Spatio-Temporal Multiple Geo-Location Identification on Twitter

Kambiz Ghoorchian
School of Electrical Engineering
and Computer Science (EECS)
KTH Royal Institute of Technology
Stockholm, Sweden
ghoorian@kth.se

Sarunas Girdzijauskas
School of Electrical Engineering
and Computer Science (EECS)
KTH Royal Institute of Technology
Stockholm, Sweden
sarunasg@kth.se

***Abstract*—Twitter Geo-tags that indicate the exact location of messages have many applications from localized opinion mining during elections to efficient traffic management in critical situations. However, less than 6% of Tweets are Geo-tagged, which limits the implementation of those applications. There are two groups of solutions: content and network-based. The first group uses location indicative factors like URLs and topics, extracted from the content of tweets, to infer Geo-location for non geo-active users, whereas the second group benefits from friendship ties in the underlying social network graph. Friendship ties are better predictors compared to content information because they are less noisy and often follow the natural human spatial movement patterns. However, their prediction's accuracy is still limited because they ignore the temporal aspects of human behavior and always assume a single location per user. This research aims to extend the current network-based approaches by taking users' temporal dimension into account. We assume multiple locations per user during different time-slots and hypothesize that location predictability varies depending on the time and the properties of the social membership group. Thus, we propose a hierarchical solution to apply temporal categorizations on top of social network partitioning for multiple location prediction for users in Online Social Networks (OSNs) like Twitter. Given a large-scale Twitter dataset, we show that users' location predictability exhibits different behavior in different time-slots and different social groups. We find that there are specific conditions where users are more predictable in terms of Geo-location. Our solution outperforms the state-of-the-art by improving the prediction accuracy by 16.6% in terms of Median Error Distance (MED) over the same recall.**

***Index Terms*—Geo-Location Identification; Graph Partitioning; Social Network Analysis; Spatio-Temporal Analysis.**

I. Introduction

GPS-Tagging is an invaluable functionality, recently added to many Online Social Network (OSN) platforms like Twitter and Facebook. It allows users to instantly share their exact geographical location information in the form of Latitude and Longitude. Numerous services, known as Location Based Services (LBSs), have been recently developed that benefit from this information for various applications like targeted advertisement [1], traffic control, and disaster management [2]. Twitter added Geo-Tagging to their services in 2010. However, studies show that only 6% of tweets are Geo-tagged [3]. Thus, the research question is ***"How can we infer Geo-location of a user in Twitter using her publicly available information."***

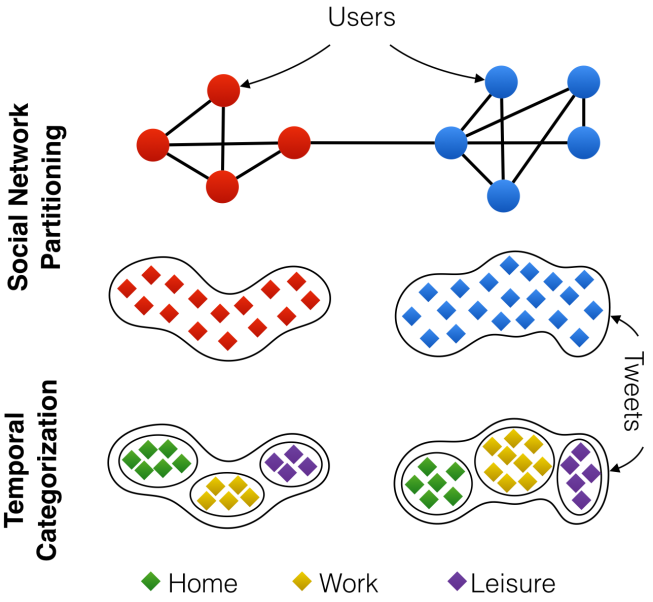


Fig. 1: A hierarchical algorithm for multiple Geo-location identification on Twitter constructed from two layers: *Social Network Partitioning* and *Temporal Categorization*. The algorithm first, partitions users and their corresponding tweets into multiple groups based on the topological structure of their social graph to account for users friendship locality. Then, it categorizes the tweets in each group into multiple sub-groups depending on the time-stamp of the tweet (e.g., home 0-7, work 8-18 and leisure 19-23) to consider the temporal dynamics of their behavior.

Current solutions either use location indicative factors like toponyms (location names), URLs and time-stamps, extracted from the *content* of tweets, or leverage the underlying *social network* graph to design their Geo-location prediction model. The studies show that social network-based approaches outperform content-based solutions in their prediction results.

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

Social network-based approaches often rely on two main assumptions: (i) each user has a single location, called *home location*, that is static and frequently visited; (ii) the home location of friends are close to each other. For example, Compton et al. [4] calculated a single location for each geo-active user as the geometric median of all the locations of her tweets and proposed a method to infer Geo-location for non geo-active users using the extracted home locations of their geo-active friends.

