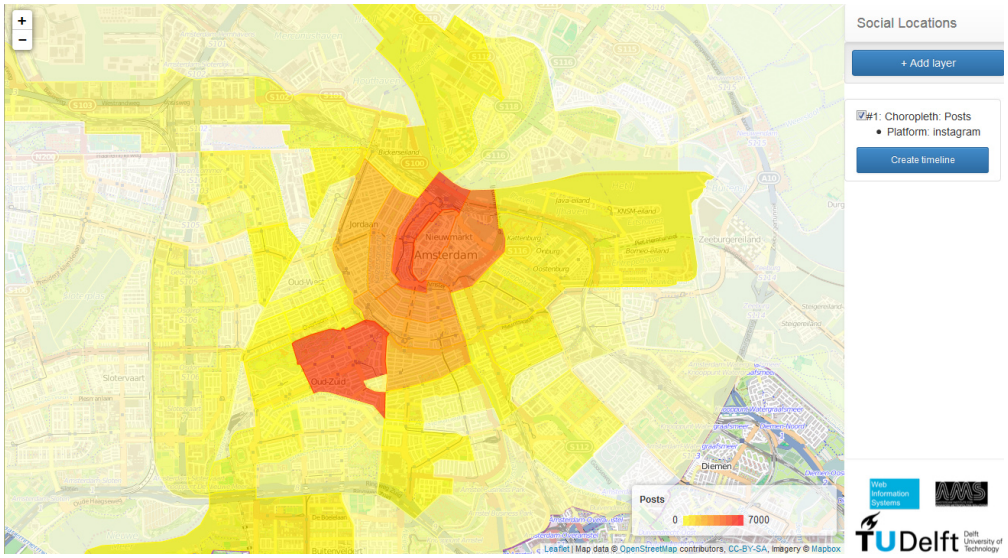


Figure 4.8: Choropleth Posts visualization.

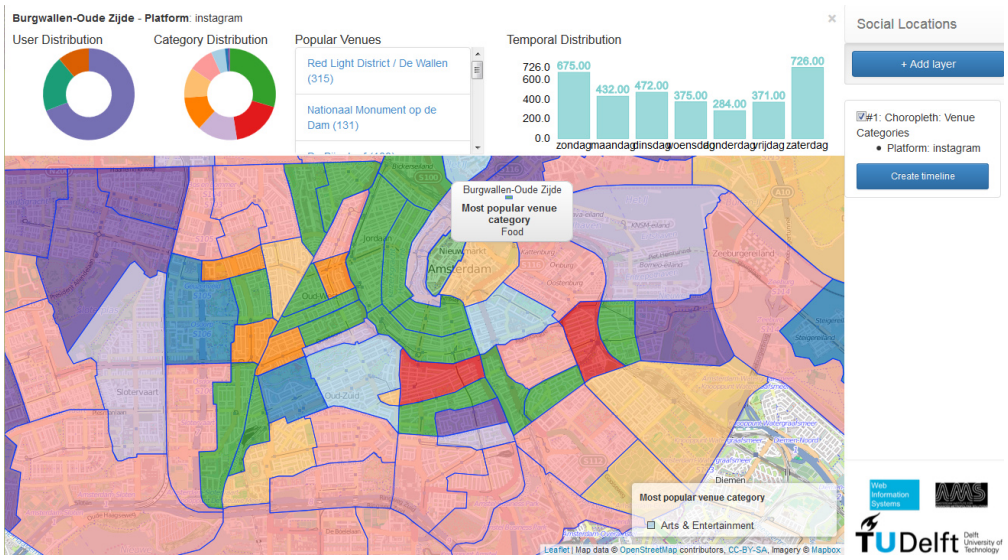


Figure 4.9: Choropleth Posts visualization. Clicking on a district opens an overlay at the top of the screen, showing the distribution of user roles and venue categories, the most popular venues, and the temporal activity in that district.

Chapter 5

Urban Analysis: Comparing Cities

In this chapter we reports on the analysis of city usage through social media, performed on four European cities. To enable and support the analysis we used our novel system, thus empirically validating it. As the possible combinations of attributes to analyze is potentially endless, we focus ourselves on key research questions. These research questions follow from the research questions inspired by the questions defined in the introduction chapter:

RQ2 In what ways do different social media differ in regards to location data?

RQ3 Can we find differences in city usage between cities using social media data?

RQ2 will be answered by employing our system to compare two social media platforms, namely Twitter and Instagram, comparing the results of each research question between the two platforms. We know that Twitter is primarily a microblog service, and Instagram a photo-sharing service, however both platforms do have an overlap. By looking into the different locations visited by users of these platforms, and their temporal activity we hope to answer exactly how the location data extracted from these platforms differ. This also ties in with RQ1, as once we know the differences between the social media platforms, we can also know which platform is best suited for a particular use case of research.

RQ3 is the ”end result" of our system and design, where we see if the attributes we selected can actually provide significant insight into city usage. We do this by analyzing the *who* (users), *what* (PoI's), and *when* (temporal activity) aspects of city usage.

We divide this chapter into 5 key areas: general Twitter and Instagram usage analysis, venue analysis, user analysis, path analysis and path pattern analysis. For each of these areas we answer several research questions that we hope to provide valuable and significant differences between cities, demonstrating the value of our approach.

5.1 Experimental Setup

We use two main social media sources to perform our analysis, Twitter and Instagram. To create a dataset for our analysis we collected tweets and Instagram posts from four

	# posts		# users	
	Instagram	Twitter	Instagram	Twitter
Amsterdam	61,774	50,530	10,520	6565
London	411,223	460,992	59,814	49,234
Paris	249,302	352,166	32,292	17,828
Rome	104,507	69,584	15,888	5671

Table 5.1: Dataset numbers

different cities, Amsterdam, London, Paris, and Rome, during a three week period from February 20th to March 12th 2014.

We extracted paths for every user in the city during the crawling period, and also mapped all posts to Foursquare venues. For all active users we calculated the home location and the other various attributes. We define active users by ranking all the users by the average number of posts per day the users made and then selecting the top third for each city.

We summarize the relevant numbers in Table 5.1.

5.2 Instagram vs Twitter usage

While Twitter and Instagram share similarities, there are differences in how people use them. The main difference we expect between them is the main focus of Instagram, pictures instead of (only) text. We suspect this will influence when and what people post. Another difference which we discovered by studying a small selection of our dataset content, is that when users take a picture via Instagram they do not have to immediately post it online. Users can also edit and upload pictures to Instagram which they made with their regular phone-camera app. We refer to this as "delayed posting". An example of this can be seen in Figure 5.1. We are interested in how these differences in usage present themselves when looking at temporal patterns of Instagram versus those of Twitter.

Figure 5.2 shows the activity of users over a day, for all the cities combined. During the night and morning, the activity between Twitter and Instagram does not differ significantly. However during the late afternoon and evening we start to see a difference. The peak of Instagram usage is during the late afternoon while at that same time, Twitter usage slows down. This could indicate a difference in the type of users, as we see Twitter usage increase in the evening, i.e., after work. It could also indicate a difference in places visited, perhaps Instagram photos are primarily made at tourist attractions, meaning most people would visit those during the day. Later on in this chapter we will investigate this further.

To see whether we can see the effect of "delayed posting" we look at the distribution of distance vs time between posts. We suspect that users who post pictures at a later time may post more than one picture at a time. Think of the situation where an Instagram user comes home after a day in the city, and then goes through his pictures that he took during the day, and then uploads them. The Instagram app works in such a way that the timestamp of these uploads will be at that time, but the location will be at the place the picture was taken. To discover this anomaly we look at the distance

Figure 5.1: An example of a "delayed" instagram post. This was posted at 2014-03-11 22:21:21 in Amsterdam. The photo was clearly made during the day, but was posted during the evening.

covered between consecutive posts. In Figure 5.3 we plot the distance between consecutive posts that have been made within 15 minutes of each other. Both histograms have a very skewed distributions with a very long tail. We see no apparent difference here, except a small spike for Twitter, that can be explained by the presence of automated accounts such as emergency services who post many geo-located tweets. The average human walking speed is 5 km/h, which means we need to look at distances greater than 1250m. We show two more histograms, this time cut off at 5000m in Figure 5.4. Instagram distances have more spread than tweets, but after the aforementioned 1250m, the probability densities drop below .0001.

Thus the occurrence of delayed posting can not be easily seen. We can conclude that while we have seen cases of delayed posting, the statistical impact of those cases is small.

5.3 Venue Analysis

5.3.1 Venue mapping validity

As mentioned in Section 4.2.1, we use the Foursquare API to map coordinates tweets and posts to Foursquare venues. To get a feel of how accurate these mappings are we conducted a small sample test of randomly selected tweets and posts.

We randomly selected 100 posts from both Twitter and Instagram, that were made in Amsterdam. We then let people manually annotate this small sample set. We presented them the content of the tweet or post itself, the location and info of the mapped venue, and the exact location of the post. We also showed the list of other possible venues returned by Foursquare. The person then had an option to select one of the following options:

Correctly Mapped When the mapping was a 100% match, the user was at that venue.

