2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

# Identification of Key Locations based on Online Social Network Activity

Hariton Efstathiades, Demetris Antoniades, George Pallis and Marios D. Dikaiakos
Dept. of Computer Science, University of Cyprus
Nicosia, Cyprus
Email: [h.efstathiades, danton, gpallis, mdd]@cs.ucy.ac.cy

***Abstract*—Ubiquitous Internet connectivity enables users to update their Online Social Network profile from any location and at any point in time. These, often geo-tagged, data can be used to provide valuable information to closely located users, both in real time and in aggregated form. However, despite the fact that users publish geo-tagged information, only a small number implicitly reports their base location in their Online Social Network profile. In this paper we present a simple yet effective methodology for identifying a user's key locations, namely her home and work places. We evaluate our methodology with Twitter datasets collected from the country of Netherlands, city of London and Los Angeles county. Furthermore, we combine Twitter and LinkedIn information to construct a work location dataset and evaluate our methodology. Results show that our proposed methodology not only outperforms state-of-the-art methods by at least 30% in terms of accuracy, but also cuts the detection radius at least at half the distance from other methods.**

## I. Introduction

The massive adoption of mobile devices that offer Internet connectivity, geo-location capabilities, and continuous access to online social networking services (OSNs) has enabled users to contribute content to OSNs on a continuous basis, from different locations and at different times of the day. Based on this ubiquitous OSN activity, it is now possible to sketch the mobility trajectories of users and to pinpoint their visited locations. However, few users (34%) publicly reveal, in their OSN profiles, explicit and accurate information about their physical locations at a granularity that is higher than city level [8]. In recent years, the automatic mapping of users to their "key" visited locations of interest (e.g., home, work, leisure), based on their online social presence, has been of great interest for the research community [15, 17]. Information about the key locations that users visit and from which they contribute content to OSNs, has applications in a variety of research fields like understanding user movement [2]; investigating the relation among real-world human activities and interactions, physical spaces, and OSN structure and dynamics [1]; and exploring the challenges to user privacy protection. Moreover, the combined knowledge that can be mined from this information, can be of tremendous help for a diverse number of applications aiming at improving habitats and daily activities in cities, from event identification and recommendation to urban city planning.

In this paper we use geo-tagged Twitter activity traces to identify the key locations of users, namely their Home and Work places. Our method is based on two basic observations: First, users tend to spend a significant, but distinct, amount of their time during an average day in these two key locations. For example, in week-days, users tend to stay at home for most of the evening hours. Second, Home and Work locations are much more likely to appear in a user's geo-tagged OSN activity during these specific timeframes, than locations that are not embedded to the user's routine.

To evaluate our approach we focus on geo-located data from Twitter, one of the most popular OSNs. We collect more than 1 billion tweets, posted by more than 1 million users from three different geographical areas: the country of Netherlands, the city of London and Los Angeles county. We selected these areas as they exhibit a high Twitter usage [16] and belong to different nations, cultures, and time zones. Furthermore, we apply our method on a USA dataset used in recent literature by Yuan et al. [20]. Our results show that our method can accurately identify the key locations of users with precision values close to 80%, in post-code granularity. This result is not only an improvement of 30% over the state-of-the-art, but also offers key location identification in a much finer grain granularity over the 10Km and city level range identified in previous work [15, 17].

The contributions of this paper can be summarized as follows:

1) We present a simple but effective method for identifying key locations of a user based on geo-tagged Twitter data.
2) We present an extended evaluation of our method where we show that it can identify the user's Home location with an accuracy of more than 80%, giving a 30% improvement over the state-of-the-art.
3) We construct a work location identification dataset by using user reported information to both Twitter and LinkedIn OSNs.
4) We present an evaluation of user workplace identification with an accuracy of 63% at post-code level and more than 80% for a radius of 10Km. To the best of our knowledge this is the first study that constructs a dataset and performs analysis for workplace location identification.
5) We use the proposed method to perform a broader key location identification for all users in our dataset and compare that with socio-economic open data for the areas of interest. The comparison shows a clear mapping between our identified locations and the ground truth.

The remainder of this paper is structured as follows: We define our research question in Section II and present related work in Section III; In Section IV we describe the datasets used in the paper, while in Section V we perform a characterization of user activity in regards to her key locations. We present

ASONAM '15, August 25-28, 2015, Paris, France
© 2015 ACM. ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2808877

our method in Section VI and its evaluation in Section VII; Section VIII summarizes our findings and concludes our work.

## II. Problem formulation

Given as input the geo-tagged Twitter activity $T_u$ of a user $u$ we are interested in the identification of the user's key locations, namely Home and Work locations, denoted as $H_u$ and $W_u$ respectively. The tweet information we are interested in is represented by the vector $< p, t_p >$, where $p$ denotes the geographical coordinates ($< long, lat >$) the user tweeted from at time $t_p$. The set of all location visited by user $u$ can then be denoted as $P_u$. Our research then tries to give an answer to the question: *Can we identify a user's u home and work location simply by observing the locations and time the user tweeted from?* In the following sections we introduce a method to answer this research question providing the highest key location identification accuracy and also minimizing the detection radius granularity to as low as possible.