One important factor missing in those assumptions, which limits their prediction granularity, is the fact about human mobility in time. More specifically, users tend to visit multiple locations (e.g., home and work) frequently through their daily life, as suggested by [5]. Therefore, the frequent locations extracted for geo-active users are not literally always their home locations. They are, in fact, a combination of home, work, and other frequently visited locations, which are not necessarily close to each other. Thus, predicting a user's home location by looking at a mixture of non-relevant locations from her friends can cause dissemination of noise in the prediction results. Following this intuition, we hypothesize that the prediction can achieve finer granularity if the temporal aspects and the group membership characteristics of the users are taken into account.

Our approach. We propose a solution to overcome this limitation by considering multiple Geo-locations per user. Thus, we model a user's mobility pattern by taking the temporal aspect of her tweets into account. In particular, we distinguish between different days of the week and different hours in a day by presenting the concept of *Time-Slot* and assume that during a time-slot a user and her friends are in a geographically bounded location and stay there for an interval of time. These assumptions are based on the following intuitions:

- Users tend to appear and stay for a short period of time in a handful set of locations **frequently** and **periodically** in time [5].
- Strong ties (reciprocal relations) [6] are influenced by proximity [7].

The solution was developed in a hierarchical structure constructed from two layers: Social Network (SN) and Time (TI) as shown in Figure 1. The first layer, SN, uses a graph partitioning algorithm to partition users and their corresponding tweets based on the topological structure of the underlying social graph. This layer extracts the highest prediction granularity relative to the expansion properties of the social graph. The second layer, TI, divides the tweets in each partition into multiple sub-groups depending on the time-stamp (e.g., day and time) of the tweets.

To find the best features for temporal categorization, we analyzed the temporal distribution of tweets and realized that the temporal behavior of users follows a specific pattern depending on days of the week and hours of the day. Based on that, we considered three different time-slots, *Home (0-7)*, *Work (8-18)*, and *Leisure (19-23)*, during two types of days, *Weekdays* (from Monday to Friday) and *Weekends* (Saturday and Sunday). Then, we applied our solution over these temporal classes to identify the groups where users are most predictable with respect to their geographical location.

Contribution. The main contribution of this paper is to propose a hierarchical solution for Geo-localization of users on Twitter by combining the properties of their social network graph and the temporal aspects of their tweets.

We run multiple experiments on a large-scale Twitter dataset to show the efficiency of our approach. The dataset, collected between 2010 and 2014, contains 9.8 million tweets. The results show that users' geo-locations are most predictable during working hours on weekdays compared to other time-slots (e.g., weekends or home hours during weekdays). Our approach achieves up to 16.6% improvement in terms of MED compared to the stat-of-the-art [4]. In addition, we show that by limiting the prediction to larger social network partitions MED can be reduced by 42% on around 33% of the users.

II. Related Work

Current solutions for location identification on Twitter can be divided into two groups (i) Content-based, and (ii) Network-based.

Content-based Solutions. focus on the content of the tweets and other location indicative factors (like IP address, URLs, toponyms, etc.). Their main assumption is: ***Lexical structure of the text is influenced by geographic location [8]***. Wang et al. [9], Eisenstein et al. [10], Cheng et al. [3], and Hong et al. [11] proposed different probabilistic models base on various combinations of topic and location as latent variables and infer location using variational inference methods like *Expectation Maximization (EM)* [12]. Eisenstein et al. [10] reported $494km$ prediction results in terms of MED on a Twitter dataset over the entire US. Hong et al. [11] extended their approach by considering users' topical tendency. They assumed multiple locations per user and developed a probabilistic model based on that. They improved the previous results to an Average Error Distance (AED) of $120km$ on an open-source dataset, called CMU [10]. These solutions consider location as a continuous variable in their prediction models. Therefore, they exhibit scalability issues when it comes to larger geographic areas like country or continent level.

Next group of content-based solutions, known as semi-supervised approaches, tried to overcome this problem by discretizing prediction from exact geographic locations (e.g., Latitude and Longitude) to geographic-regions (also known as Geo-scope). The solutions in this group proposed models to map text into discrete locations like geodesic grids [13] (uniformly distributed isometric blocks on the surface of the earth [14]) or a list of cities [3], or publicly known locations (e.g., restaurants, tourist attractions, etc.) [15] and [16]. The results, even though exhibit notable improvements over the previous group, e.g.: 51% recall within $160km$ by [3] or 83% recall within $40km$ by [15], are still too far from the true geographic location.