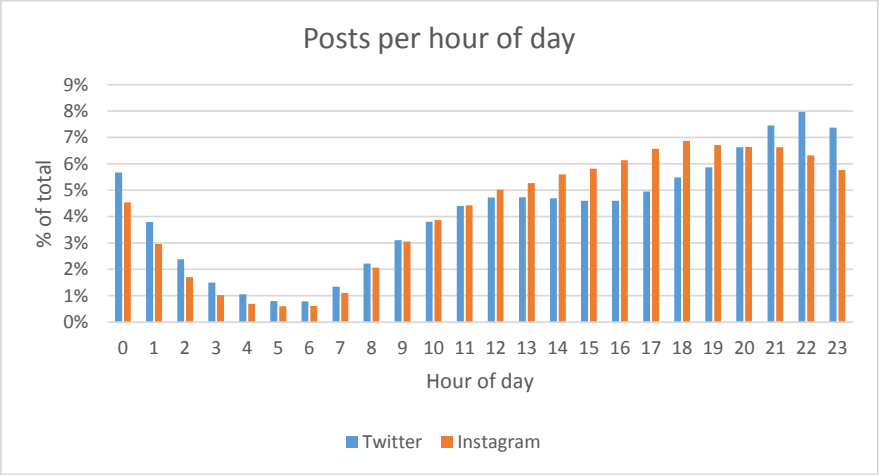
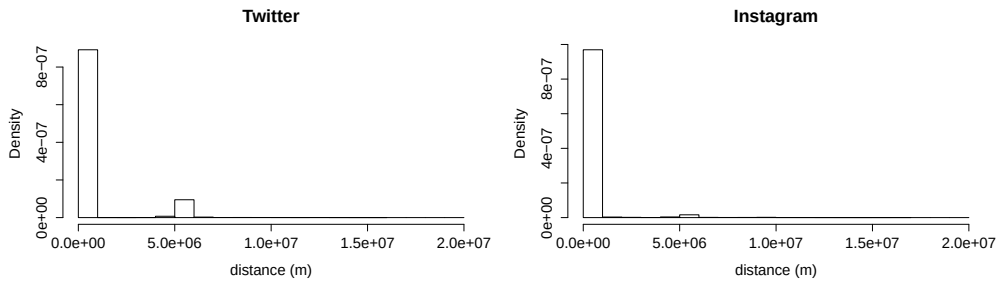
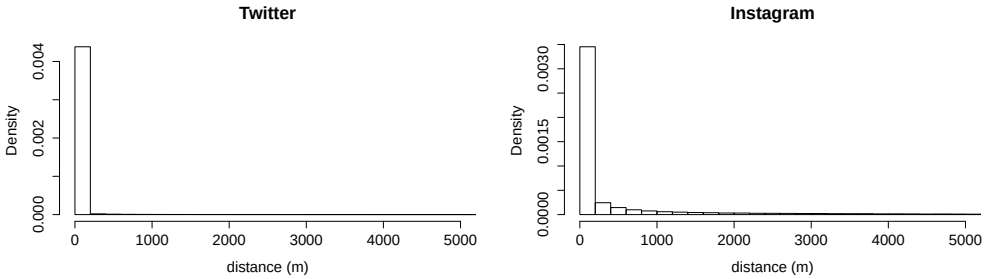


Figure 5.2: The aggregated daily activity for Twitter and Instagram



(a) Twitter (b) Instagram

Figure 5.3: Distribution of distances for posts that are made within 15 minutes of each other



(a) Twitter (b) Instagram

Figure 5.4: Distribution of distances for posts that are made within 15 minutes of each other, zoomed in on distances below 5000m

Better alternative When a better venue exists in the list of possible venues returned by Foursquare. This is an incorrect mapping.

Residential area When the mapped venue is incorrect, but the post was made in a residential area.

Office area When the mapped venue is incorrect, but the post was made in a residential area.

Passing by The mapped venue is very nearby, but the user is most likely only passing by. According to our definition of a visit, this is a correct mapping.

Completely Wrong The mapped venue is nowhere close to the post, a completely incorrect mapping.

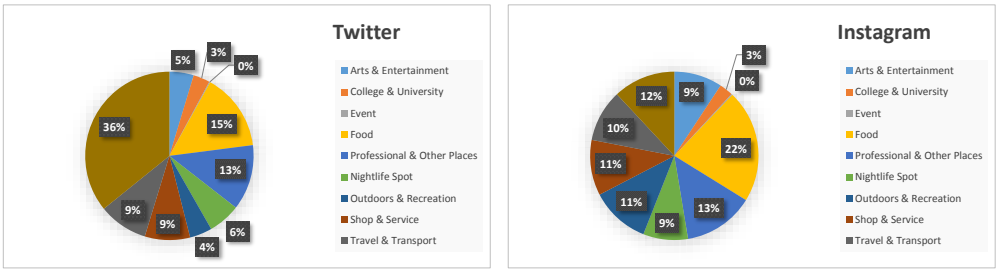
Not sure When it cannot be determined whether this is a correct mapping.

The results of this can be seen in Table 5.2.

We see that about a third of the posts were correctly mapped to the correct venue. The wrongly mapped tweets that were in a residential area are in a later stage fixed by mapping to the home venue of the user, however we still mark them as "incorrect" here as if this is a standalone module. The *passing by* mappings are also considered correct mappings. Only a very small percentage of the mappings were completely wrong (7% and 2%). Investigating the mappings that were classified as *not sure*, it turns out these posts are often inside buildings, or in very remote locations, where the annotator could not determine what the actual venue was, as most of our annotators are not from Amsterdam. In this case we have to trust the Foursquare mapping. If we

	Instagram	Twitter
Correctly Mapped	33%	39%
Better alternative	14%	16%
Residential area	7%	10%
Office area	0%	1%
Passing by	7%	15%
Completely wrong	7%	2%
Not sure	29%	16%

Table 5.2: Correctness of venue mapping



(a) Twitter (b) Instagram

Figure 5.5: Distribution of venue categories for Twitter and Instagram

summarize these results in terms of valid *visits* (*correctly mapped* + *passing by* + *not sure*), we have 69% and 70% of valid visits for Instagram and Twitter respectively.

We have to keep in mind however that this is a very small sample set, and these numbers only provide an indication of the mapping validity. In future work, the validity of the venue mapping must be further improved.

5.3.2 Twitter vs Instagram

Continuing in our study in how Twitter and Instagram usage differ, we can also look into what kind of places users of both platforms visit. Figure 5.5 shows the distribution of venue categories for both platforms, combined over all cities.

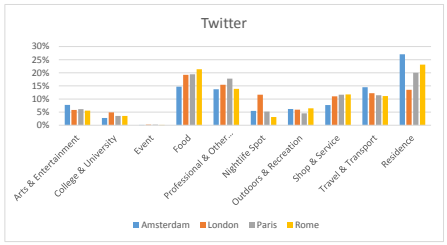
The most significant difference that can be seen here is the large percentage of posts in the Residence category for Twitter. This would imply that Twitter users tweet most frequently from home. This is supported by the the fact that the most Twitter activity is in the evening, i.e. after work, as we saw in Figure 5.2. For Instagram the Residence category popularity is average, and here the Food category (restaurants, cafes) is the most popular.

5.3.3 How diverse are the venues visited in each city?

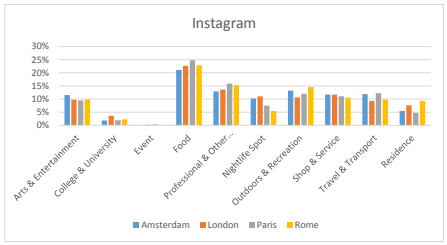
Now we will focus our comparison more towards the differences between cities. Our first question here is if we can see differences in the venues that are visited in each

	Twitter		Instagram	
	Distinct	Relative	Distinct	Relative
Amsterdam	8246	0.172	10,483	0.163
London	40,950	0.105	36,204	0.120
Paris	27,189	0.079	23,384	0.132
Rome	9401	0.147	16,478	0.172

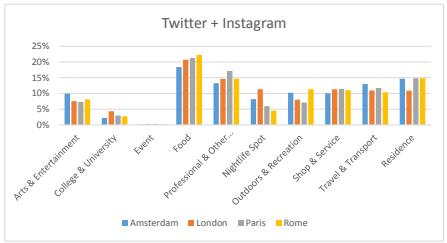
Table 5.3: Distinct venues visited in each city (Distinct column), and compared to the total number of visits (Relative)



(a) Twitter



(b) Instagram

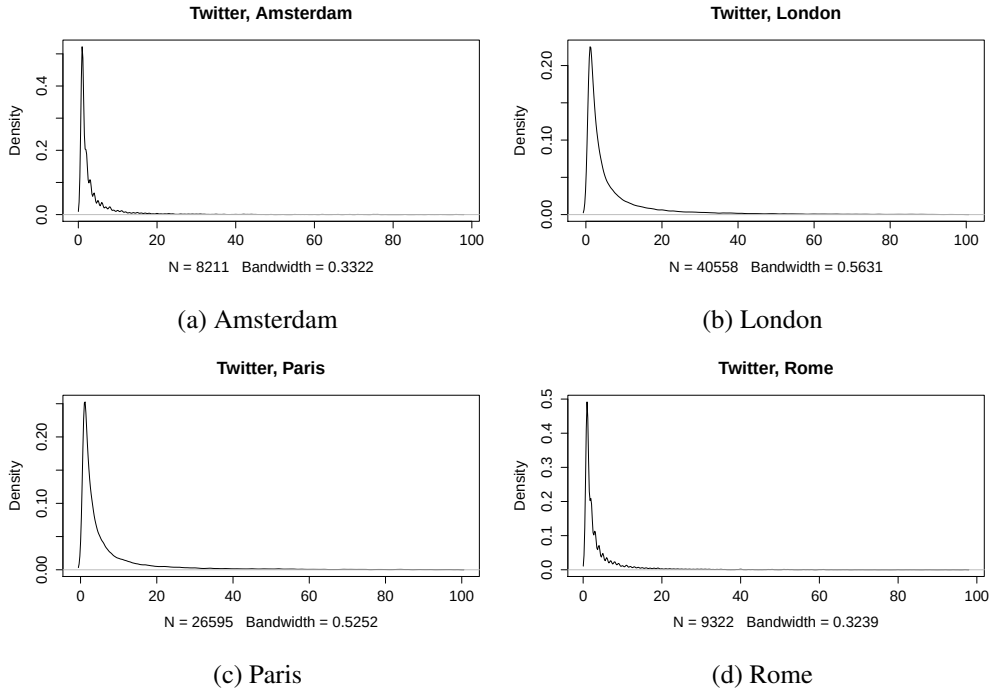


(c) Aggregated values of Twitter and Instagram

Figure 5.6: Distribution of venue categories for Amsterdam, London, Paris and Rome

city, in other words, how diverse are the venues that are visited. Is the city more aimed at nightlife? Or is arts more popular? How many distinct venues are visited? Table 5.3 shows a first indication of diversity. The table shows the total distinct number of venues that were visited, as well as the relation between the distinct venues and the total number of visits. The higher the relative number, the more venues are visited compared to the total number of venues. We see that this holds for the smaller cities Amsterdam and Rome on both Twitter and Instagram.

