What do Geotagged Tweets Reveal about Mobility Behavior?

Pavlos Paraskevopoulos[1] and Themis Palpanas[2]

[1] George Mason University and LIPADE, Paris Descartes University
pparaske@gmu.edu
[2] Paris Descartes University
themis@mi.parisdescartes.fr

Abstract. People's attention tends to be drawn by important, or unique events, such as concerts, demonstrations, major football games, and others. Many individuals are even willing to travel long distances in order to attend events they regard as important. As a result, the everyday patterns that a person has, changes. This includes changes in the normal mobility patterns of this person, as well as changes in their social activities. In this work, we study these phenomena by analyzing the behavior of social media users. We investigate the activity and movement of users that either attend a unique event, or visit an important location, and contrast those to users that do not. Furthermore, based on the online activity of users that attend an event, we study the information that we can extract related to the mobility of these users. This information reveals some important characteristics that can be useful for a variety of location-based applications.

Keywords: social networks, social ties, geolocation, movement, Twitter

1 Introduction

The mobility of people and the reasons that cause them to move has been an interesting research topic [1–3] that could be used in order to lead to more efficient urban planning, and to a better understanding of human behavior with regards to unique events. Unique events, such as concerts or football games, tend to attract much attention, while many people are willing to travel long distances in order to attend them.

These events may also affect the social media activity of the people who attend them, as well as the activity observed in the areas they take place, an aspect that is used by some studies [4–7]. The observed increase in (geolocalized) posts on social media from the location a unique event takes place in [5], indicates that social media users tend to share with their friends moments that make them happy or excited, often times also sharing their locations, creating geotagged posts, contrary to their normal patterns. This increase of the social media activity reveals the social ties [8] created between users, simply as a result of attending the same event.

In this work, we try to understand what forces users to make geotagged posts, by observing their mobility through the geotagged tweets. We also investigate if unique event attendants share normal activity and mobility patterns. Finally, we examine the number of the users needed to reveal some important characteristics such as routes or the shape of a country. In order to achieve our targets, we propose a set of methods, which we evaluate using a dataset consisting of geotagged posts from Twitter.

The contributions we make in this paper can be summarized as follows.

1. We employ user samples of different sizes, and study how the sample size affects the information on the most important mobility and activity patterns of users.
2. We examine the difference of the activity and mobility behavior of people who attend an important event, as opposed to the general user population, and show that attendance of certain events imply increased mobility for these users.
3. We present results indicating that user presence in special events or locations is related to the activity patterns of the user, and increases the likelihood of making geotagged posts.

The results of our analysis can be useful for a variety of applications, such as in marketing. In this case, the advertisers can choose target groups depending on their mobility characteristics, which can in turn be determined by knowing some specific locations and/or events that a user has visited.

2 Related Work

2.1 Social Media and Social Network Analysis

Smart devices give users the opportunity to use social media regardless of their location: house, office, or street. They also give the choice to the user to mark her position when posting a message or a photo, creating geotagged posts. In our study, we concentrate on the analysis of data that derive from Twitter. Twitter is a social network that gives to the user the chance to express feelings, or make comments, by using a 140-character text. Although only around 2% of all the tweets are geotagged [9, 10], these are enough to extract important events and their locations, while also increase the volume of the geotagged information by geolocalizing non-geotagged posts [11].

A study that observes movement of people by checking geotagged tweets is presented by Balduini et al. [12]. In this study, the authors analyze geotagged tweets originated from London, and more precisely close to the Olympic stadium during the Olympic games, and identify the exact movement of the crowd, especially during the opening ceremony. Other works have also proposed tools for the analysis and visualization of such geotagged information [13, 14].

Observing the movement of the crowd is very interesting, but it is not the only question that researchers have tried to address. Some studies focus on the extraction of local events by the analysis of the text posted in tweets. Such

a study is presented by Abdelhaq et al. [15]. The target of this study is to identify local events. In order to achieve its target, it initially uses both geotagged and non-geotagged tweets for identifying keywords that best describe events. Then it keeps only the geotagged tweets and extracts the local events. Another interesting study that uses tweets in order to identify events and to explain social media activity during interesting events is presented in [16]. In [4] the authors try to identify where an earthquake happened by only analyzing the activity on Twitter.

Cho et al. [17] develop a framework for analyzing periodic and not periodic movement of the users of social networks, using data of social networks and mobile data. Another interesting study that identifies both aggregated mobility patterns and mobility patterns for unique users is presented in [18]. The authors of this paper use data from online social media such as Twitter and Facebook in order to get how the popularity of a location affects the destination of the user. In [19], Hu et al. present a method that targets to predict future location based on what a user posts, when it's posted and from which location. On the contrary to the previous methods that rely on Twitter or Facebook, Noulas et al. in their work presented at [20] analyze the spatio-temporal patterns of the users' activity and their dynamics using check-ins from Foursquare.