Network-based Solutions. proposed to overcome limitations in content-based approaches leveraging the information in the underlying social network graph. The main assumption supporting these approaches is: ***A user's social network structure is influenced by locality***. based on that, the solutions in this group proposed models to infer users' location from the locations of their geo-active friends in the underlying social network graph.

Earlier approaches used probabilistic modeling and classification in their prediction models. Li et al. [17] developed a probabilistic model based on *Latent Dirichlet Allocation (LDA)* [18] to combine the *"following network"* (network of the followers) with topic modeling. Their system was able to locate 54% of users within $31.5km$ from their true geographical location. McGee et al. [6] investigated the structure of the social network to identify multiple features that distinguish between the tie-strength among users and developed a classification model on those features. They reported a prediction error of $33km$ with a higher recall, 80%, compared to [17]. Rout et al. [19] first, trained an SVM classifier on various features representative of different aspects and characteristics of users network. Then, they used the trained model to infer location for non geo-active users. Their model is applied to a dataset across the UK and their prediction level is explicitly mentioned to be in the city level granularity.

A more recent group of solutions focused on graph-partitioning to improve the prediction results. Jurgens [20] proposed a solution based on label propagation [21] to predict location for non geo-active users by propagating the location of geo-tagged tweets from their geo-active friends through the social network. Their solution significantly improved the previous results with an estimated MED of $10km$. Kong et al. [22] developed a solution based on social-tie strength. They weighted ties using local clustering coefficient (number of common friends). Compton et al. [4] extended Jurgens' [20] approach using a weighting mechanism based on the frequency of mutual mentions among friends to distinguish between strong and week ties in social graph and prevent the dissemination of noise through weak ties. They achieved the best-reported results of $MED = 6.8km$ for 80% of users in their dataset. We compare our approach with their method to show the efficiency of our solution by considering the temporal aspects over social network partitioning.

We also found a solution by Sadilek et al. [23] who proposed a method for combining the temporal aspects with social network properties. Their approach is very similar to ours with respect to the temporal characterization of the Tweets. However, they make a strong assumption stating that each user has at least one geo-active friend among her direct neighbors in the social graph and they predict location using only one iteration of the label propagation algorithm. Therefore, their approach cannot be considered as a real social network based solution thus, not comparable with ours.

III. Solution

The main idea is to identify users' location by dividing them into location-specific groups using their friendship graph and the time of their tweets. The underlying assumption is that: ***a user's social network friendship structure is affected by her spatio-temporal pattern***. In particular, we assume that Twitter friends tend to appear in (hence tweet from) same locations during same time-slots. Based on that, we designed a hierarchical solution to first, extract spatio-temporal similarity groups by combining partitioning and temporal categorization and then, infer the geographical location of non geo-active users from their geo-active friends in each group. Let's present two definitions, *Time-Slot* and *Location* that are required to explain the details of the algorithm.

A. Definitions

Time-Slot. Time has two components: a linear that represents continuity (frequency), and a circular that captures periodicity in human behavior. Studies [5] show that people appear statically (e.g., for a short period of time) in a same set of locations and repeat this behavior frequently and periodically in time. For example employees of a company frequently visit their work location between 8-16 periodically during weekdays. Our goal is to capture these spatio-temporal patterns of users' behavior on Twitter and use it to predict their location. Thus, we define *Time-Slot (TS)* as the smallest division in time during which the spatio-temporal behavior of users fulfills the following two properties:

- They are geographically static, meaning that their longest travel distance (longest distance between their messages) does not exceed a specific threshold.
- They frequently and periodically appear in the same location during the same TS.

For example, *Monday 8 AM to 6 PM* is a valid TS during which users often appear and stay in a specific bounded geographical location (e.g., their work locations, schools,

universities, etc.) and they repeat this behavior periodically and frequently in time (e.g., every Monday during most of the weeks of a year).

Location. The main intention behind geo-localization is to find the places that a user visits and spends time in a regular manner. Knowing this information is vital for applications like targeted advertisement. For example it is intuitively accepted that a day worker often has her lunch in a place close to her working area and therefore, identifying her work location is important to send her correct restaurant advertisements.

Based on that, we define a location as a place or an area with a limited geographical boundary, where a user tend to visit and stay during a specific TS. We consider multiple locations per user depending on the number of TSs. Each location is represented as the *Geometric Median* of the locations of the tweets published by the user during the corresponding TS. Geometric median is defined as the location with the minimum total distance to all GPS locations of messages in a specific TS [20]. Given a set of geo-locations $L = \{l_1, l_2, ..., l_n\}$ of tweets published by a user U during a specific TS (e.g., during working hours on weekdays), the user's location in this TS (e.g., work location) l_u, is then calculated as,

$$\underset{l_u}{\operatorname{argmin}} \sum_{i=1}^{n} D(l_u, l_i) : \forall u \in L$$

where D is the geographic distance between two geo-locations. We measured D using the *Haversine* [24] formula, which determines the distance along the surface of the earth.