In Figure 5.6 we plot the venue category distribution for each city. Here we see a number of differences between cities. Amsterdam has the highest percentage of *Arts & Entertainment* visits for both Twitter and Instagram, and also the highest percentage of *Residence* visits for Twitter. In contrast, Amsterdam scores lower on *Food*, with that



(a) Amsterdam (b) London

(c) Paris (d) Rome

Figure 5.7: Distributions of venue popularity for Twitter

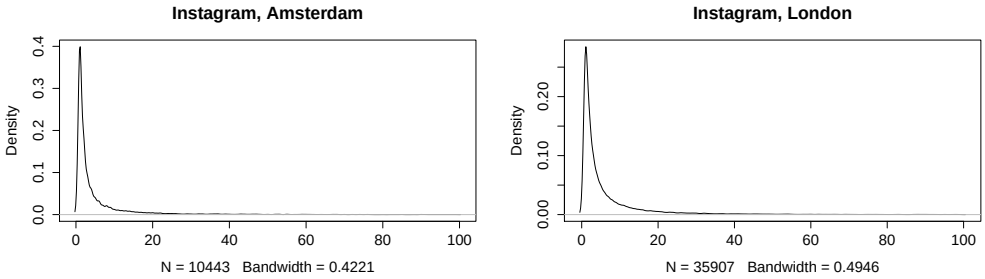
category being the most popular in Rome and Paris. The relatively low percentage for *Nightlife spot* in Rome and Paris is noticeable. This would imply that people are more often visit restaurants, than bars, pubs and clubs, compared to London and for a lesser extent Amsterdam.

Another aspect of diversity of venues, is the popularity (the visit count). Are only a few venues popular or not. In Figure 5.7 and 5.8 we plot the kernel density functions of the popularity of venues for each city, for both Twitter and Instagram. We clearly see that cities that are similar in size (Amsterdam and Rome, London and Paris) have similar density plots. The larger the city the more "even" the distribution is, meaning the larger cities have more popular venues than the smaller cities, and smaller cities have more venues which are only visited a handful of times. This seems to correspond to the relative numbers in Table 5.3, which implied that more venues are visited in the smaller cities than in larger ones. In terms of *diversity* this means that people in the smaller cities Amsterdam and Rome visit more distinct venues.

5.3.4 Can we characterize activities of cities based on the venue categories?

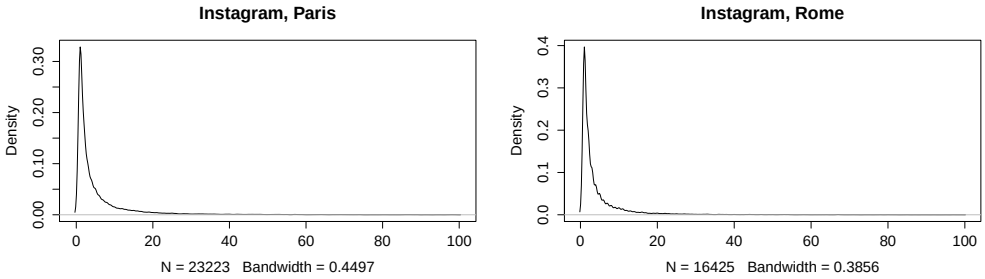
One of the main questions we ask ourselves is how a city is *used*. We will study this by looking at two aspects, *when* are people most active, and *what* do they visit.

We show the overall activity of visits during the day in Figure 5.9. On Twitter we see that Amsterdam and London have a similar level of activity over the whole day, while Paris and Rome peak during the late evening. On Instagram however the activity for all four cities is similar, with only Amsterdam showing again that it is



(a) Amsterdam

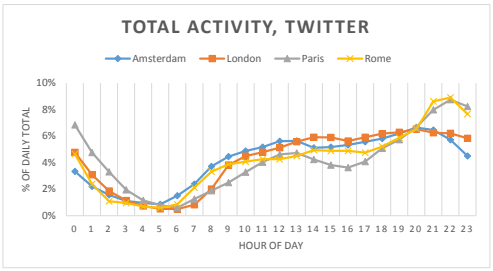
(b) London



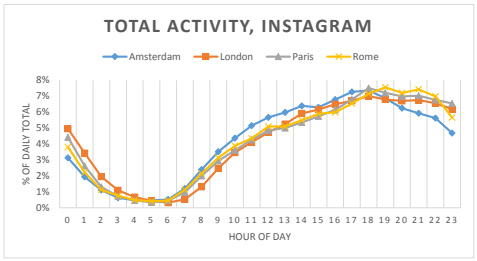
(c) Paris

(d) Rome

Figure 5.8: Distributions of venue popularity for Instagram



(a) Twitter

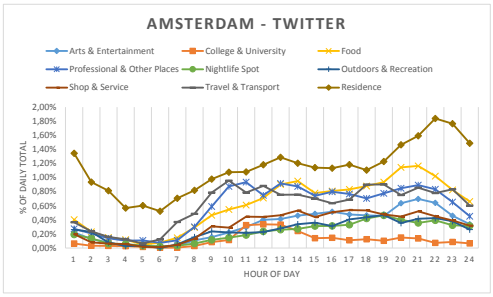


(b) Instagram

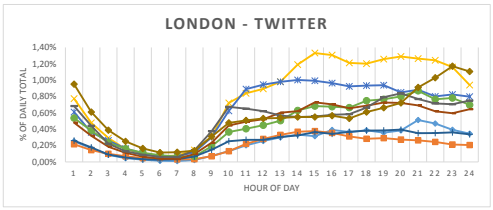
Figure 5.9: Overall activity during the day for Amsterdam, London, Paris, and Rome, on Twitter and Instagram

more active during the day, with a steady decline in activity after 18:00. To gain a better understanding of what these visits are, we look at the *what* aspect of the visits. In Figure 5.10 we show the activity for all the venue categories for each city in Twitter, and in Figure 5.11 for Instagram. To better compare the activity of the venue categories per city, we plot the activity of a selection of categories in each city, *Food* in Figure 5.12, *Residence* in Figure 5.13, and *Nightlife Spots* in Figure 5.14.

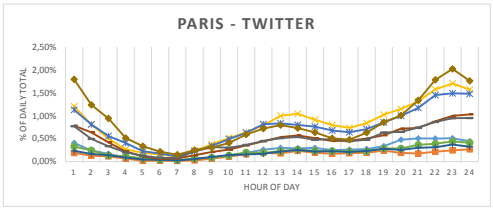
The main difference we see in regards to the platforms, is the large amount of *Residence* posts in Twitter where it is one of the most active categories in all four cities. Whereas on Instagram this category is one of the least popular. We see a



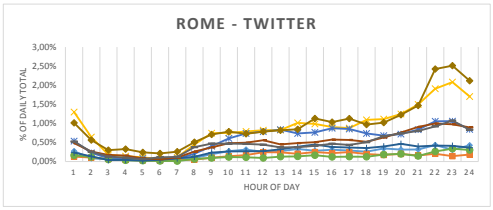
(a) Amsterdam



(b) London



(c) Paris



(d) Rome

Figure 5.10: Distribution of venue categories during the day for Amsterdam, London, Paris, and Rome, on Twitter

expected activity for Residence in Figure 5.13 for Twitter, where the category is most active in the evening and night. For Instagram we see a much less clear trend with a steady rise of activity over the whole day.

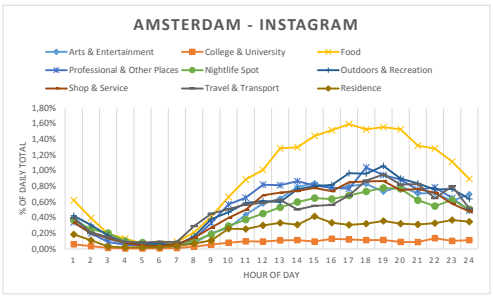
The most popular category for all cities on both Twitter and Instagram is *Food*. Now we can see that the decrease of activity in Amsterdam that we saw in Figure 5.9 is influenced by the similar drop in activity in the *Food* category. It should be noted that Coffee Shops are part of the Food category. These are a very popular destination in Amsterdam, and can almost be considered a tourist attraction. This explains why the Food category shows different behavior compared to the other three cities.

The *Nightlife spot* category (bars, pubs, clubs) shows an interesting difference on Twitter. We see in Figure 5.14 that this category is in fact more popular during the day both in Amsterdam and London. On Instagram the difference is smaller, but still noticeable.

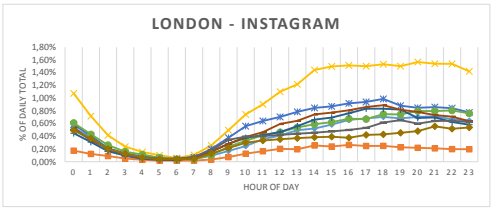
5.4 User Analysis

Now we will aim our attention to the users of the cities, and how they differ between cities and platforms.