In [21], Crandall et al. investigate the social-ties two users have, based on the co-occurrences they have at a set of different locations. In order to achieve this, they apply a spatio-temporal probabilistic analysis on geotagged photos collected by Flickr. Finally, a study that analyzes the demographics of the people who participate at the movement "#blacklivesmatter" is presented by Olteanu et al. [22]. In this study, the authors investigate the demographics of the users, creating groups of the users based on their activity. On the contrary to this study, we analyze the activity of the users taking into consideration only their id and their geotagged tweets, achieving a more privacy aware analysis.

We note that the studies mentioned above either do not analyze movement, or if they do so, it is at the granularity of a city. Our target is to analyze the mobility caused by unique events at a country level. Furthermore, we study the characteristics of the activity and mobility patterns of different users, and how these are affected by unique events.

2.2 Studies on the Mobility of the Users

Apart from the studies based on social media and social networks, there are also several studies related to mobile phone usage data, GPS devices, or even bank note distribution, aiming to predict the mobility of users, or to analyze the differences of the activity of an area, based on user movement.

Ashbrook et al. [1] present a two level model that applies a clustering at the location recorded by car GPS devices and a probabilistic model in order to find the next location of the user. The target of this study is to identify the most important locations, while also predicting the movement of users. GPS traces are also used by Krumm et al. [2]. In this work, the authors present their algorithm, which uses a probabilistic model and historical data in order to predict the

destination of the user, while also identifying deviations from the user's normal patterns. Do et al. [23] present a probabilistic model that predicts the location of a user at a future time interval, by using GPS data from smartphone devices. The study presented in [24] identifies mobility patterns based on trajectories that are created from anonymized mobile phone users and the travel distances of each user. The authors of [3] on the other hand, identify a set of features using a supervised method on GPS data, extracting mobility patterns.

Most of the studies previously described operate on GPS data. The study presented by Scellato et al. [25] proposes the "NextPlace" framework, which operates using either GPS or WiFi data. The target is to identify the location of a user, based on a spatio-temporal analysis of the data of the network. Chatzimilioudis et al. [26] present a set of algorithms that use trajectories for achieving a crowdsourcing analysis. Their framework can be applied in both outdoor and indoor environments, while their results target to help in cases such as minimizing of energy consumption of networks. Thanks to the mobile devices, users can call or send messages any time, creating Call Detail Records (CDRs). A study that uses CDRs is presented in [7], targeting to identify mobility patterns, while also explain the differences of the activity of a location based on CDRs and an event dataset. This study operates on cell-tower granularity. Other studies use data that are not so obvious they can reveal the users' mobility. Such an example is the study presented by Brockmann et al. [27] that targets to identify users' mobility by applying a spatio-temporal analysis of the banknote distribution.

All the methods described above target to predict user mobility based on either the individual user's patterns, or on some identified general patterns. In our study, we focus on the analysis of the mobility *differences* between groups of users that share some special characteristics (such as a common location at a specific time interval). As a result, we study deviations of normal patterns, and the reasons these deviations appear. Furthermore, the methods previously described operate on datasets where each user has a lot of points. In our study, in order to achieve our movement analysis, we use geotagged posts from Twitter, which results in a very sparse dataset (a user can have just 1-2 geotagged tweets in a period of 4 months), limiting the amount of available information.

3 Proposed Approach

3.1 Problem Description

The problem we want to investigate in this study is the identification of differences in the social media activity between users that attended an important event (e.g., a concert) and those who did not. In addition, we want to study the reasons that force a user to generate a geotagged message, as well as the corresponding mobility patterns. Finally, we would like to examine the extraction of a sample of users in a social network, that could allow us to reproduce the main routes that the users follow.

In the context of this work, we concentrate on users who attend major events or sights, such as concerts, or an important touristic attraction. Furthermore,

Algorithm 1 Get Representative Sample and Characteristics

INPUT: Temporal and Spatial parameters.
OUTPUT: A representative sample of users and its activity and movement.

1: $P_{WinInterest}, Q_{WinInterest} \leftarrow GetUsers(FGL, win_{ev}, CGL, WinInterest)$ ▷ get the users from the event *location* and the CGL and their activity
2: $users, activity, movement \leftarrow$ Percentage of top uses in P, Q ▷ get the representative users' sample
3: **return** $users, activity, movement$

we focus on Twitter, a social network that has more than 313M users, 80% of which are on mobile devices[3].