B. Social Graph Partitioning

The first step is to divide users into social groups by extracting topological partitions from their friendship graph. We use a well-known graph partitioning algorithm, called *Louvain* [25], to extract the social partitions from the underlying friendship graph. The algorithm is based on modularity optimization and tries to find the best partitioning by maximizing a value called *Modularity*. Modularity is the average ratio between edges inside a partition to the edges outside a partition over all partitions in the network. The value of modularity is calculated using the following formula:

$$Q = \frac{1}{2m} \sum_{vw} [A_{vw} - \frac{k_v . k_w}{2m}] \delta(c_v, c_w)$$

where m is the number of edges in the graph, $\frac{k_v . k_w}{2m}$ is the probability of an edge between two nodes v and w in a random graph with m edges. A_{vw} is 1 if v is connected to w and 0 otherwise. δ indicates community membership of the two nodes, v and w, and its value is 1 if both are in the same community and 0 otherwise.

The algorithm tries to maximize the modularity by first assigning each node to its own partition and then, aggregating two partitions if their aggregation causes higher modularity value. Then, the algorithm replaces the two aggregated partitions with a single node in the graph and repeats this process until the modularity stops improving. Finally, the algorithm returns the resulting partitioning as the partitioning of the graph. A common practice to avoid local optima is to run the algorithm multiple times and choose the partitioning with the highest modularity value among all runs. In our experiments, we choose the value 10 for the number of runs for each partitioning task in each experiment.

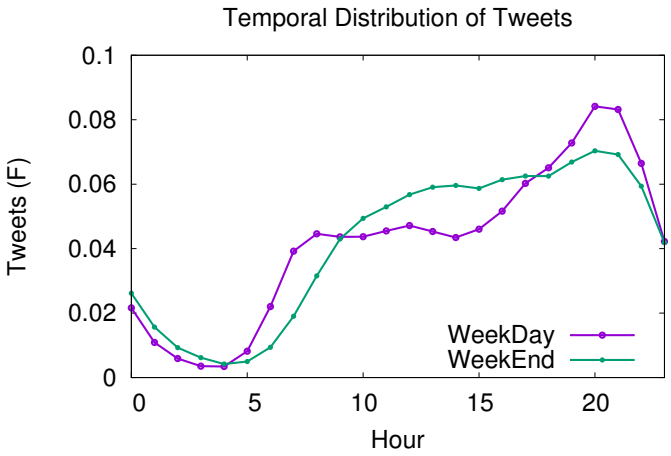


Fig. 2: Temporal Distribution of tweets in the dataset. The sharp changes in separating hours between different time-slots (work, home, and leisure) are clear. Also, an explicit shift during home hours 0-7 can be distinguished between weekend and weekday, which is an indication of variation in users temporal behavior between the two groups.

C. Temporal Categorization

After partitioning the social graph, the next step is to categorize users in each partition into different classes based on the time-stamp of their messages. The goal is to find a TS that captures stability and periodicity in the user's temporal behavior. To achieve this goal we made an analysis of the temporal behavior of the users by observing the temporal distributions of their tweets. Figure 2 shows a comparison between two distributions of the average frequency of tweets per hour over weekdays and weekends in our dataset. We observed two traits that were considered as clues in our temporal categorization model. First, the notable difference between the two distributions during daytime (e.g., 5 to 19), which is an indication of different periodic behaviors (e.g., working vs staying home). Second, the significant variations in the frequency of messages in specific hours of the day (e.g., 8 or 18), which indicate the shift in the static behavior of users in time.

Based on this analysis, we consider two levels for temporal categorization: *Weekly*, and *Daily* to account for frequency,

periodicity and static properties of a TS. In particular, we consider two types of days during a week (i) weekdays (from Monday to Friday) and (ii) weekends (Saturday and Sunday). Then, for each day we divide the time of the day into three time-stamps: Home (from 0 to 7), Work (from 8 to 18), and Leisure (from 19 to 23). Since we want to show the effect of the temporal distinction between two types of weekdays, we consider the same temporal division (home, work, and leisure) during both weekdays and weekends. Even though it does not make sense to consider work location during the weekend but the literal meaning is not considered and it is just a matter of labeling.

Furthermore, home and work are TSs when a user is in specific bounded areas commonly known as her home and work locations, respectively. Whereas, leisure represents a TS when the user spends her time in multiple locations other than her home and work. Note that the leisure time is still considered as a valid TS as it satisfies the corresponding properties. In particular, we believe that even though users tend to visit more than one location during this TS but they are still static during their visiting time and they tend to visit the same location frequently (e.g., the same restaurant close to their home location) and periodically (e.g., every weekend or every evening). The main drawback of considering multiple locations during a TS is that it can affect the prediction error, which will be discussed further in the Section V.