## III. Related Work

Identification of user key locations is of high interest for researchers in Online Social Networks analysis, who focus either on estimating these locations or enriching their datasets and using them for further analysis. In a large part of the literature, researchers are interested in identifying user home locations. Recent studies present approaches that are focused on estimating a user's key locations based on geo-tagged OSN activity or/and content that the user publishes in her profile. In this section we present studies for both and summarize common methodologies that have been identified.

**Geo-tagged based-** Georgiev et al. [6] aim to study users' geographic activity patterns using data retrieved from Four-square. In their study they aimed at estimating user home locations and investigate the influence in events participation. They assume that a user's home place is his most popular place as estimated by the number of Four-square check-ins. Jourdak et al. [11] investigate the influence of home location to user mobility patterns, also by marking the most frequently visited location as home place. This approach is probably the easiest in inferring home places, however it lacks on accuracy and granularity. A similar approach, proposed by Cho et al. [2], divides the geographical space in 25 by 25Km cells and define the home location of a user as the average position of him in the cell with the most check-ins. Hawelka et al. [7] investigate global mobility patterns with the use of Twitter. For that purpose they need to estimate users' residence country. They mark as country of residence the country where the user published most of her tweets. All the above use the average, median or most popular coordinate to estimate the location of the user, without considering her different daily habits. A simple drawback of this approach is that a regularly visited place, like a cafeteria or the cinema, would have a significant impact on the identified location. Our method takes into account the different hours of the day that the user will most likely reside in a key location, thus eliminating, to a high degree, the influence of user's hot spots.

Sadilek et al. [18] described an approach to infer the location of Twitter users for a given time period, using geo-tagger information of their ego network. Their approach showed the need for at least 2 geo-active friends in order to predict the user's location with an accuracy of up to 77%. To succeed this their approach needs at least 100 geo-tagged tweets for a one month period, from the user's friends.

**Content based-** Mahmud et al. [15] propose an approach that based on a location dictionary for places all over the United States, manages to infer 57% of users Home Location using their Tweets at city-level granularity. They present a hierarchical classification approach which narrows down the granularity from timezone, state or region and then city. Ryoo and Moon [17] propose a content-based approach that aims in identifying Twitter users' home place in a granularity of 10Km, using the tweet textual contexts. They use a probabilistic model to assign location data to popular words in Twitter and then use the popularity of these words to identify the location of the users that tweet them. Their approach manages to identify up to 57% of the users in a 10Km radius. Li et al. [14] present an effective location identification approach based on information collected from multiple microblogs, combined and utilized in order to identify the top-k candidate locations of a user. They show an accuracy similar to that of Ryoo and Moon. All the above approaches use the aforementioned geo-tagged methods as ground truth for the users' home location. As mentioned before these approaches lack in terms of accuracy and thus are not able to provide valid ground truth information. Our approach significantly improves the accuracy of those techniques, thus can be additionally used to provide the ground truth information needed from context-based approaches.

## IV. Dataset

We used a variety of OSNs to collect geo-location information about the users. Regarding workplace location, we introduce a novel method, that combines a variety of OSNs, and dataset.

### A. Home Location

For home location identification we turn to Twitter and search for users that include geographical information in their tweets. To avoid extensive crawling of the Twitter network we first visit Twitter's live stream for three different geographical areas, namely, the country of Netherlands (March 2014), the city of London, UK and LA county, CA, USA (November 2014). We use the geographical boundaries of these areas and collect geo-tagged tweets within these boundaries. For each of these tweets we collect public information about the user that posted the tweet. This information includes the past tweeting activity of the user, her ego network, followers and followees, and her profile information. To expand our dataset we use the users collected from this process as *seeders*. For each seeder we randomly crawl users belonging to her ego network and collect the same information. We keep only users that have at least one geo-tagged tweet from the three areas of interest, and add them to the seeders list for further crawling.

**Data cleansing:** One major concern for any Twitter dataset is to avoid bots, which act differently than most regular Twitter users, biasing the analysis. The nature of our analysis also requires to focus on individual users, removing from our dataset Twitter accounts that are linked with company or professional profiles. These accounts are mainly used to

| Name | Location | Users | Tweets | Geo-tagged Tweets |
|---|---|---|---|---|
| TW-NL | Netherlands | 702,593 | 668,684,891 | 16,445,151 |
| TW-LA | LA County | 350,637 | 532,738,302 | 35,645,531 |
| TW-LO | London | 182,272 | 232,331,077 | 35,406,092 |

TABLE I. HOME LOCATION DATASET: NUMBER OF USERS, NUMBER OF TWEETS AND GEO-TAGGED TWEETS, FOR EACH OF 3 REGIONS OF THE RESULTED DATASET.

| Name | Post-code areas | Average area radius ($Km$) | Ground Truth Users |
|---|---|---|---|
| TW-NL | 286 | 2,68 | 1414 |
| TW-LA | 62 | 2,75 | 370 |
| TW-LO | 151 | 2,37 | 760 |