3.2 Methodology

In this section, we describe the method we developed for tackling the problems previously described (for the general schema, refer to Algorithm 1). Our method is based on the creation of social ties [8], where as social ties we define the connection between users that may not have common characteristics, except for visiting (independently) the same location during a specific time period. Initially we set the temporal and spatial parameters we are interested in. We then remove the spam and bot accounts based on the observed activity of the account. Finally, we follow the geolocalized posts users send during a predefined period of time. In the following sections, we elaborate on the methods discussed above.

Setting the temporal and spatial parameters.
We start by setting the temporal and spatial parameters we are interested in:

1. FGL: the location the event is going to take place in (Fine-Grain location)
2. win_{ev}: the time window during which users visited FGL
3. CGL: the Coarse-Grain location (e.g., city, country) in which we will observe the movement of users
4. $WinInterest$: the period of time we will follow the users' geotagged posts

Get the Event and CGL users.
In order to get the initial sample of our users, we use the spatio-temporal parameters and we check our dataset for users that posted at least one geotagged tweet from the event location, before, during or after the event (win_{ev}). Afterwards, we get all the geotagged tweets these users posted, for a predefined period of time ($WinInterest$).

Having already extracted the users who attended the event, we get the rest of the users from our CGL that have at least one geotagged tweet during the win_{ev} and they have no tweets from the FGL during this time interval. The steps that we follow in order to get the users we are interested in, are presented in Algorithm 2.

[3] https://about.twitter.com/company

Algorithm 2 Get Users

procedure GetUsers($FGL, win_{ev}, CGL, WinInterest$)
  $U_{FGL,win_{ev}} \leftarrow$ all users at FGL at time-window win_{ev}
  for all $u \in \{U_{FGL,win_{ev}}\}$ **do** ▷ get first sample of users in FGL and their activity
    $P^{u,CGL}_{WinInterest} \leftarrow$ all tweets user u posted from CGL during time-window $WinInterest$
  $U_{CGL,win_{ev}} \leftarrow$ all users at CGL at time-window win_{ev}
  for all $u \in \{U_{CGL,win_{ev}}\}$ **do** ▷ get all users in CGL and their activity
    if u not in $U_{FGL,win_{ev}}$ **then**
      $Q^{u,CGL}_{WinInterest} \leftarrow$ all tweets from user u at time-window $WinInterest$
  $P_{WinInterest} \leftarrow$ SpamFilter($P^{CGL}_{WinInterest}$) ▷ clean spam and bot accounts from $P^{CGL}_{WinInterest}$
  $Q_{WinInterest} \leftarrow$ SpamFilter($Q^{CGL}_{WinInterest}$) ▷ clean spam and bot accounts from $Q^{CGL}_{WinInterest}$
return $P_{WinInterest}, Q_{WinInterest}$

Cleaning the Dataset.
There are a lot of accounts that are either bots sending posts with the same content for a long period of time, or accounts that are sending posts with different content, from the exact same location. These accounts do not offer any useful information for our problem (they actually induce noise), therefore, we filter them out. More specifically, for a given account we check if at least 30% of the posted messages have the same prefix, latitude, or longitude. If an account meets at least two of the three conditions, we filter out the account.

Activity and Movement Comparison.
After the extraction of the datasets of the location place and the CGL, we compare their activity using the cumulative distribution function (CDF). Using the CDF, we can compare the activity between the users who visited the event locations and those who did not. Furthermore, we check the distribution of the other locations they visited during the time interval of interest, $WinInterest$. In order to achieve this, we compare the difference between the maximum and minimum latitude and longitude the user appeared in.

The hypothesis we want to verify using the above analysis is that users tend to travel long distances in order to visit a unique event or a unique location. In addition, we want to verify a second hypothesis, that users are more willing to share their location in case they attend important events, as opposed to their normal activity patterns.

4 Experimental Analysis

In order to evaluate our ideas, we used geotagged posts from Twitter. The datasets we used contain events such as major concerts and important touristic

locations. In this section, we present a set of activity and movement analytics, while we provide the reader with visualizations of the location we get the tweets from.

Dataset Description

For the evaluation of our methods, we used a dataset containing geotagged tweets generated from Italy (as defined by a bounding box) for the period between 1st of June and 20th of October 2016. In this dataset, we have 1.460.083 geotagged tweets, posted by 173.182 unique users. We focused on important locations and events that took place during these time intervals. More precisely, we targeted users who posted geotagged posts from *Vatican* in Rome, and the concert of Bruce Springsteen in Rome, which in our experiments is referred as *Concert*.