It is clear that defining the TS based on the global temporal distribution of the messages only captures the dominant behavior among a majority of users and not everyone. Therefore, more sophisticated mechanisms are required to develop a categorization model that represents various temporal behaviors among all users. One approach is to use time-series analysis to extract common temporal patterns among users and combine those patterns with graph partitioning for location identification. Another approach is to combine both features by enriching the underlying social network graph using temporal information such that different spatio-temporal patterns construct different partitions in that graph and extract the patterns using graph partitioning algorithms. Such complex modifications and designs can improve the geo-localization results. However, making these analyses is beyond the scope of our work. Thus, we only focus on our hand-tagged non-automatic temporal categorizations that still provide a relatively significant improvement over the-state-the-art.

D. Location Prediction

The next step, after extracting the socio-temporal partitions, is to apply prediction on each partition in order to compare predictability among different groups. In this step, we apply 5-fold cross-validation following common practice, 80% training, 20% testing. First, we divide each socio-temporal partition into training and testing groups. Then, we extract the geometric median of the users in the training group as prediction location for users in the test group. Finally, we calculate the error for each user as the distance between her true location and the predicted location and compare the median error distance among users in different groups.

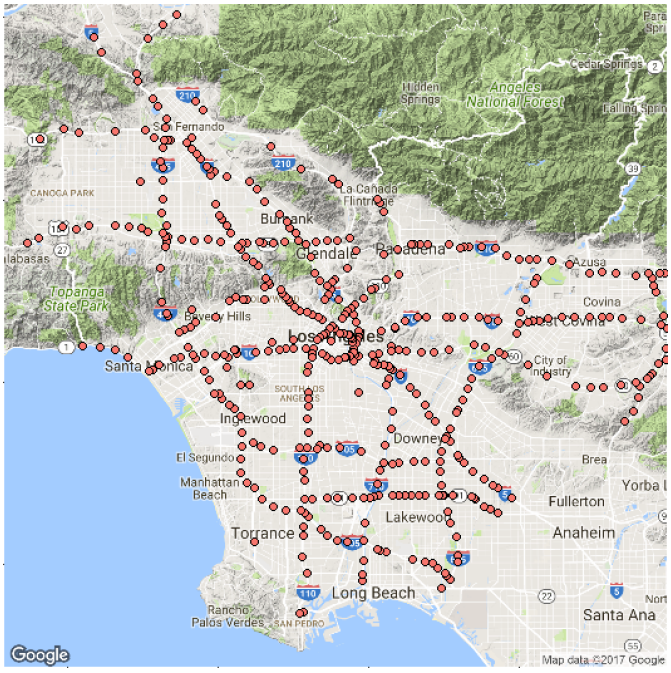


Fig. 3: A set of geographical locations of tweets published by a sample user account called, *Total Traffic LA* during autumn 2014 over the city of Los Angeles. As we can see the locations cover almost the entire urban area of the city.

IV. Experiments

A. Datasets

We perform our experiments on a Twitter dataset collected from 2010 to 2014. The dataset contains $9.8M$ tweets published by $59K$ users. To collect the dataset first, we crawled user statuses using Twitter Streaming API. Then, we extracted all user Ids and created the corresponding social network by crawling the list of friend and follower Ids for each user using Twitter search API [26]. Studies [26] show that Twitter Streaming API allows the collection between 1% to 40% of tweets in near real-time. However, to guarantee the credibility of our dataset and make sure that our dataset reflects the size of a real Twitter dataset we used a distributed cloud-based mechanism to collect a large number of geo-tagged messages, in parallel, from multiple machines with different IP addresses, where the messages were geographically bounded over the entire US.

Data Cleansing. Social network accounts do not only belong to individual users with normal spatio-temporal human behavioral patterns, explained in our assumptions. In fact,

TABLE I: A small sample of user accounts who published more than 95% of their tweets from an exact same geographical location. The accounts belong to advertisement, forecasting, and public news services also known as social bots.