TABLE II. HOME LOCATION DATASET: NUMBER OF POST-CODE AREAS AND AVERAGE AREA RADIUS IN $Km$, FOR EACH OF 3 REGIONS OF THE RESULTED DATASET.

advertise their owner and are clearly differentiated from Twitter accounts used by "regular" users [5, 19]. Filtering individuals from a list of Twitter profiles is an open research problem that we aim to target in our future work. For the purpose of this work we randomly sampled 1,000 users, from our dataset, and manually marked the individual users. For this sample we evaluated a number of different profile features to identify the distinguishing factors for individual users. These features included the number of friends and followers, number and frequency of tweets etc. Our analysis showed that the cardinality of the intersection between the sets of followers and friends of a user is a satisfactory distinguishing factor for identifying individual users. Reciprocal relationships are also used to identify close friends [10], which is a characteristic of individual users. Based on this result we use this feature and remove all "corporate" and bot accounts from our dataset.

**Collected data:** Table I summarizes the collected data for each geographical area after data cleansing is performed. Overall we retrieved information for more than 1 million Twitter users. This information contains around 1.5 billion Tweets, 6% of which contain geographical information. This number is significantly larger than most related work [12].

**Ground truth dataset:** We used public information contained in Twitter user profile, manually inserted by the users, in order to create a ground truth dataset for evaluating our approach. To this end, we search the profile information location field for exact geographical coordinates or user-reported post-code information. Then, we use either of these values to map the user to a post-code, considering that to be the user's home location. Table II details the number of users contained in our ground truth dataset, for each area of interest. The table also lists the number of unique post-codes for which we have users and the average geographical area covered by each post-code. The latter value also constitutes the average granularity in which we can actually locate a user's key locations.

**Previous work dataset:** To further strengthen the evaluation of our method and compare against state-of-the-art approaches we apply our methodology to the dataset retrieved by [3] and used by Yuan et al. [20]. This dataset includes geo-tagged micro-blogging activity and home location ground truth for USA 9,475 users. We refer to this dataset as *GeoText*. [1]

| Name | Users | Tweets | Geo-tagged Tweets |
|---|---|---|---|
| TW-LinkedIn-Work | 317 | 915,933 | 73,003 |

TABLE III. WORKPLACE LOCATION DATASET: NUMBER OF USERS, NUMBER OF TWEETS AND GEO-TAGGED TWEETS.

### B. Workplace Location

In contrast to home location, work location is not usually clearly stated by a Twitter user in her personal profile. The reason for this is that Twitter profiles are used for a completely different purpose than career-related tools. LinkedIn on the other hand, is a professional social network where users publish career related information, including (among others), their current location and place of work.

To construct a work location dataset we use FriendFeed, an online OSN profile aggregator tool. FriendFeed allows its users to aggregate information posted into multiple OSNs by adding their profile accounts to a central service. For our dataset we collect FriendFeed accounts, whose owner have added both their Twitter and LinkedIn profiles, from FriendFeed's public stream during January 2015. We then used Twitter and LinkedIn APIs to retrieve the public profile information of the collected users, concluding to a list of 3,285 users. For these profiles we were able to collect both the geo-tagged activity of the user (Twitter) and the user's work location (LinkedIn). To the best of our knowledge, this is the first study that builds a dataset for user work location identification.

**Data Cleansing:** Despite the fact that the majority of LinkedIn profiles include information about a user's current employer, details regarding the exact geographical location of a company is limited. Additionally, when such geographical information is available, usually is related to the company's global headquarters and not the exact branch where users work at. For that reason we performed a pre-processing analysis in order to identify the exact branch of the company where a user works, along with its (self-stated) location at post-code level. As a first step we used users location field from her LinkedIn profile, that provides information about users' locations at city level. We then aimed to find the companies with the same name, as the one in the user's current employment field, in the area close to users reported location. If the location is not identified we discard the user profile from our analysis set. Users who do not include information about their employer were also discarded. Following this approach we managed to identify geo-location information for the workplace of 317 different users from different countries and map them to their corresponding post-code area.

**Collected data:** Tables III and IV summarize the data collected for inferring users workplace locations. Our sample is multi-cultural as it contains users from a variety of countries of origin who are working in different industries.

## V. USERS KEY LOCATIONS

Most previous work in user location identification from Twitter ignores two important observations that actually characterize users daily routine, not only in their online activity

[1] Geo-tagged Microblog Corpus: http://www.ark.cs.cmu.edu/GeoText/

| | | Percentage |
|---|---|---|
| **Country of origin** | United States | 34.7 |
| | Great Britain | 11.3 |
| | Italy | 5.7 |
| | Spain | 5.1 |
| | Canada, France, Turkey | 4.7 (each) |
| | Other(23) | 29.1 |
| **Industry** | Internet | 21.8 |
| | Information Technology | 16.4 |
| | Marketing and Advertising | 11.7 |
| | Computer Software | 8.2 |
| | Online Media | 7.6 |
| | Other(51) | 34.3 |