4.1 Important Event and Location Activity Analysis

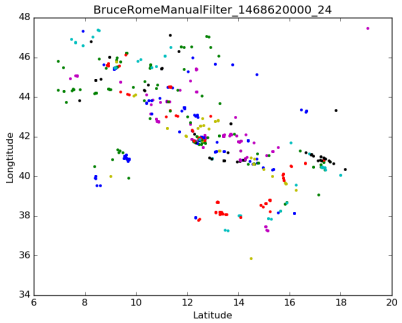
People Attending *Concert*

We initially focused on an important event that took place in Rome and attracted a lot of people. This event was the *Concert* that took place at the location "Circus Maximus" on 16 of July 2016. We found the users that visited this location and posted a geotagged post since the midnight of the previous day. The time windows that we used were 24 hours and 48 hours (it was a 2-day concert), searching for posts initially posted up to the end of the concert (i.e. 24 hours) and afterwards also the following day (i.e. 48 hours). Having identified the users who generated messages from this location during our window, we followed all their geotagged posts for the period between June 1st and October 20th, 2016.

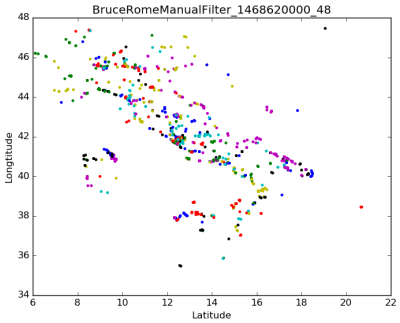
After further analyzing the activity of these users, we found that it was a sample of 67 non-spamming users. The statistics of these users' activity is presented in Table 1. As we can see in this table, when decreasing the number of users in our sample, keeping a percentage of the most active ones, the standard deviation of the activity of the users is not affected much, while the mean activity of the users decreases. This fact implies that the distribution of the activity of the users is similar for the majority of the users in our sample, and especially for the most active users.

In Figure 1a we depict the locations these 67 users "appeared" at, while in Figures 1b,3a,1c we can see respectively the locations the 75%, 50% and 25% most active users posted geotagged tweets from, for the period June to October. In all the plots we present in this section, each color represents a different user[4]. As we can see in Figure 1a, the combination of mobility and activity patterns of these 67 users cover the entire country of Italy. Furthermore, after manually checking the position of highways in Italy, we found out that these 67 users are able to form the main shape and the main routes of the country of Italy. This is still true when we consider the 50% most active of these users (see Figure 3a),

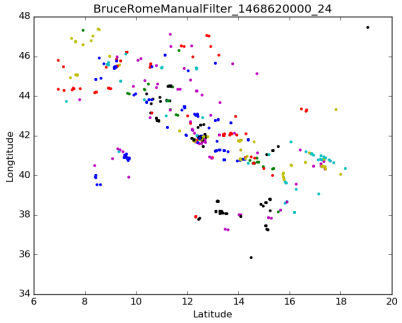
[4] Due to the relatively high number of users, different users may share the same color.



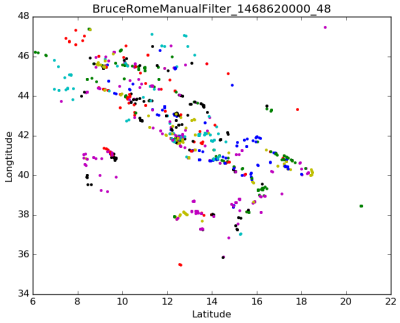
(a) all users (67 users)



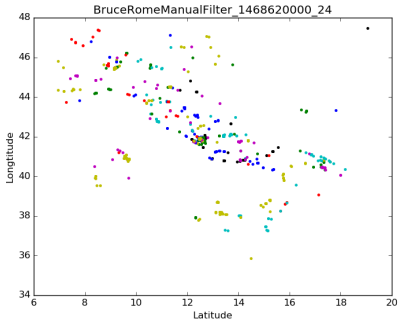
(d) all users (144 users)



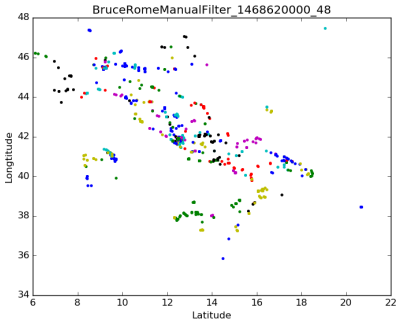
(b) 75% of users with highest activity (50 users)



(e) 75% of users with highest activity (108 users)



(c) 25% of users with highest activity (17 users)



(f) 25% of users with highest activity (36 users)

Fig. 1: *Concert*, 24-hour (left) and 48-hour (right) windows

and almost true even when we limit the number of the users to 17 (25% of most active, Figure 1c).