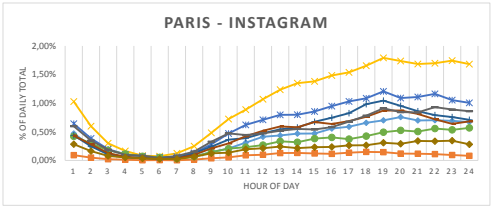
We begin by summarizing the results of the gender recognition in Table 5.4. We managed to successfully determine the gender of 41% of the active users on Instagram,



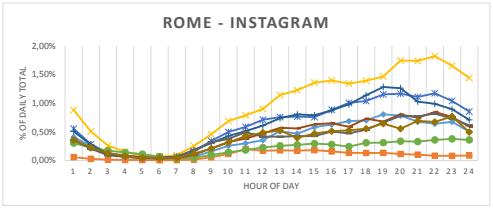
(a) Amsterdam



(b) London

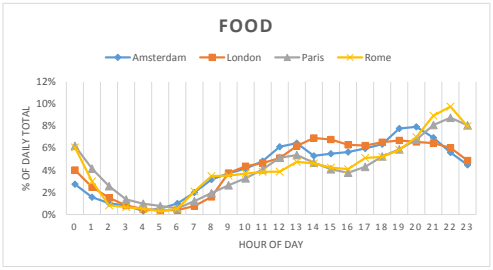


(c) Paris

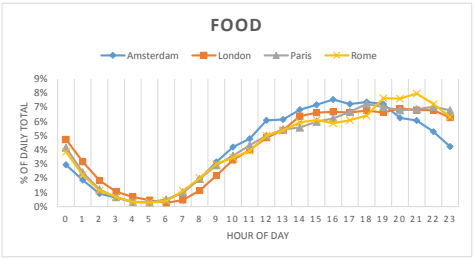


(d) Rome

Figure 5.11: Distribution of venue categories during the day for Amsterdam, London, Paris, and Rome, on Instagram

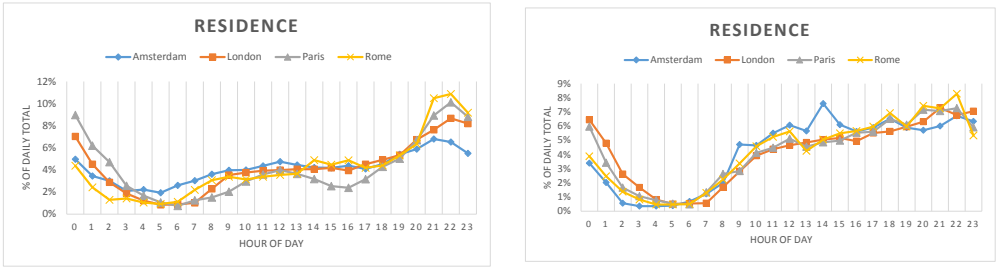


(a) Twitter



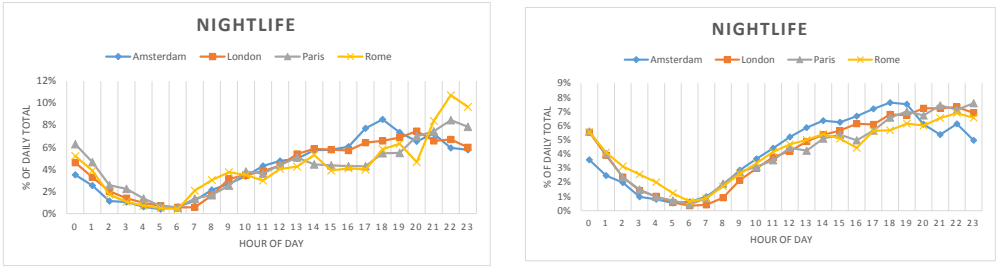
(b) Instagram

Figure 5.12: Overall activity of *Food* category venues during the day.



(a) Twitter (b) Instagram

Figure 5.13: Overall activity of *Residence* category venues during the day.



(a) Twitter (b) Instagram

Figure 5.14: Overall activity of *Nightlife Spots* category venues during the day.

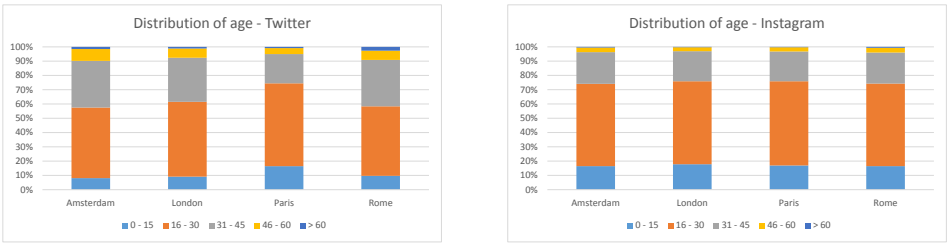
and 58% of the active users on Twitter. We see that the distribution of gender does not differ significantly between cities. Between platforms there is a significant difference however, with Twitter a larger percentage of male users, whereas on Instagram there are more female users.

In Figure 5.15 we show the distribution of age across the four cities. The largest differences in age between cities we see in Twitter. Here Paris has a considerable larger amount of visitors in the age ranges 0-15 and 16-30, and a much lower amount in the range 31-45 compared to the other three cities. On Instagram the differences between cities are much less apparent. We do see however that for Instagram a much larger amount of visitors are between the ages of 0 and 30, showing that the Instagram users are younger than Twitter users.

Finally we show the distribution of the user roles in Figure 5.16. On Instagram we see a much larger percentage of Foreign Tourists than on Twitter for all cities. Amsterdam has the largest share of tourists that are active in the city, nearly 80% on both Twitter and Instagram. For London we see on both platforms that it has the highest percentage of Residents. London is by far the biggest city under analysis, and this has its influence here, there are simply much more people who live in London compared to the number of tourists.

	Twitter		Instagram	
	Female	Male	Female	Male
Amsterdam	42%	58%	57%	43%
London	45%	55%	61%	39%
Paris	50%	50%	61%	39%
Rome	46%	54%	60%	40%

Table 5.4: Distribution of gender



(a) Twitter (b) Instagram

Figure 5.15: Distribution of age

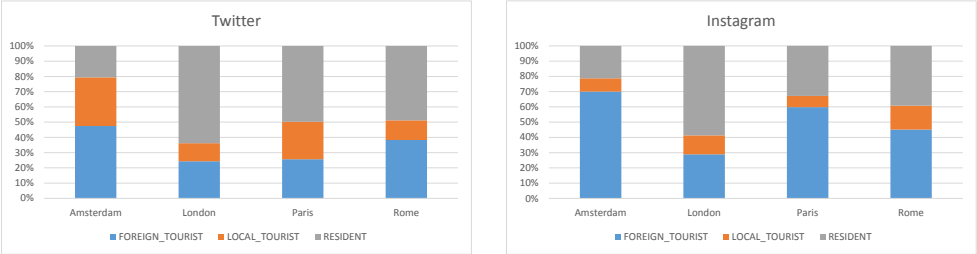
Knowing who is a foreign tourist, we can characterize a city based on the different nationalities of the tourists. For each city we selected the top 15 countries of origin of *foreign* tourists, and plotted them in Figure 5.17. We see people from the United Kingdom are the most active tourists in Amsterdam, Paris, and Rome. The United States and France are also well represented in each city. Russia and China represent approximately 10% and 5% of the foreign tourists on Instagram, however on Twitter their share is negligent, with only 2% for Russia, and China is in none of the top 15 countries. On first glance this is remarkable for such large countries, however this can most likely be explained the presence of much more popular micro-blogging services in those countries (Vkontakte in Russia [12], Sina Weibo in China [22]), whereas in western countries Twitter is by far the most popular such service. This is an important factor to keep in mind when dealing with these distributions of foreign tourists.

It is also interesting to see that the four cities do not differ significantly in regards to the countries of origin, given the fact that there are only 20 different countries in the top 15 of all 4 cities on Twitter and 18 on Instagram. Overall Instagram has a more diverse set of countries as seen in Table 5.5.

5.4.1 Are there differences between mobility between genders across cities?

Now that we have determined the demographics of our active users, we can study how these demographics influence mobility.

One of the measurements of mobility that we defined is radius of gyration. We will look into the gender of our users, and if we can see differences in mobility, and if



(a) Twitter (b) Instagram

Figure 5.16: Distribution of user roles

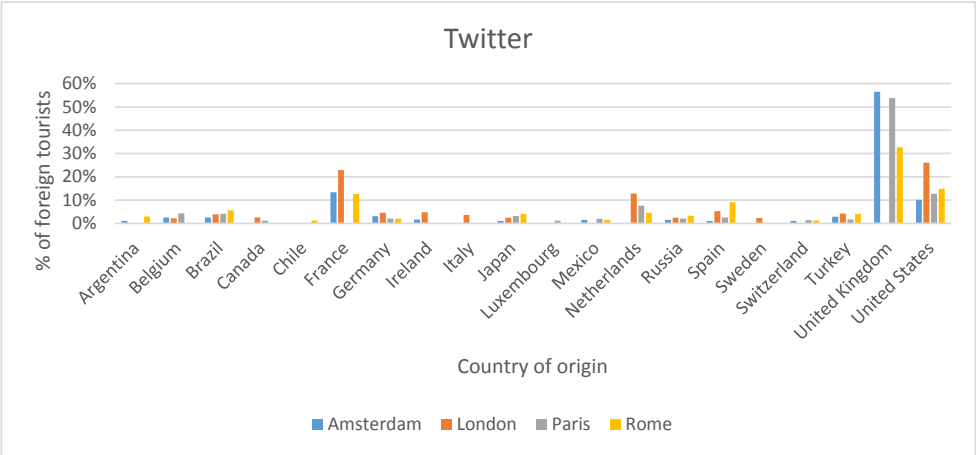
	Twitter	Instagram
Amsterdam	60	79
London	71	85
Paris	63	83
Rome	60	93

Table 5.5: Number of distinct countries of foreign tourists in Amsterdam, London, Paris, and Rome.