Id	#Tweets	Title	**Description** (extracted from the description section in their respective profiles)
1	2772	Loans-N-Homes	A real estate company
2	2112	Global Consultant	Beauty, Luxury, Education & Sales
3	1735	Loans-N-Homes	Providing Real Estate Services in Southern California for over 10 years.
4	995	Los Angeles Now	Breaking news and weather updates from Los Angeles.
5	812	Lawndale Weather	Weather updates, forecast, warnings and information for Lawndale, CA.
6	607	Espectaculos al Dia	Visit live nation entertainment for concert tickets.
7	502	Call Touba	Recharge the mobile phones of your loved ones back home OR recharge your own...
8	458	LakersSRH	Lakers SportsRoadhouse, Lakers Breaking news, Links, & Licensed Merchandise.
9	421	Creative Galina	CEO and jeweler behind DumbbellJewelry

there are *social bots* (like weather forecast, traffic report and security alert) or *professional accounts* (like organizations, news agencies, and celebrities), which are used for public news dissemination or business advertisement purposes. Figure 3 shows one such account, called *Total Traffic LA*, that is used for continuous reporting of Traffic incidents and their locations over the Los Angeles. The figure shows a small sample of reports during autumn 2014 over the city of Los Angeles where the locations cover almost the entire urban area of the city. Table I shows another example of social bots, which belong to organizations and public news agencies that publish more than 95% of their messages from an exact same location. As we can see, the spatio-temporal behavior of the users in both groups is in contrast with our expected pattern, a few high-frequent and static locations, for a normal human user. Such users significantly bias the analysis when it comes to geo-location identification.

To prevent the negative effect of these accounts we must identify and remove them from our dataset. However, removing these accounts is not a trivial problem. In fact, it is an open research question by itself. For the purpose of this work, we follow the methods used by [23] and [27]. First, for each user we group all her locations within 100 meters of each other and call it a *unique* location. Then, we remove all users with less than 5 and more than 50 unique locations which are $15km$ away from each other. This cleaning resulted in the pruning of around 30% of the messages and about 8% of the users. Thus, the remaining dataset contains $6.4M$ tweets that were published by around $54K$ users.

B. Evaluation Metrics

MED. We use *Median Error Distance (MED)* for evaluating the results. To extract the MED, first, we calculate the error distance between the predicted and the expected geo-locations of the users in each specific socio-temporal group. Then, we extract the median of the predicted errors among all users in that group as the evaluation of our prediction model.

Another measure used by some approaches (like [3], [11] and [28]) is called *Average Error Distance (AED)*, which uses the mean value among prediction errors. We do not use AED since it is more sensitive to anomalies in the dataset compared to MED. For example, assume the following set of prediction results $E = \{3, 4, 3, 2, 6, 8, 50, 3, 1, 100\}$ that contains the prediction errors from the true location in kilometers for 10 sample users. As we can see, the $MED = 3.5km$ provides a much better estimator of the true distribution of the errors compared to $AED = 18km$ among all users.

C. Experimental Settings

We run two sets of experiments *Temporal Categorization* and *Partition Feature Analysis* to examine the effect of temporal aspect of the tweets and the characteristics of the social network partitions on users' geo-location identification, respectively. We use a python implementation [29] of the Louvain [25] algorithm for social network partitioning in our method. All experiments are executed on a machine with 48 cores of 2GHz CPUs and 100GB RAM.

Temporal Categorization. To show the effect of temporal categorization on geo-location prediction we run a set of experiments over different TSs and compare the results with the baseline method. The experiments are made in two different groups: weekday and weekend and we consider three different TSs in each group: Home (H), Work (W) and Leisure (L), as explained in Section III. We refer to these experiments as "SN_TI_H", "SN_TI_W" and "SN_TI_L", which stand for temporal categorization over social network partitioning during home (0-7), work (8-18) and leisure (19-23) TSs.

To further analyze the effect of TSs on geo-location predictability we run another set of experiments over a different set of TSs. We consider the same day types, namely, weekday and weekend over TSs of smaller and fixed span of 3 hours. We run experiments over 8 different temporal categories of 3 hours each $0-2$, $3-5$, $6-8$, $9-11$, $12-14$, $15-17$, $18-20$ and $21-23$. The experiments are shown with an $X-Y$ label, where X represents the starting time and Y shows the ending time of the publication of the tweets in the corresponding group. For example, $9-11$ shows the prediction results over tweets published from $9am$ to $11am$.

Partition Feature Analysis. One important limitation against the improvement of the prediction accuracy roots

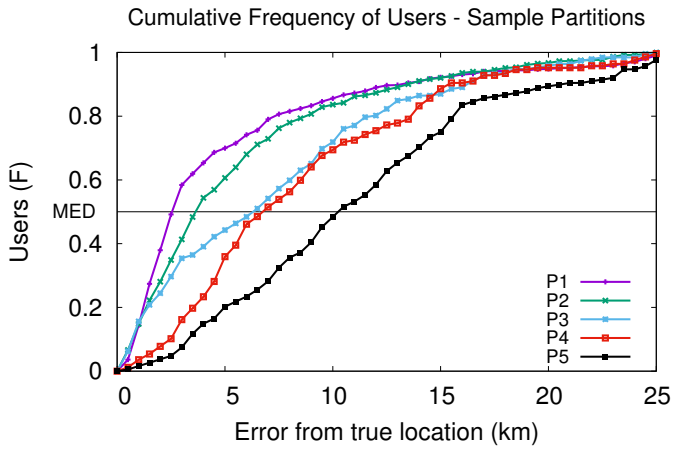


Fig. 4: A sample of 5 social network partitions over the entire dataset to show the effect of *partition_size* on geo-location predictability for different social network partitions. The partitions are sorted by their size from largest $P1$ to smallest $P5$. As we can see, the prediction accuracy significantly reduces by reducing the size of the partitions.