These results reveal some very interesting characteristics of our dataset and users. They indicate that an extremely small number of users is mobile enough in order to cover the entire country. Recall that the users in the sample we

	Concert (1 day)				*Vatican* (1 day)			
User%	Users	Act%	Mean	Std	Users	Act%	Mean	Std
100%	67	100%	25	31	48	100%	23	42
75%	50	98%	33	32	36	98%	30	47
50%	34	91%	45	32	24	91%	42	54
25%	17	69%	68	33	12	75%	69	69

Table 1: Statistics for most active users of *Concert* and *Vatican* for 1 day

	Concert (2 days)				*Vatican* (2 days)			
User%	Users	Act%	Mean	Std	Users	Act%	Mean	Std
100%	144	100%	19	27	91	100%	25	56
75%	108	98%	25	29	68	95%	32	64
50%	72	95%	36	31	46	91%	45	74
25%	36	75%	57	33	23	79%	79	96

Table 2: Statistics for most active users of *Concert* and *Vatican* for 2 days

examined belong to a particular demographics group, namely, they all attended a specific music concert. Nevertheless, this observation can lead to interesting marketing applications, since we can now target users with particular mobility patterns.

After having checked the activity and the locations of the people identified using the 24-hour window, we analyzed the people identified by the 48-hour window. The volume of the sample was increased to 144 users. The statistics of this sample are presented in Table 2.

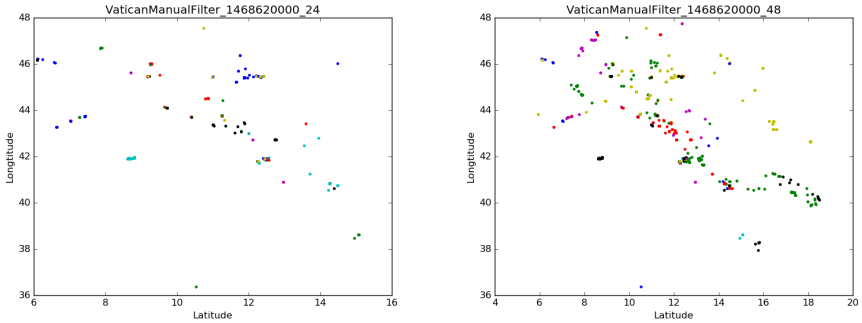
As we noticed in the case of the 24-hour window, sub-sampling with the most active users does not affect much the standard deviation of the activity. Furthermore, the mean activity is slightly decreased compared to the one of the case of the 24-hour window, while the standard deviation is similar. This implies that the activity of the 68 users identified at the concert location during the second day, does not differ to the activity of the users of the first day.

In Figure 1d, we can see the locations of the 144 users identified at the concert for the 48-hour window. The fact that we increased the window, appending users to our dataset, provided us with more geotagged tweets. Due to this, we have more points in our plots, showing more precisely the map of Italy and the main highways. Furthermore, comparing Figures 3a (which is formed by 34 users) and 1f (which is formed by 36 users) we notice that the shape of Italy formed by the 36 users is much more representative. This is due to the fact that the users, whose activity is depicted in Figure 1f, have in general (slightly) higher activity.

Finally, in order to check the impact of the concert to the area, we slightly modified our parameters, targeting users that visited the concert area one week before the concert took place. Even though the area is located in the center of Rome, only 6 users had posted geotagged messages from this location during a 24-hour window. This means that the concert was indeed the reason that the users posted geotagged posts (as we also verified by further analyzing the content of the posts).

People Visiting *Vatican*

Having analyzed the activity of the users who attended an important unique



(a) 24 hours (48 users) (b) 48 hours (91 users)

Fig. 2: *Vatican* Visitors

event such as a concert, we turned our focus on one of the most important locations of Rome, the *Vatican*. We followed exactly the same procedure we did in the case of the concert, modifying only the location whose visitors we were interested in.

After analyzing the activity of the visitors' of *Vatican* using the 24-hour window, we found that 48 users posted geotagged tweets from *Vatican* during this window. The activity of these users is presented in Table 1. Contrary to the case of the *Concert*, the standard deviation of the activity of the users that visited *Vatican* is affected when limiting the sample to the most active users.

In Figure 2, we depict the locations of the users that visited *Vatican*, the same day that the concert was, and posted a geotagged post from *Vatican* for a 24-hour and 48-hour window, consecutively, while in Table 2 we present the statistics of the user sample we took using the 48-hour window. As opposed to the case of the users who attended the *Concert*, the shape of the map of Italy that is formed is not very clear. This difference is more obvious when comparing the Figure 1b, which was created using a sample of 50 users, with the Figure 2a, which is created using a sample of 48 users. In the case of the 48-hour window, this comparison is possible between the Figures 1e (108 users) and 2b (91 users). Possible explanations for this behavior include the fact that the majority of the users who visited *Vatican* are tourists whose home-location is outside of Italy. Nevertheless, these results highlight the different mobility and activity behaviors of these two different samples of users.