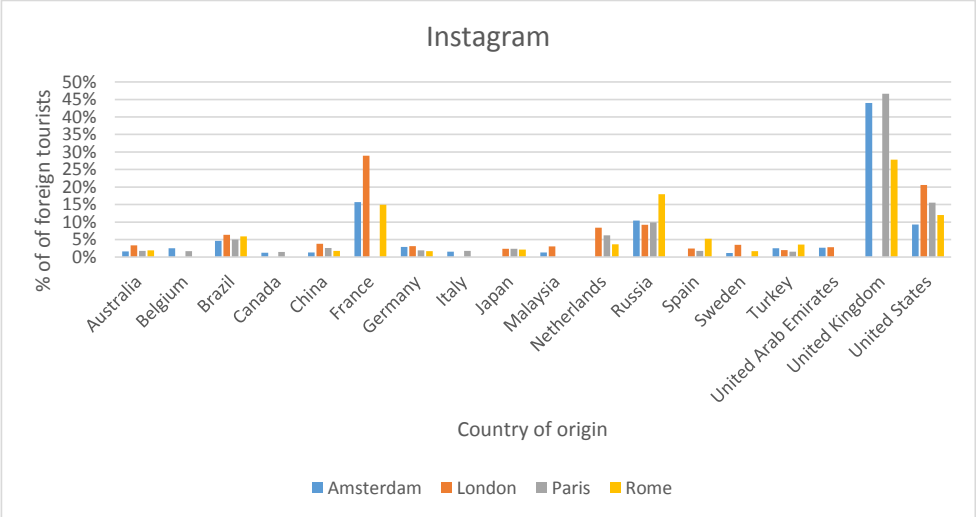
there is a difference between the mobility of genders across cities. One would expect that in a country where there is a very traditional male-female hierarchy, the mobility of females would be lower than those of males. In other words, one could give an indication of emancipation based on mobility.

To this end we calculated the average radius of gyration for each gender, for users *whose home city* is the city under study. We present these results in Figure 5.18, and more detailed in Table 5.6. The distribution of the radius of gyration resembles a power law distribution, corresponding with the findings by Cheng et al.[8], explaining the very large standard deviations. The overall average radius of gyration on Twitter and Instagram correspond to the user roles, where we saw that Instagram has a larger percentage of tourists, resulting in a larger radius of gyration. Comparing cities, we see that while Amsterdam is the smallest city of the four, it has a considerable high radius of gyration on both platforms. This, and the noticeably low radius of Rome in Instagram are interesting findings that could warrant further research. Where do the people of Amsterdam travel to? Do they work further away from home than in other cities? Or is the high radius because of longer trips abroad?

The differences between males and females however are much less descriptive. The differences are very small, and Twitter and Instagram give opposite results for each city. This is further illustrated in Table 5.7 where we show the p-values of a two-sample Kolmogorov–Smirnov (K-S) test we performed comparing the male and female distributions of each city, where H_0 is defined as the two samples are drawn from the same distribution. For London and Paris on Twitter, and Amsterdam and Rome on Instagram $p < 0.05$. This means that for the other combinations there is

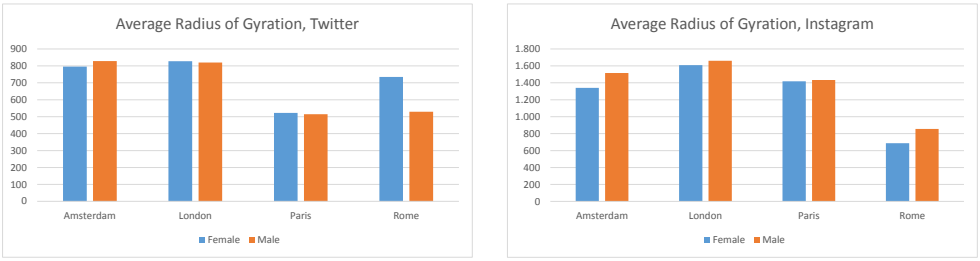


(a) Twitter



(b) Instagram

Figure 5.17: The distributions of the countries of origin of the foreign tourists in Amsterdam, London, Paris, and Rome.



(a) Twitter (b) Instagram

Figure 5.18: Radius of gyration for male/female residents of Amsterdam, London, Paris, and Rome.

		Twitter			Instagram		
		mean	**sd**	**median**	**mean**	**sd**	**median**
Amsterdam	Female	796.02	1,322.18	117.94	1,341.32	1,566.01	622.63
	Male	828.89	1,400.26	173.70	1,515.68	1,556.64	954.71
London	Female	828.02	1,359.53	130.13	1,608.94	1,784.17	838.68
	Male	819.97	1,296.57	163.65	1,660.73	1,787.68	916.86
Paris	Female	522.49	1,094.80	95.53	1,418.52	1,576.89	653.49
	Male	514.43	1,252.11	60.44	1,433.55	1,651.90	614.56
Rome	Female	735.05	1,253.09	126.51	687.27	1,192.64	216.23
	Male	528.98	981.05	108.37	855.10	1,284.69	300.87

Table 5.6: Mean, standard deviation and median of the radius of gyration for male/female residents of Amsterdam, London, Paris, and Rome.

	Twitter	Instagram
Amsterdam	0.684	0.036
London	0.046	0.548
Paris	0.048	0.369
Rome	0.097	0.004

Table 5.7: P-values of a two-sample K–S test between the male and female radius of gyrations for each city.

no significant difference between male and female radius of gyrations. Although even here we see that there is no overall conclusion to draw, as the values differ for each city and platform. For example, in Amsterdam Twitter has a high p-value, but on Instagram a very low one, however for London and Paris Instagram has a much higher p-value.

The small and contradictory results are most likely because of the data source we use. People who use social media are most likely more “modern” and emancipated. Social media are therefore less suited to study gender differences in terms of mobility.

	Twitter		Instagram	
	Paths	Path Patterns	Paths	Path Patterns
Amsterdam	1375	54	2900	88
London	17087	799	17970	945
Paris	10997	795	13521	994
Rome	1952	653	3896	469

Table 5.8: Extracted paths and path patterns

5.4.2 Is there a significant difference between the social activity of residents and tourists across cities?

The user roles we defined in relation to a city can provide a better understanding in how different people use a city. In this section we will study the differences in (social) activity between the residents, local tourists, and foreign tourists, again focusing first on the *when* and then on *what*.

Figure 5.19 and 5.20 show the activity during the day of each user role in each city. The first thing we notice is that the activity between roles differs much more in Twitter, than on Instagram.

We will therefore focus our discussion of these results on the Twitter activity. We see that *Local Tourists* in all four cities post the most during the day between 12:00 and 15:00. Amsterdam is the only city where local tourists have a second significant activity spike during the evening. This might be explained by the geography of The Netherlands, where many cities are close by compared to other countries, and thus people are more inclined to for example have a night out in another city. This can be further investigated by looking at the venue categories that are visited by local tourists.

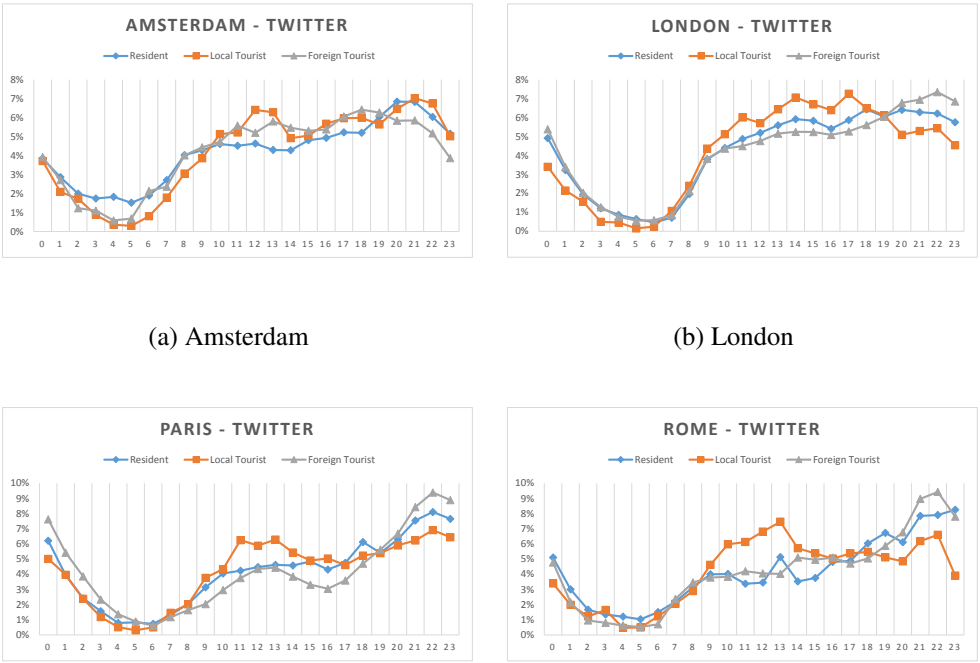
The *Foreign Tourists* in Paris, London, and Rome are the most active during the late evening, while in Amsterdam there is a decline as we have seen before.