TABLE IV. WORKPLACE LOCATION DATASET: DEMOGRAPHIC CHARACTERISATION

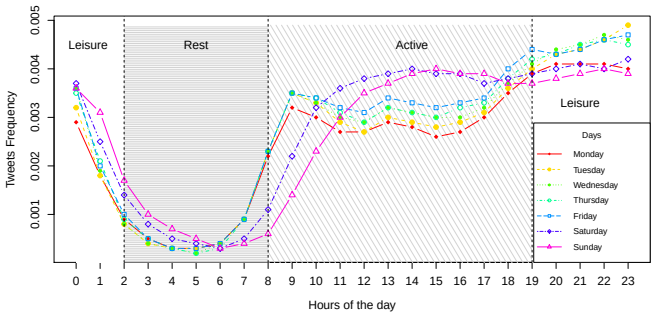



Fig. 1. Tweets publishing activity during a week. Based on differences in behavior, day is divided in different *time-frames*. Rate is calculated divided by the total tweets quantity of the whole week.

but also in their real life habits. These are: $(i)$ users tend to spend a significant, but distinct, amount of their time during an average day in two key locations namely their Home and Work; $(ii)$ these two locations are much more likely to appear in the user's geo-tagged activity during these specific timeframes, than locations that are not so frequent in user routine.

These observations are intuitive for users when considering our real life interactions. Since we are interested in key location identification we use the ground truth dataset, described in the previous section, to evaluate whether these observations are also present in users' Twitter life. Figure 1 plots the percentage of Twitter activity (y-axis) for the different days of the week (lines) and the different time of each day (x-axis). We can clearly see the diurnal pattern in tweeting activity. Early morning hours show less activity than hours in the morning-afternoon and evening hours. Additional, we can observe the points in which user behavior seems to change, i.e. around 2 AM and 7-9 AM. [2] Furthermore, we can observe a slight shift in the tweeting activity of the users during weekends, as compared to weekdays. This shift denotes differences in the behavior of the user during weekends, an observation also made by Herder et al. [9], when analyzing user trajectories. Due to this observation we decide to ignore weekend activity when searching for the user's home and work locations. We include this activity at a later state when we want to analyze the Leisure locations a user visits.

[2]Similar behavior has also been observed by [4]

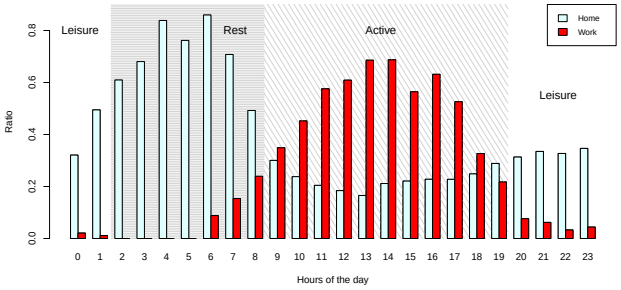



Fig. 2. Ratio of tweets published from user's reported Home and Work locations on an hourly basis. Y-axis represents the portion of total geo-tagged Tweets that have been produced during the specific hour.

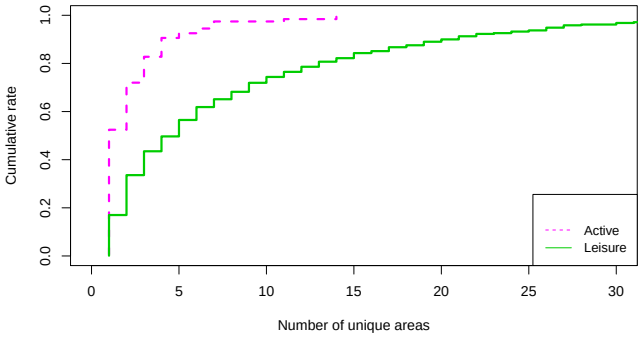



Fig. 3. Number of different locations from which a user tweets during *Active* and *Leisure* hours.

Based on the above observations, we argue that user's activity can be split in three different time frames related to the place that the user might be during that period. These timeframes are: $(i)$ *Rest* time, between 2 and 8 AM, the time that the user most likely resides at her *Home* location $(ii)$ *Active* time, during 8 AM and 7 PM, denoting the time that the user will most likely be at *Work* and $(iii)$ *Leisure* time during the rest of the day, where the user spends her free time most probably outside the home and work environment.

We expect that a user will mostly be posting tweets from a single location during the *Rest* and *Active* timeframes. Figure 2 examines this hypothesis for the *Home*(cyan) and *Work*(red) key locations. Using our ground truth dataset, for each case, we plot the ratio of user tweets sent from her reported home/work location during different hours of the day. The ratio is calculated as the fraction of tweets user $u$ posted at each specific hour during the day from her home/work location over the total number of tweets of the user for that hour. As we can see from the results, the probability tends to increase significantly during (and close to) the *Rest* timeframe for the Home location, and during the *Active* timeframe for the Work location. Our observations also agree with the results of an analysis performed on a single user from Yuan et al. [20].