4.2 Most active and Random Users from Italy and Rome

Following the analysis of the users who either attended an event (i.e., concert) or visited an important location, we wanted to compare their activity with the people who haven't been located in one of the previous cases. In order to achieve our target, we have identified and followed users either from Italy or from (only) Rome that at the day of the *Concert*, were not located at the location the *Concert* took place. In order to make the comparison fair, we keep only n users,

where n is the volume of the sample of users who attended the *Concert*. Finally, having extracted the appropriate number of the users to be used, we experimented with the cases of the n Random people from the *Concert* and Italy/Rome, and Top n users from the *Concert* and Italy/Rome.

Impressively, after having manually analyzed the user activity, we found out that there is a non-spam user with 26756 tweets that exchanges messages having his location identification "on".

Rome Visitors Compared to *Concert* Attendees

In this part, we present the plots with the comparison of the locations between the n Random and n Top users from the *Concert* and Rome.

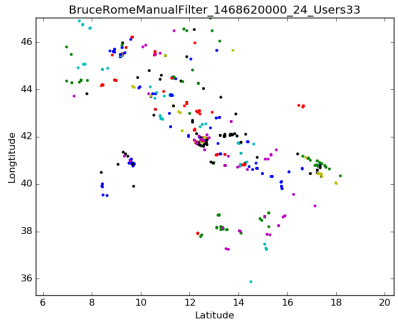
We initially use 50% of the volume of the users posted geotagged post from the location of the *Concert*, where 50% equals to 33 users. The depiction of the 33 users' location is depicted in Figure 3. As we can see in the figure, the representation of both the routes and the map of Italy is very accurate compared to the real map of Italy, regardless of the small number of users our sample has. As expected, in case we increase the volume of the users to 50 (75% of the users attended the concert) or 67 (100% of the users attended the concert), the representation of the map and the routes become even more clear. This relies on the fact that we have more users and as a result, more tweets.

After further analyzing the spreading on the map of the locations the users posted geotagged posts from, we found out that the spreading of the locations of the users who attended the *Concert* is much higher compared to those who posted geotagged post from Rome. This strengthens the assumption we previously did, that the users travel from other locations in order to attend a unique event such as a concert.

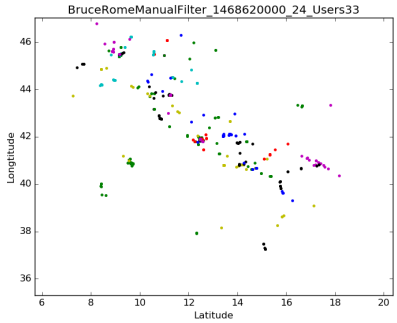
Italy

Having compared the activity and the location between the users of Rome and those of *Concert*, we wanted to compare the n Random and n Top users from the *Concert* and Italy. Similarly to the case of the comparison of the two groups of users in the case of Rome and *Concert*, when using the 50% of the volume of the users who attended the concert (i.e. 33 users), the representation of the map of Italy and the routes are depicted clearly in the case of the most active users of Italy. On the contrary to the case of the users from Rome, even the representation we get using the 33 random users from Italy is much more clear. The reason of this difference relies on the movement of the users of Italy (Figure 3f). In Figures 3a and 3b we can see the representation of the locations of the users.

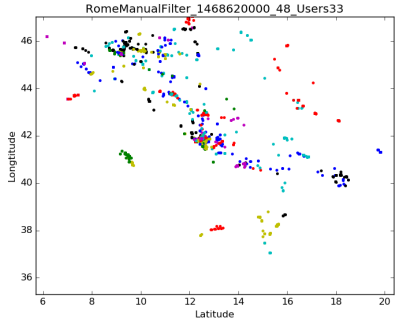
The representation of the routes becomes more clear when increasing the number of the most active users of Italy to 67 (100%). Regardless this increase of the number of the users, the map of Italy is still not as clear as it is in the case of the users who attended *Concert*. After further analyzing the locations of the users from Italy dataset, we find out that the spreading of the locations on the map is still much smaller than the one the users who attended the concert have.