Residents in Amsterdam are more active during the night compared to other cities, but the increase in activity in the morning already begins at 06:00, leveling out around 09:00, while in London, Paris, and Rome this happens an hour later. Amsterdam residents can be considered early risers, while the residents of Paris and Rome go to bed later and also wake up later.

Now like we did in the Venue Analysis section we will also look into *what* these different user roles have visited. For the sake of brevity we will only look at the differences between groups, and not cities. We show this in Figure 5.21. Finding significant differences is difficult. One consistent result is that foreign tourists are consistently more active in the *Travel and Transport* category. This category consists mostly of public transportation venues (train stations, airports), as well as hotels, so this result makes sense. We also see that tourists, both foreign and local, are more active in *Arts & Entertainment* than residents.

5.5 Path Analysis

To characterize the mobility of a city as a whole, we defined paths and path patterns.



(a) Amsterdam

(b) London

(c) Paris

(d) Rome

Figure 5.19: Activity of different user city roles during the day for Amsterdam, London, Paris, and Rome, on Twitter

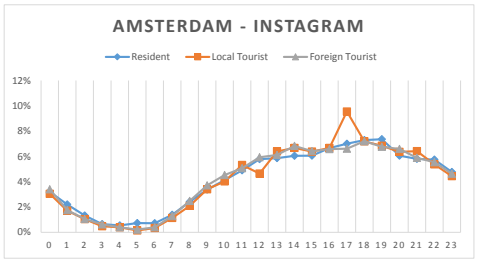
In Table 5.8 we show the number of paths and path patterns we extracted from all four cities. Instagram overall has more paths than Twitter, and the same holds for path patterns (with the exception of Rome). When a city has many path patterns, it implies that there is a larger set of paths that are frequently traversed by users. In that sense the number of path patterns is another indication for the aforementioned *diversity* of city usage. Out of these results Rome is most noticeable for having a high number of path patterns compared to the total number of paths. This is in agreement with our findings in Section 5.3.3 where we concluded that people visit more distinct venues in Rome. We see now that this is also the case for paths users traverse, they are more diverse than the other cities.

5.5.1 How do different demographics of people influence the paths they take?

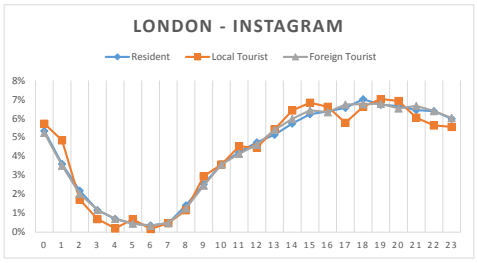
Building on our previous question of how demographics influence mobility, we will now look at how demographics influence the paths the people take. More specifically, we will study the user city role with relations to paths.

In Figure 5.22, 5.23, and 5.24 we show the average duration, length, and number of PoI's on paths for each city, for each user role.

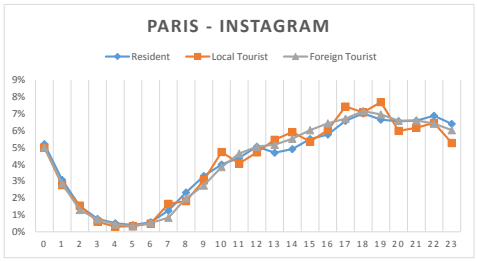
Looking at the duration of paths in cities, the most telling results are the local



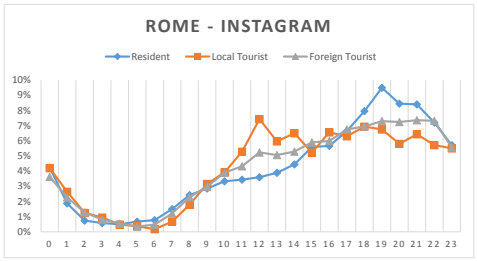
(a) Amsterdam



(b) London

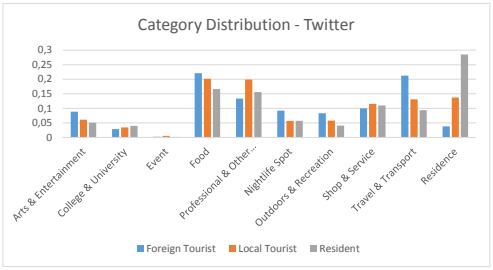


(c) Paris

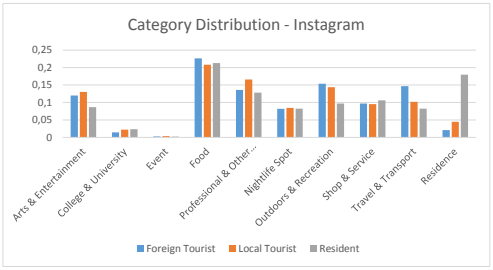


(d) Rome

Figure 5.20: Activity of different user city roles during the day for Amsterdam, London, Paris, and Rome, on Instagram

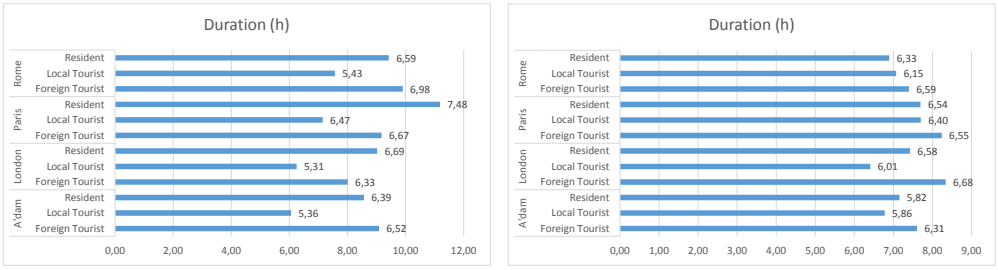


(a) Twitter



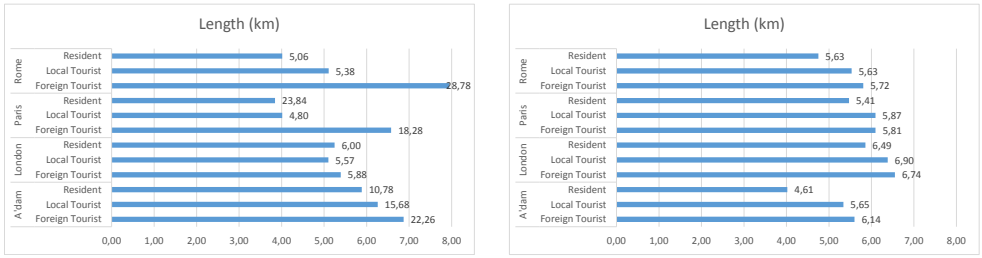
(b) Instagram

Figure 5.21: Distribution of PoI categories for each user role



(a) Twitter (b) Instagram

Figure 5.22: Average duration of paths in hours for each user role. The value next to each bar is the standard deviation.



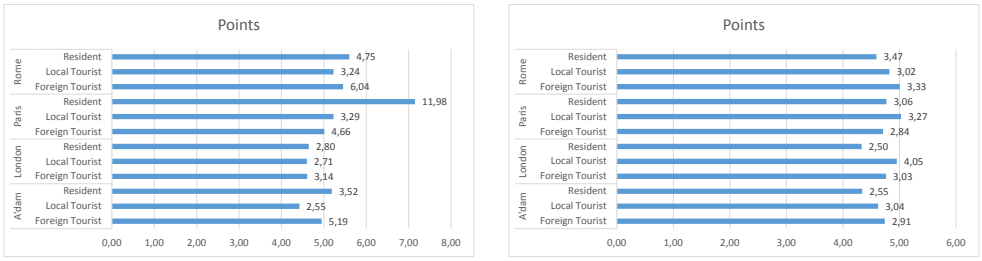
(a) Twitter (b) Instagram

Figure 5.23: Average length of paths in km for each user role. The value next to each bar is the standard deviation.

tourists, where we can clearly see they spend less time in the city than foreign tourists.

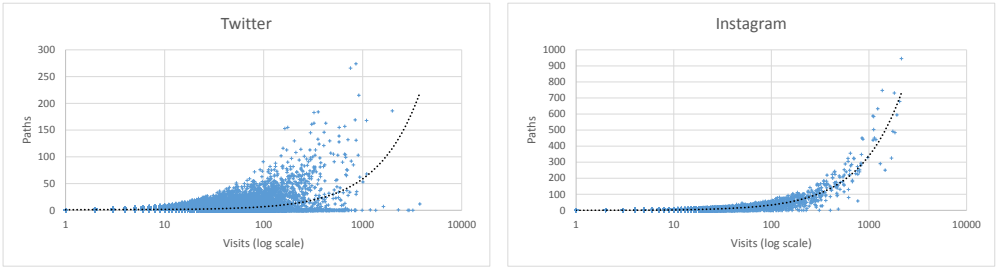
We see that in regards to the length of the paths, we see that the tourists (Foreign and Local) traverse the longest paths. It is also noticeable that the average length of paths between cities does not differ significantly, especially on Instagram. This means people do not necessarily traverse longer distances if they are in a larger city. Comparing the lengths of Twitter paths and Instagram paths, we see that the lengths in Instagram do not differ significantly, except for residents in Amsterdam. The standard deviations of the lengths on Instagram are also similar, while they vary much more on Twitter. The

The number of PoI's on paths show less differences. For each user role, for each city the number of PoI's are between 4 and 5. The standard deviations also indicate small differences and fairly even distributions. The outlier here are residents in Paris on Twitter, this is most likely the result of a few very active users (many posts per day and every day) which raises the average of the full population.



(a) Twitter (b) Instagram

Figure 5.24: Average number of PoI's on paths for each user role. The value next to each bar is the standard deviation.