Figure 3 examines the number of different locations the user tweets from during *Active* and *Leisure* timeframes. We excluded the user's reported *Home* location from this analysis.

| Dataset | *Rest* | *Leisure* |
|---|---|---|
| TW-NL | 0.744 | 0.362 |
| TW-LA | 0.735 | 0.357 |
| TW-LO | 0.737 | 0.354 |

TABLE V. PROBABILITY OF *tweeting from Home* DURING *Rest* AND *Leisure* TIMEFRAMES FOR THE 3 DIFFERENT DATASETS.

We observe that in 90% of the cases the user will post, at max, from a handful of locations during *Active* timeframe. Having in mind that the user spends most of this time at her workplace, we expect it to be the most popular of these locations. The Figure also plot the CDF of different locations a user tweets from during the *Leisure* timeframe. The number of different locations is significantly higher in this case. Around 50% of the users tweet from more than 10 unique locations during this timeframe. This observation clearly demonstrates the different habits the users have in the different timeframes.

**Discussion**. Our analysis shows that the majority of users demonstrate temporal activity patterns on Twitter highly related with their home and work locations. By analyzing the geo-tagged information we can conclude that tweeting activity during *Rest* timeframe is more likely to be generated from *Home* location. During *Active* timeframe activity is mostly likely to be generated from *Work* location. This result clearly indicates that actual information about a user's key locations can be inferred from Twitter activity.

## VI. KEY LOCATION IDENTIFICATION MODEL

The above observations verify our hypothesis that the user is much more likely to tweet from her *Home* location during *Rest* hours and from her *Work* location during *Active* hours. Based on these remarks, we define our key location identification method as follows: Given a set of geo-tagged tweets $T_u$ of user $u$ and the place $P_u$ the tweets where posted from, we first split this set into three subsets, $R_u$, $A_u$ and $L_u$ containing the tweets during *Rest*,*Active* and *Leisure* timeframes respectively. We then estimate the *Home* and *Work* locations of the user by finding the most "popular" location during "non-working" ($R_u$ and $L_u$) and "working" ($A_u$) hours, respectively. The popularity in each case is calculated as follows

$$W_u = \arg\max(\forall p \in P_u | t_p \in A_u : \sum_{i=day_1}^{day_n} A_u(i,p)) \quad (1)$$

$$H_u = \arg\max(\forall p \in P_u | t_p \in (R_u \cup L_u) : \sum_{i=day_1}^{day_n} w_r \times R_u(i,p) + w_l \times L_u(i,p) \quad (2)$$

Equation 1 calculates the most popular place, in number of unique days, among all the places the user tweeted during the *Active* timeframe, $A_u$. Equation 2 calculates the most popular place, also in number of unique days, among all the places the user tweeted during both the $R_u$ and $L_u$ timeframes. According to Figure 2 users tweet from Home with higher probability during $R_u$. To take this observation into account we apply a different weight $w_r$ to the popularity of a place $p$ if the tweet is included in $R_u$, and $w_l$ if the tweet is included in $L_u$.

We calculate the weights by estimating the average, amongst all users, fraction of tweets from the home location over the total number of tweets during the two different timeframes. Table V shows the weight values for our three home location ground truth datasets. We can observe that the weights for all areas are almost identical. This shows that our method can easily be adapted to any area of interest without changing the weights. We use the average of all three in the evaluation of our method.

## VII. EVALUATION

We evaluate our key location identification method, proposed in the previous section, at post-code granularity both for Home and Workplace locations. For the Home location case, we evaluate our method using two different approaches. First, we compare the identified user Home locations with the user reported home location, as extracted from the user's profile entry. Second, we compare our results with publicly available socio-economic data. We compare the post-code population density in Home locations, with the ones that we derive by applying our method in our Twitter dataset. We compare the estimated workplace locations against exact workplace locations identified both from LinkenIn and Twitter data.

### Metrics and Methods

We validate our approach based on well established metrics used in literature. These are:

- ACC *Accuracy* gives the percentage of correctly inferred users' key locations over the total sample size [13, 15, 17].
- ACC@R *Accuracy within radius* (R) gives the percentage of correctly inferred users' key locations identified within R Km from users reported locations [13, 15, 17].
- AED *Average Error Distance* defines the distance, in Km, between the inferred location (center of the post-code in our case) and user's reported location [13, 17].

Using the above metrics we evaluate our method and compare it with the state-of-the-art geo-tagged data user location methods as those are defined in related work. These are:

- MP *Most Popular* marks as home location the most popular location, in number of geo-tagged tweets, visited by the user [6].
- MC *Median Clustering* marks the user's home location by calculating the median value of location the user tweeted from [17].
- TF-C *TimeFrame - Clustering* is the method proposed in this paper. The method takes into account the fact that the user usually resides in different locations during different times of the day and week.