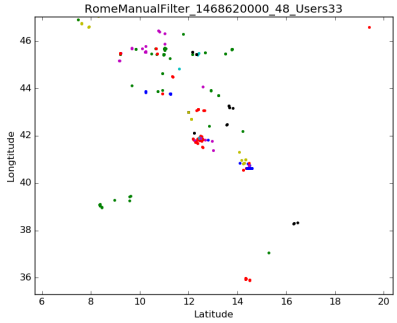
(a) Concert Users (Highest)



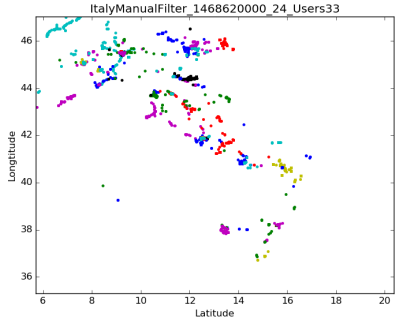
(b) Concert Users (Random)



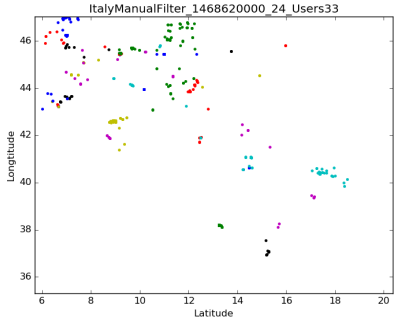
(c) Rome Visitors (Highest)



(d) Rome Visitors (Random)



(e) Italy Visitors (Highest)



(f) Italy Visitors (Random)

Fig. 3: *Concert*, Rome and Italy Visitors (Random 33 VS Top 33)

4.3 Cumulative Distribution Function and Movement

In this part, we investigate the cumulative distribution function and the movement of the users who attended *Concert*, as opposed to those who did not.

As we can see in Figure 4, the activity of the users who attended the concert differs from the activity of the users of Italy. More precisely, the percentage of the users who attended the concert and has a unique tweet is double com-

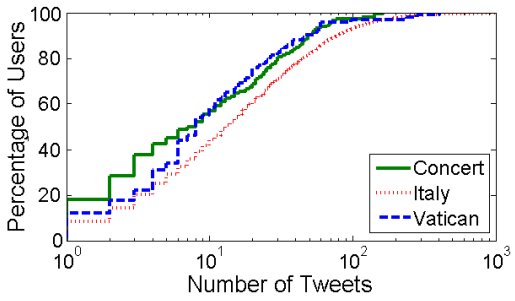


Fig. 4: CDF: Comparison of Number of Tweets

	Lat Dif Median	Lon Dif Median
Concert	247	247
Vatican	218	163
Rome	209	228
Italy	181	231

Table 3: Distances between the furthest locations that users traveled to (in km)

pared to the percentage of the users who were located in Italy, but not in the concert area. Furthermore, we notice that the cumulative distribution function (CDF) of the users who attended the concert is very similar to those who visited $Vatican$, while the same happens with the users who were visiting Rome. After manually checking the tweets of the users who were at $Vatican$ or Rome in general, we found out that the posts generated by the users who had posted only a few geotagged tweets, had been posted from unique locations such as $Vatican$, Colosseum or other historical monuments of Rome.

Furthermore, we compared the movement of the users who attended the concerts and those who did not. We found out that the median difference of the maximum and minimum latitude and longitude that the concert users appeared is 247 km in both dimensions. In the case of the users who were located in Italy but not at the concert the median of the latitude and longitude difference is 181 km and 231 km respectively. Regarding the users of Rome who haven't attended the concert, these numbers become 209 km and 228 km. Finally, regarding the users who visited $Vatican$, the median differences get reduced to 218 km and 163 km. These differences of the locations the users visited indicate that the users who intend to attend a unique event, either have different mobility patterns than the rest of the users, traveling a lot, or they are willing to travel from long distances in order to attend an event such as a concert. In Table 3 we present the differences of the movement each group of users had.

The difference at the locations the users of each group appeared, constitutes one more hint, reinforcing our initial hypothesis that users who attend important unique events, such as concerts, tend to travel from other locations in order to attend the event. Furthermore, these numbers combined with the distribution of the locations the attendees of a concert appear, indicates that the users who attend unique events also tend to travel more.

4.4 Discussion

Overall, our results show that the movement of the users who attend a unique event (i.e. *Concert*) is much higher than the rest of the users, indicating that the attendants of a unique event are willing to travel from long distances in order to attend the event. The differences at the mobility between those who attend a unique event and those that do not attend it, is so high that by depicting the locations where 33 event's users appeared during a period of 4.5 months, is enough to reveal the shape of a country and its highways.

Furthermore, the analysis of the activity patterns of the users, indicate that even though the sample of the users who attend an event shares only one common characteristic (i.e. attended an event), their activities follow specific patterns. The cumulative distribution function indicates that a sizable percentage of the users (around 20%) is willing to post geotagged information from the location the event takes place, which is opposite to their normal patterns. The analysis of the activity of the concert location before and during the concert, reveals the effect that a unique event has to the activity of that location.

The differences between the activity and mobility patterns of the users who visited $Vatican$ and those who attended the *Concert*, indicate that unique events attract visitors that may come from far away, and could be a better choice of advertising when we want to advertise with a country-wide coverage (e.g., electronic devices). On the other hand, unique locations, such as the $Vatican$, attract mostly local visitors, and are a better choice if we want to advertise something that refers to a city level (e.g., a restaurant).