(a) Twitter (b) Instagram

Figure 5.25: Each PoI with the number of visits of that PoI vs the number of occurrences on paths

5.5.2 Is there a significant relation to the popularity of a PoI, and how often they are on a path?

Trying to learn more about the collected paths, we will now look at the PoI's on those paths. We will look into how the popularity of a PoI influence the paths of a city.

In Figure 5.25 we plot each PoI in our dataset (all cities combined), and see how the popularity of that PoI (ie. amount of visits) relates to the number of times this PoI is on a path. We see that for Instagram this relation is almost exactly linear, with very few outliers. The distribution for Twitter is more diverse where we see more venues that have many visits, but are not on any paths. It seems that on Instagram, almost every visit to a venue is part of a path.

5.6 Path Patterns Analysis

We now take a closer look at the path patterns we extracted. To give a taste of the kind of patterns that are found, we show the top 10 path patterns for both Twitter and Instagram that we found in Amsterdam and London in figures 5.26 and 5.27.

Dam → Madame Tussauds (22)
Amsterdam → Dam (20)
Madame Tussauds → Dam (19)
Amsterdam → Heineken Experience (17)
Dam → Amsterdam (16)
I Amsterdam → Heineken Experience (14)
Rijksmuseum → I Amsterdam (13)
I Amsterdam → Rijksmuseum (13)
Amsterdam → I Amsterdam (12)
Amsterdam → Anne Frank Huis (12)

(a) Instagram

Station Amsterdam Centraal → Amsterdam Airport Schiphol (AMS) (16)
Amsterdam Airport Schiphol (AMS) → Station Amsterdam Centraal (16)
Amsterdam Eco Tours → Hunter's Coffeeshop (10)
Hunter's Coffeeshop → Amsterdam Eco Tours (9)
McDonald's → Ten Katemarkt (8)
Ten Katemarkt → McDonald's (8)
Praxis → CS Digital Media (7)
Umeno → CS Digital Media (7)
Broederooooaagestt → Umeno (7)
Umeno → Broederooooaagestt (7)

(b) Twitter

Figure 5.26: Top 10 path patterns in Amsterdam for Twitter and Instagram

We immediately see a siginificant difference in the venues that are on these patterns. For Instagram we see well known (touristic) venues, while the venues on the Twitter patterns are less well known.

5.6.1 In what way do the paths on different social media platforms differ?

To further answer one of our main research questions regarding the difference of location data between Twitter and Instagram, we will now investigate what kind of patterns the different platforms produce. We already saw the different kind of venues in the top 10 path patterns in Amsterdam and London. We take a more in-depth look at these differences by studying the venue categories on path patterns, the popularity of venues on path patterns, and the support and length of path patterns.

Venue categories

In Figure 5.28 we show the distribution of venue categories that are on path patterns (for all cities combined). We see that the distributions are significantly different. The large percentage of *Outdoors and Recreation* on Instagram compared to Twitter makes sense, as pictures are often made outside. Looking at the subcategories of this root

Big Ben → The London Eye (109)
The London Eye → Big Ben (81)
Royal Academy of Arts → Sensing Spaces (58)
Sensing Spaces → Royal Academy of Arts (56)
Royal Academy Of Arts: Building The Revolution → Sensing Spaces (40)
Buckingham Palace → Big Ben (40)
Big Ben → Westminster Bridge (39)
Sensing Spaces → Royal Academy Of Arts: Building The Revolution (38)
Westminster Bridge → Big Ben (38)
Olympia Conference Centre → Olympia Grand Hall (36)

(a) Instagram

Chick Freak → Anar Persian Kitchen (19)
Chick Freak → Athlone Gardens,Anar Persian Kitchen (19)
Athlone Gardens → Anar Persian Kitchen (19)
Chick Freak → Athlone Gardens (19)
Chick Freak → Anar Persian Kitchen,Athlone Gardens (19)
Anar Persian Kitchen → Athlone Gardens (19)
Athlone Gardens → Chick Freak → Anar Persian Kitchen (18)
Anar Persian Kitchen → Chick Freak (18)
Athlone Gardens → Chick Freak (18)
Anar Persian Kitchen → Chick Freak → Athlone Gardens (18)

(b) Twitter

Figure 5.27: Top 10 path patterns in London for Twitter and Instagram

category we see that *Plaza* and *Bridge* are the most popular venue types on Instagram (Bridge is popular because of the Tower Bridge in London). *Arts and Entertainment* is a popular category on both platforms, which is especially interesting when we compare it to the overall popular venue categories that we showed in Figure 5.5, where this category has a low activity.

PoI popularity and frequency

The difference we saw in overall category activity compared to the occurrence of those categories in path patterns, could indicates that if (types of) places are popular, it does not necessarily follow that they occur on many path patterns. We have seen that this holds for venue categories, now we will see if this also holds for the popularity of the venues. In order to see how popular the venues on path patterns with a high support are, we calculate the average number of visits of all the venues on patterns of a particular frequency. This is show in Figure 5.25. We clearly see that for patterns of the same frequency, the venues on Instagram are more popular than those on Twitter. This corresponds to our observations of the top 10 path patterns we showed.

Path Pattern Frequency & Size

In Figure 5.30 we plot the frequency (i.e., the support) of path patterns, and in Figure 5.31 we show the distribution of the number of PoI's per path pattern. The overall low

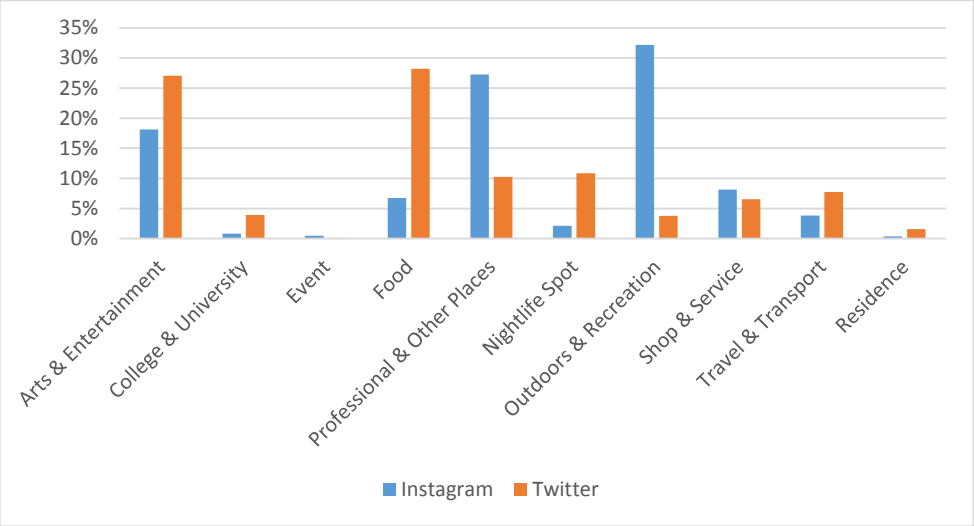


Figure 5.28: Distribution of venue categories on path patterns

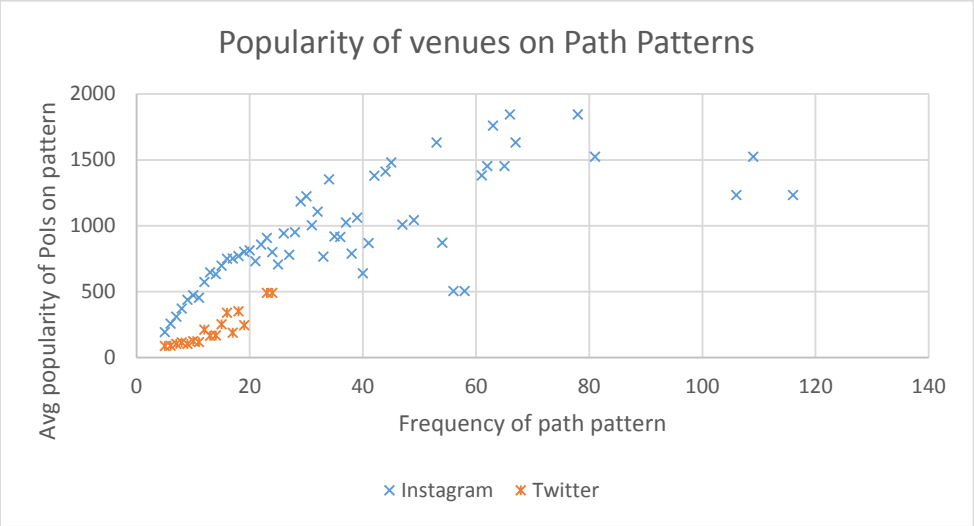
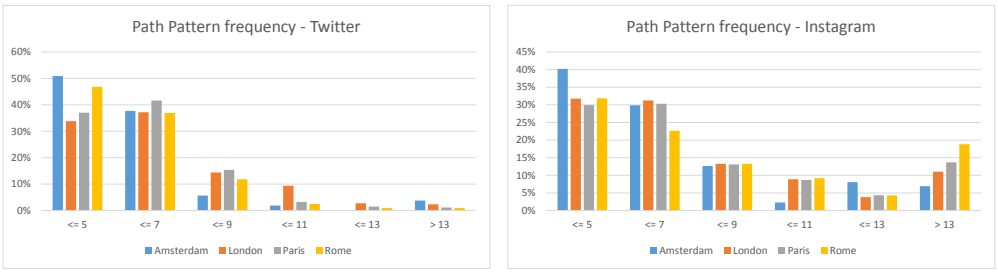


Figure 5.29: Average popularity of venues vs the frequency of patterns

value of the frequencies can be explained by the relatively short timespan in which these patterns were extracted. However, we do again see differences between Twitter and Instagram. The Instagram patterns are overall more frequent, however they are also shorter. The differences in pattern frequency between Twitter and Instagram fit with the other observations we made in the chapter: Twitter users tweet at home, have more activity by residents, and also visit less touristic venues, and travel shorter distances, we now also see this in the path patterns. Twitter has less patterns, they visit less popular venues, they visit less touristic categories and now we also see that the patterns are traversed by less people than on Instagram. The larger number of tourists influences the patterns traversed in a city significantly.