### A. Home Location identification

### Data pre-processing

Before applying our method to either dataset we first do a pre-processing pass over the data, to eliminate common well known locations and bring all geo-tagged information to a

| Method | TW-NL | TW-LO | TW-LA |
|---|---|---|---|
| | ACC | | |
| MP | 0.69 | 0.47 | 0.55 |
| MC | 0.67 | 0.19 | 0.39 |
| **TF-C** | **0.81** | **0.68** | **0.701** |
| | AED | | |
| MP | 3.21 | 4.13 | 6.05 |
| MC | 3.93 | 5.21 | 8.15 |
| **TF-C** | **2.77** | **2.05** | **2.63** |

TABLE VI. HOME-LOCATION IDENTIFICATION PERFORMANCE MEASURED IN ACCURACY(ACC) AND AVERAGE ERROR DISTANCE (AED) IN KM, FOR 3 DIFFERENT APPROACHES IN 3 DIFFERENT AREAS.

common format at post-code granularity. Popular locations are referred in Twitter as *Points Of Interest* (POI). These locations define specific attractions, local businesses, landmarks etc. POIs are not used to define a user's home place, and for this reason we decide to remove such places, marked with a specific tag in the tweet location field, from the user's Twitter stream.

In a second step we map geographical coordinates contained in the tweet location field to the closest post-code in terms of euclidean distance. We choose postcode level over other forms of mapping, i.e. city or arbitrary geographical boundaries [3], since it is a well defined and official boundary on one hand and much more precise on the other.

### Evaluation with ground-truth data

**Results**. Table VI presents the evaluation of TF-C in correctly identifying the *Home* location of the user, for the three different geographical locations, along with the comparison with the aforementioned state-of-the-art methods. Overall TF-C outperforms the other methods, in both metrics presented in the table. In terms of accuracy TF-C can identify more than 80% of the user's home locations, in the country of Netherlands, while in any case it can identify more than 70% of the user's home. In comparison with the other methods, TF-C performs 20-50% more accurate.

In terms of the AED metric we can see, from Table VI, that TF-C locates the user closest to her Home location, with values always being less than 2.7Km from the center of the user defined post-code. Recall, from Table II, that the average area radius for the post-codes in our dataset is also around our method's AED values. All other methods identify the user at least 3.2Km from her defined location, and in some cases reach error distances close to 8Km.

Figure 4 compares the evaluated approaches in terms of the ACC@R metric for the TW-LO dataset. The figure plots the total accuracy of each method as a function of the distance from the center of the user defined postcode. From the results we observe that *TF-C* can identify more than 95% of the users in less than 10 Km from their center location, and more than 80% in less than 5Km. The *MP* and *MC* methods reach the same level of accuracy (80%) for radius larger than 10 and 15 Km respectively. Also, *TF-C* can identify all users in less than 20 Km, versus the 30+ Km of the two comparison methods.

Figure 5 examines the number of tweets needed, by our method, to accurately identify the user's Home location. As we can see from the figure, 10 to 20 tweets are enough for

[3]Cho et al.[2] used a 25Km square boundary

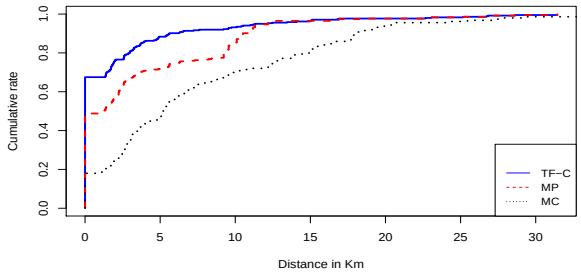



Fig. 4. TF-C performance for London dataset. Proposed methodology is able to identify the exact post-code location with 68% accuracy and performs better in lower granularities than compared approaches.

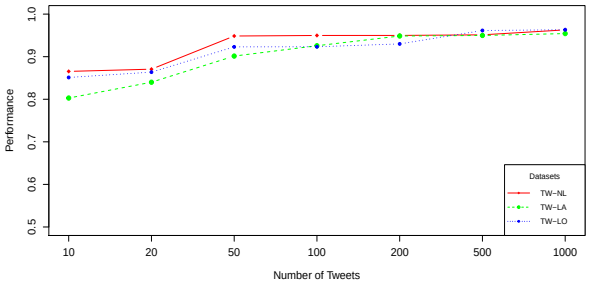



Fig. 5. Performance of proposed method in contrast to the number of recent tweets for the 3 datasets.

TF-C to identify more than 85% of all identified users. Recall, that TF-C is able to accurately infer a user's key locations when her tweeting activity follows a distribution similar to the ones presented in figure 2. With the above numbers in mind it is clear that our approach can provide high accuracy both for new and old twitter users, using only a small amount of their tweet activity.

**Discussion**. Results show that TF-C outperforms the state-of-the-art in geo-tagged data based key location identification methods by at least 15% and up to 50% in terms of accuracy. Also our method can detect user's home location in a radius smaller than 10Km in most of the cases. *MP* and *MC* are both methods used to provide ground truth data for social community based [18] and content-based [15, 17]. All these methods result in low detection accuracy, between 20 and 70%, and also detect users in a much higher radius, more than 10Km in all cases. Our results show that TF-C provides a more accurate ground truth for user's home location, that will help improve both the methods themselves and their detection accuracy. In future work we plan to both evaluate our approach against such methods and quantify the improvements a better ground truth dataset can provide.