down to the partitioning itself. In fact, partition feature analysis is a research area by itself that is used in many other applications like user profiling [17] [30] or location identification [20]. We hypothesize that not every social network partition is a good representative of geo-location proximity and following that, we analyze a set of partition-specific features to identify the most effective feature with respect to geo-location predictability. In particular, we run a set of experiments on multiple partition-specific criteria including *message_count*, *number_of_users*, *relative_partition_density* and *partition_size*. The experiments reveal a strong correlation between the *partition_size* and geo-location predictability. More specifically, we find that there are a large number of smaller partitions that result in a low prediction accuracy. Also, we realize that pruning those small partitions from the analysis and their corresponding users and tweets from the dataset significantly improves the prediction accuracy on other partitions. Figure 4 shows the prediction error on a sample of 5 partitions, sorted by size, from the dataset, during working hours on weekdays. Each curve shows the prediction accuracy for the users in the corresponding partition. As we can see, the smaller the partition size, the lower the prediction accuracy becomes. Following this intuition, we designed this experiment to prune the dataset by increasing the minimum partition size and examine the effect on average prediction accuracy while reducing the recall.

Baseline Method. We compare the results of our algorithm with one of the best solutions, based on social network partitioning, developed by Compton et al. [4], which reported the best results on geo-location identification, to our knowledge. The algorithm was developed following the explanations in the paper. We refer to these experiments as *SN*, hence the name *Social Network* based approach. SN, as explained in Section I, only uses a social network partitioning algorithm for geo-location prediction. The algorithm is based on label propagation through the friendship ties over the underlying social network graph.

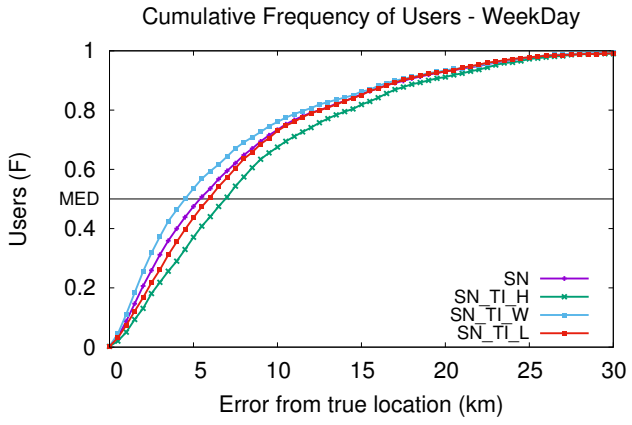
V. Result and Discussion

Temporal Categorization. Figure 5 shows the cumulative frequency of users sorted by their prediction error. The figure depicts two sets of experiments: $\{SN\}$ as the baseline and $\{SN\_TI\_H, SN\_TI\_W, SN\_TI\_L\}$ representing our hierarchical algorithm with temporal categorization. 5-a shows the results during weekdays and 5-b during weekends. The horizontal black line in the middle represents the MED on both charts.

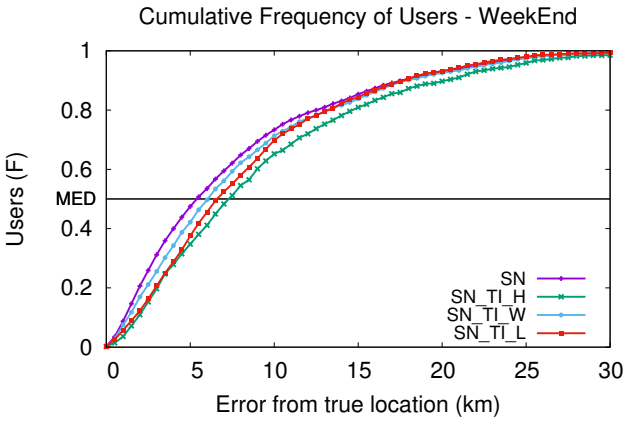
The most important outcome in this experiment is the variation of the prediction accuracy during different TSs that confirms our hypothesis regarding the dynamics of the human spatio-temporal behavior. In addition, we can see that our solution outperforms the baseline method SN during working hours in *weekdays*, (SN_TI_W), which indicates the integrity of considering temporal aspects. In particular, we achieve $MED = 4.5km$, which is 16.6% improvement over the same experiment on SN approach with $MED = 5.4km$. This result is significantly confident with $p < 0.01$ on 95% confidence interval over 100 runs of the partitioning.