(a) Twitter (b) Instagram

Figure 5.30: The frequency (support) of path patterns in each city, for Twitter and Instagram

5.7 Discussion

In our analysis we saw several differences between the usage of Twitter and Instagram, and the location data the platforms reveal. The daily activity on Twitter peaks around noon, and rises even more during the evening, while Instagram activity is steady during the afternoon and then falls in the evening. This, together with the fact the most popular PoI category is *Residence* (39% on Twitter compared to 12% on Instagram) shows that most tweets are made at home, while the categories on Instagram are more diverse. We also saw that the Instagram user base has slightly larger percentage of female users than on Twitter. Instagram users are also younger. The most noticeable demographic difference we saw between platforms, was in the user roles. In all cities the percentage of foreign tourists was 10-30% higher on Instagram than on Twitter.

This combined with the interesting differences in the extracted path patterns, where we saw that Twitter patterns have less popular venues, and are also much less frequent than Instagram patterns, show that Instagram location data is much more focused on main attractions and landmarks in cities, while Twitter location data is more ”personal" in nature, and thus the locations visited are not necessarily important in the city.

All these findings show us that in the context of city analysis, Instagram is more useful when interested in popular attractions in the city, especially those interesting for

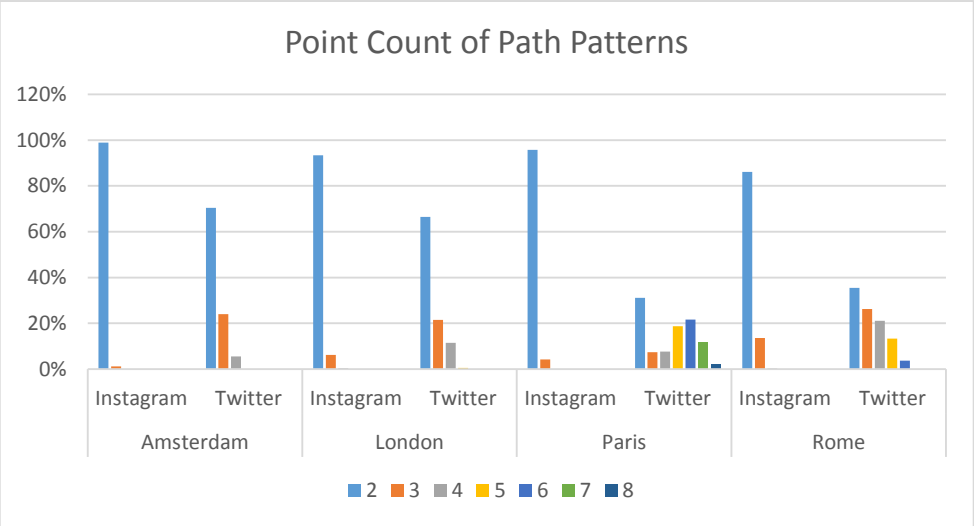


Figure 5.31: The number of PoIs on path patterns in each city, for Twitter and Instagram

tourists. Twitter is a more suited indication of every day life, with its high percentage of resident users.

We compared the cities of Amsterdam, London, Paris, and Rome. The most noticeable difference we discovered, is the activity in the evening hours. We saw that Amsterdam slows down after 21:00, while Paris and Rome have the highest percentage of activity in the evening. In Amsterdam the *Residence* category is the most active during the entire day, while in London, Paris, and Rome we only see that category peak in the evening. Another interesting finding, was that we saw no significant differences in the paths users traverse between cities, in regards to size, length and number of PoIs. These properties are thus not significantly affected by the size of a city.

Chapter 6

Conclusions

In this work we have designed and presented a system for city analysis using social media. The system consists of an extraction part, which can extract several properties regarding points of interests, users, and paths, that can be used for city analysis, as well as a visualization tool and analysis part. The system is designed to be extensible, which allows us to further expand the functionalities of the system in future work.

In our implementation we have used Twitter and Instagram as data sources. We used Foursquare venues as points of interest, and used the Foursquare API to map tweets and Instagram posts to Foursquare venues.

We tested our system using real world data, by analyzing the Twitter and Instagram usage over a three week period in Amsterdam, London, Paris, and Rome. This analysis showed promising results, both in showing the differences between Twitter and Instagram location data, and in the ability to discover significant differences in user city usage.

Overall we can conclude that our approach for using social media for city analysis gives encouraging results. Even in the small subset of possible combinations of analysis one can do with the attributes we have extracted from social media, we already discovered significant differences between cities.

In future work we aim to improve the validity of the attributes we extracted such as the venue mapping, and the age and gender recognition. We also intend to extend our visualization tool to enable it to provide more views of the data, and present some of the analysis we performed automatically. In addition to this it also our goal to extend the analysis framework itself, by for example also including the content of the social media posts, in order to better support domain-specific analysis of social media. Possible fields of application could be transport as well as environmental monitoring.

Bibliography

[1] Marco Balduini, Alessandro Bozzon, Emanuele Della Valle, Yi Huang, and Geert-Jan Houben. Recommending venues to visitors of city scale events by continuous predictive social media analytics. *IEEE Internet Computing*, page 1, 2014.

[2] Marco Balduini and Emanuele Della Valle. Tracking movements and attention of crowds in real time analysing social streams the case of the open ceremony of london 2012. 2012.

[3] Marco Balduini, Emanuele Della Valle, Daniele Dell'Aglio, Mikalai Tsytsarau, Themis Palpanas, and Cristian Confalonieri. Twindex fuorisalone: Social listening of milano during fuorisalone 2013. In *The Semantic Web: ESWC 2013 Satellite Events*, pages 327–336. Springer, 2013.

[4] Jie Bao, Yu Zheng, and Mohamed F Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 199–208. ACM, 2012.

[5] Ranieri Baraglia, Cristina Ioana Muntean, Franco Maria Nardini, and Fabrizio Silvestri. Learnext: learning to predict tourists movements. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 751–756. ACM, 2013.

[6] Murat Ali Bayir, Murat Demirbas, and Nathan Eagle. Discovering spatiotemporal mobility profiles of cellphone users. In *World of Wireless, Mobile and Multimedia Networks & Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a*, pages 1–9. IEEE, 2009.

[7] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[8] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. Exploring millions of footprints in location sharing services. *ICWSM*, 2011:81–88, 2011.

[9] Justin Cranshaw, Raz Schwartz, Jason I Hong, and Norman M Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM*, 2012.

[10] P. Fournier-Viger, A. Gomariz, A. Soltani, and T. Gueniche. Spmf: Open-source data mining library. `http://www.philippe-fournier-viger.com/spmf/`, 2013.

[11] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 330–339. ACM, 2007.

[12] Alyssa Kritsch Hootsuite. Hootsuite - the state of social media in russia. `http://blog.hootsuite.com/social-media-in-russia/`, year=2014. Accessed: 2014-08-10.

[13] Shan Jiang, Joseph Ferreira Jr, and Marta C Gonzalez. Discovering urban spatial-temporal structure from human activity patterns. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pages 95–102. ACM, 2012.

[14] Felix Kling and Alexei Pozdnoukhov. When a city tells a story: urban topic analysis. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 482–485. ACM, 2012.

[15] Ryong Lee and Kazutoshi Sumiya. 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*, pages 1–10. ACM, 2010.

[16] Claudio Lucchese, Raffaele Perego, Fabrizio Silvestri, Hossein Vahabi, and Rossano Venturini. How random walks can help tourism. In *Advances in Information Retrieval*, pages 195–206. Springer, 2012.

[17] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–646. ACM, 2009.

[18] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *IEEE Internationcal Conference on Data Mining (ICDM 2012).*, 2012.

[19] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *IEEE International Conference on Social Computing*, 2012.

[20] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM*, 11:70–573, 2011.

[21] Jian Pei, Helen Pinto, Qiming Chen, Jiawei Han, Behzad Mortazavi-Asl, Umeshwar Dayal, and Mei-Chun Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 0215–0215. IEEE Computer Society, 2001.

[22] Kenneth Rapoza. Forbes - china's weibos vs us's twitter: And the winner is? `http://www.forbes.com/sites/kenrapoza/2011/05/17/chinas-weibos-vs-uss-twitter-and-the-winner-is/`, 2011.

[23] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)*, 19(3):630–643, 2011.

[24] Thiago H Silva, Pedro OS Melo, Jussara M Almeida, Juliana Salles, and Antonio AF Loureiro. A picture of instagram is worth more than a thousand words: Workload characterization and application. In *Distributed Computing in Sensor Systems (DCOSS), 2013 IEEE International Conference on*, pages 123–132. IEEE, 2013.

[25] Shoko Wakamiya, Ryong Lee, and Kazutoshi Sumiya. 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*, pages 77–84. ACM, 2011.

[26] Zhu Wang, Xingshe Zhou, Daqing Zhang, Dingqi Yang, and Zhiyong Yu. Cross-domain community detection in heterogeneous social networks. *Personal and Ubiquitous Computing*, pages 1–15, 2013.

[27] Mao Ye, Peifeng Yin, and Wang-Chien Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 458–461. ACM, 2010.

[28] Zhijun Yin, Liangliang Cao, Jiawei Han, Jiebo Luo, and Thomas S Huang. Diversified trajectory pattern ranking in geo-tagged social media. In *SDM*, pages 980–991, 2011.