### Evaluation over previous work dataset

We also evaluate our approach over the *GeoText* dataset, collected and used in previous work related to user location identification. Home location of each user in this dataset is already provided by Eisenstein et al. [3]. Our evaluation results show TF-C identifies the home location of the users in this dataset with an accuracy of 76%. Yuan et al. [20]

also used the above dataset for evaluating a user location identification method based on the tweet text. Their approach uses training and prediction of the user location and gives prediction accuracy significantly lower than TF-C.

**Comparison with open-data**

**Results**. In the previous section we evaluated the accuracy of our method and demonstrated the improvement it offers over the related work. In this section we use open data from the County of Los Angeles to derive the population of each different post-code as a function of the total population of the County. Figure 6(a) shows a heat-map of the differences in the population distribution derived from the real data compared with the population distribution as this can be derived by our method, for 200,000 Twitter users. As depicted in the heatmap, for about 87% of the areas the predicted and real post-code population rate differ only by 0.005.

**Discussion**. Nowadays, the population census procedure is performed with the use of well studied and applied methodologies, like door-to-door interviews at a sample of habitants. Despite the fact that these enumeration methodologies provide us with accurate data, they do have several limitations. [4] Such limitations are the cost of performing such a study, the time needed for its completion and the access to the sample that will be used. Thus, such demographic studies take place on a 'several years' base and usually are out-dated. Based on the accuracy provided by our methods, we believe that *TF-C* can act as a complementary and closer to real-time method for performing demographic studies. Using data available from OSNs one can quickly and in zero cost get a close to real estimate of the current trends in an area of interest, without waiting for the more complicated population census procedure.

### *B. Identifying workplace location*

In this section we proceed and evaluate our approach' accuracy in predicting a user's *workplace location* based on her interactions in Twitter. To the best of our knowledge, this is the first study where geo-location information about workplaces has been collected and used for such an analysis.

**Data pre-processing**

We use the LinkedIn-Twitter dataset described in section IV-B for this evaluation. Contrary to the home location evaluation case, we do not remove popular locations, referred by Twitter as Points Of Interest (POI), from the workplace evaluation. These attractions or local businesses were removed from the previous analysis as they are not used to define a user's home. However they could represent a user's workplace.

Similarly with home location identification, we map geographical coordinates contained in tweet location field to the closest post-code area. However, because we use a world-wide dataset and we do not have access to global post-code information, we divide the global geographical space in boundaries with radius equal to 2Km, which is less than the average post-code coverage size in Netherlands, London and LA county. We then map each tweet to the corresponding boundary area.

[4]https://www.census.gov/prod/1/gen/95statab/app3.pdf

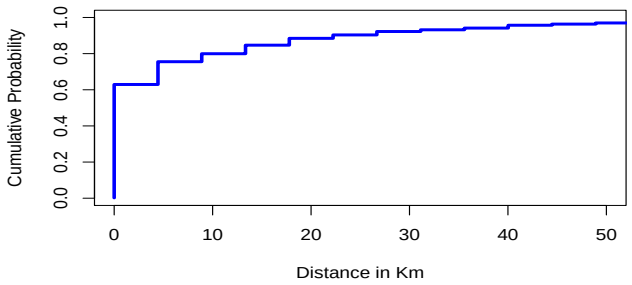



Fig. 7. TF-C performance for identifying workplace location from a global dataset. Proposed methodology is able to identify the exact workplace location at post-code granularity with 63% accuracy.

**Evaluation with ground-truth data**

**Results**. Figure 7 presents the evaluation results in terms of the ACC@R metric for the global workplace dataset. As we can see, our method is able to detect a user's workplace location with similar performance as her home location. Specifically, it is able to detect the exact post-code location with an accuracy of 63%. Additionally, in a 10Km radius, our method, is able to identify the employers location for more than 80% of the total sample.

**Discussion**. Our results demonstrate that *TF-C* achieves high accuracy in workplace location identification, on a worldwide dataset, at a granularity equal to a post-code area. From these results we can see that information about a user's workplace area can be derived from public data, despite the fact that she does not explicitly reports it. In this work we take into account only the meta-data of users activity in Twitter, taking advantage of the fact that our interactions in Online Social Networking platforms sometimes generate more information than the one we intend to share.