5 Conclusions

In this work, we presented an analysis of the differences of the activity and mobility patterns of people that attend a major event, or visit an important location. Our results indicate that users are willing to travel from far locations in order to attend a unique event. Furthermore, we investigated the number of users needed to identify main routes and locations that attract people. This led to the surprising observation that the mobility and Twitter activity of less than 35 users that attended a unique event is enough in order to shape the main routes and outlines of regions, or countries. Finally, our experimental analysis shows that user presence in special events, or locations (such as an important touristic attraction, or a major concert) affects the normal activity patterns, increasing the likelihood of making geotagged posts. In our future work, we plan to extend our analysis with more locations, events, and time periods.

References

[1] D. Ashbrook and T. Starner, "Using gps to learn significant locations and predict movement across multiple users," *Personal and Ubiquitous computing*, vol. 7, no. 5, pp. 275–286, 2003.

[2] J. Krumm and E. Horvitz, "Predestination: Inferring destinations from partial trajectories," in *International Conference on Ubiquitous Computing*. Springer, 2006, pp. 243–260.

[3] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 2008, pp. 312–321.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[5] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "Tedas: A twitter-based event detection and analysis system," in *Data engineering (icde), 2012 ieee 28th international conference on*. IEEE, 2012, pp. 1273–1276.

[6] P. Paraskevopoulos and T. Palpanas, "Where has this tweet come from? fast and fine-grained geolocalization of non-geotagged tweets," *Social Network Analysis and Mining*, vol. 6, no. 1, p. 89, 2016.

[7] P. Paraskevopoulos, T.-C. Dinh, Z. Dashdorj, T. Palpanas, and L. Serafini, "Identification and characterization of human behavior patterns from mobile phone data," *Proc. of NetMob*, 2013.

[8] S. Wuchty, "What is a social tie?" *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15 099–15 100, 2009.

[9] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook, "Mapping the global twitter heartbeat: The geography of twitter," *First Monday*, vol. 18, no. 5, 2013.

[10] V. Murdock, "Your mileage may vary: on the limits of social media," *SIGSPATIAL Special*, vol. 3, no. 2, pp. 62–66, 2011.

[11] P. Paraskevopoulos and T. Palpanas, "Fine-grained geolocalisation of non-geotagged tweets," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 105–112.

[12] M. Balduini, E. Della Valle, D. Dell' Aglio, M. Tsytsarau, T. Palpanas, and C. Confalonieri, "Social listening of city scale events using the streaming linked data framework," in *ISWfC*, 2013.

[13] City pulse,http://www.ict-citypulse.eu.

[14] P. Paraskevopoulos, G. Pellegrini, and T. Palpanas, "When a tweet finds its place: Fine-grained tweet geolocalisation," in *International Workshop on Data Science for Social Good (SoGood), in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery (ECML PKDD)*, 2016.

[15] H. Abdelhaq, C. Sengstock, and M. Gertz, "Eventweet: Online localized event detection from twitter," *Proceedings of the VLDB Endowment*, vol. 6, no. 12, 2013.

[16] P. S. Earle, D. C. Bowden, and M. Guy, "Twitter earthquake detection: earthquake monitoring in a social world," *Annals of Geophysics*, vol. 54, no. 6, 2012.

[17] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.

[18] S. Hasan, X. Zhan, and S. V. Ukkusuri, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media," in *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*. ACM, 2013, p. 6.

[19] B. Hu and M. Ester, "Spatial topic modeling in online social media for location recommendation," in *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013, pp. 25–32.

[20] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare." *ICwSM*, vol. 11, pp. 70–573, 2011.

[21] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences*, vol. 107, no. 52, pp. 22 436–22 441, 2010.

[22] A. Olteanu, I. Weber, and D. Gatica-Perez, "Characterizing the demographics behind the# blacklivesmatter movement," in *2016 AAAI Spring Symposium Series*, 2016.

[23] T. M. T. Do, O. Dousse, M. Miettinen, and D. Gatica-Perez, "A probabilistic kernel method for human mobility prediction with smartphones," *Pervasive and Mobile Computing*, vol. 20, pp. 13–28, 2015.

[24] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[25] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "Nextplace: a spatio-temporal prediction framework for pervasive systems," in *International Conference on Pervasive Computing*. Springer, 2011, pp. 152–169.

[26] G. Chatzimilioudis, A. Konstantinidis, C. Laoudias, and D. Zeinalipour-Yazti, "Crowdsourcing with smartphones," *IEEE Internet Computing*, vol. 16, no. 5, pp. 36–44, 2012.

[27] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.