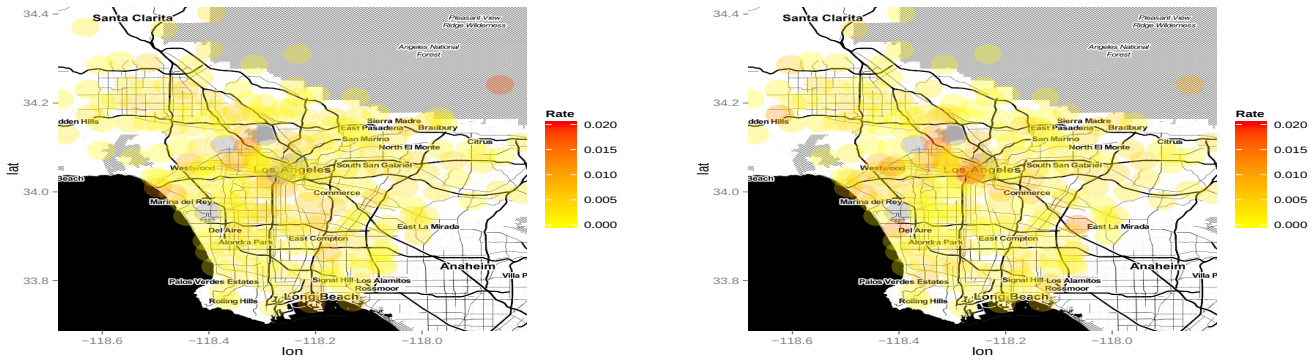
**Comparison with open-data**

**Results**. After identifying the workplace location at post-code granularity of a sample of users in Los Angeles county, we proceeded in comparing the general statistics with open-data collected from this area. Figure 6(b) presents the differences in the rates between real and predicted employees fraction over total employees of each post-code area. As we can see more than 85%, of post-code areas differ by less than 0.005, while only 5% differs by more than 0.01.

**Discussion**. As we can see with the comparison against open data, TF-C is able to provide insights to real-world studies that are more complex than population census. Methodologies that are being applied in such studies are well validated and commonly accepted, however, the identification of users key locations from their online social networking activity can also help in this effort.

## VIII. Conclusions

We presented a simple but effective methodology for the identification of a Twitter user key location. Our methodology uses geo-tagged Twitter data, and based on two main observations regarding user's real life habits manages to identify




(a) Differences between real and predicted population rate.

(b) Differences between real and predicted employees rate.

Fig. 6. Predicted population was calculated after applying the proposed model on a dataset of 350,000 users from LA county. Real population was collected from LA county's official statistics.

the *Home* and *Work* location of the users. Evaluation of our method, using data from several geographical regions, showed that it outperforms previous methods by more than 30%. Additionally, it can identify the user's key locations at post-code granularity, that is in a radius smaller than 3Km. Comparison with socio-economic open data showed that our method can correctly identify the populated areas of the geographical region of interest.

To further evaluate our proposed methodology we illustrate how one can combine information from multiple social networks, namely LinkedIn and Twitter, in order to construct a dataset that includes both the user's work location and her tweet activity. Using this dataset we evaluated our method for work location identification. Our results show an accuracy close to 80% for identification of user location in a 10Km proximity. To the best of our knowledge this is the first attempt to construct a workplace ground truth dataset and also the first workplace identification method.

Our future work plans to use the identification derived from the methodology described in this paper to derive insights for the users daily activities, how the locations visited by the user affect her social network connections, and how the user transports derived by Twitter data can be used to support city planning procedures.

## IX. Acknowledgments

This work was partially supported by the iSocial EU Marie Curie ITN project (FP7-PEOPLE-2012-ITN).

## References

[1] C. Brown, A. Noulas, C. Mascolo, and V. Blondel. A place-focused model for social networks in cities. In *Social Computing (SocialCom), 2013 International Conference on*, pages 75–80, Sept 2013.

[2] 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*, KDD '11, New York, NY, USA, 2011. ACM.

[3] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, Stroudsburg, PA, USA, 2010.

[4] D. Falcone, C. Mascolo, C. Comito, D. Talia, and J. Crowcroft. What is this place? inferring place categories through user patterns identification in geo-tagged tweets. In *Proceedings of International Conference on Mobile Computing, Applications and Services*, MobiCASE 2014.

[5] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *CoRR*, abs/1407.5225, 2014.

[6] P. Georgiev, A. Noulas, and C. Mascolo. The call of the crowd: Event participation in location-based social services. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 2014.

[7] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.

[8] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA, 2011. ACM.

[9] E. Herder, P. Siehndel, and R. Kawase. Predicting user locations and trajectories. In *User Modeling, Adaptation, and Personalization*, pages 86–97. Springer, 2014.

[10] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back?: Reciprocal relationship prediction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1137–1146, New York, NY, USA, 2011. ACM.

[11] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. Understanding Human Mobility from Twitter. *ArXiv e-prints*, Dec. 2014.

[12] D. Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships, 2013.

[13] S. Katragadda, M. Jin, and V. Raghavan. An unsupervised approach to identify location based on the content of user's tweet history. In *Active Media Technology*, volume 8610 of *Lecture Notes in Computer Science*, pages 311–323. Springer International Publishing, 2014.

[14] G. Li, J. Hu, J. Feng, and K.-L. Tan. Effective location identification from microblogs. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 880–891, March 2014.

[15] J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *ACM Trans. Intell. Syst. Technol.*, 5(3), July 2014.

[16] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981, 2013.

[17] K. Ryoo and S. Moon. Inferring twitter user locations with 10 km accuracy. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 643–648, 2014.

[18] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 723–732, New York, NY, USA, 2012. ACM.

[19] C. Yang, R. Harkreader, and G. Gu. Empirical evaluation and new design for fighting evolving twitter spammers. *Information Forensics and Security, IEEE Transactions on*, 8(8):1280–1293, Aug 2013.

[20] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: Discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, New York, NY, USA.