Other TSs show a lower performance compared to the SN model. This can be due to many reasons including the limited number of messages during home hours (Figure 2), high diversity of locations during leisure time or the fact that people are less predictive during weekends than weekdays. However, understanding the real mechanism behind those negative outcomes requires further analysis with stronger temporal modeling that we consider as a future work.

Figure 6 shows the comparison of the prediction results between the baseline method $\{SN\}$ and our approach over the fixed length TS of 3 hours (3H). In general, the results are similar to $\{SN\_TI\_X\}$ experiments where the best prediction is achieved during working hours on weekdays. However, a closer look at the results shows that the 3H experiments provide more detailed explanations and better predictions over smaller time-slots. For example, it shows that the predictions during weekdays can be improved (i) over the home time-span by excluding the results from the tweets published between $3am$ to $5am$, $3-5$, and (ii) over the leisure time-span by only including the tweets published between 18-20. Also, we notice that the tweets that published between 6-8am during the weekends provide the worst prediction results in that group, which again excluding them can improve the overall prediction results on the corresponding time-span.

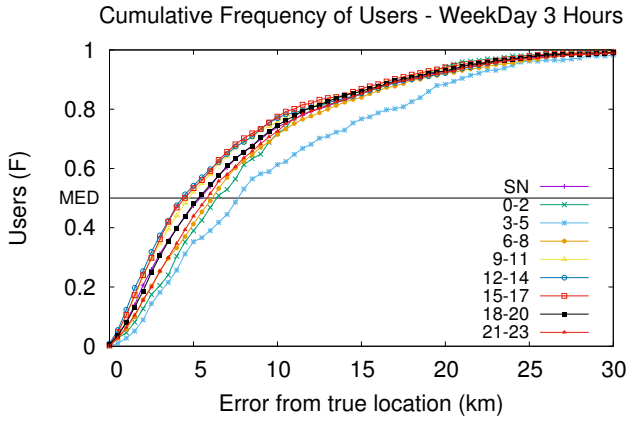


(a) WeekDay

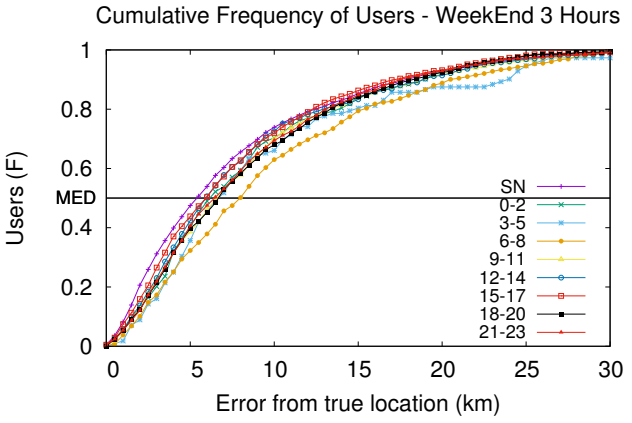


(b) WeekEnd

Fig. 5: Cumulative frequency of users sorted by prediction error from their true location. Comparison between the baseline approach, SN, and our solution over three TSs, (i) $Home(0-7) = SN\_TI\_H$, (ii) $Work(8-18) = SN\_TI\_W$ and (iii) $Leisure(19-23) = SN\_TI\_L$, during weekDays (a) and weekEnds (b).



(a) WeekDay 3 Hours



(b) WeekDay 3 Hours

Fig. 6: Comparison of the prediction accuracy between the baseline approach, SN, and our solution over the fixed length TS of 3 Hours. The experiments are labeled as $X - Y$ where X is the starting time and Y represents the ending time, during weekDays (a) and weekEnds (b).

Partition Feature Analysis. The objective is to reduce the recall by pruning the small partitions until we find a threshold where the prediction accuracy does not significantly improve. We run three experiments with different thresholds on the minimum partition sizes of 50, 500, and 800, which consequently result in 57%, 41% and 33% recall. The experiments are shown as $SN\_TI\_W\_X$ in Figure 7, where X represents the recall in each experiment. The maximum improvement in the accuracy was reached on $Recall = 0.33\%$ with $MED = 3.13km$, which is around 42% improvement over the baseline method.

This result shows that the diversity in the partition_size has a strong effect on geo-location predictability. There is a large room for improvement by correctly designing a strong partitioning that is customized for location predictability, which opens another branch of research for our future work.

VI. Conclusion

We developed a solution to improve the geo-location identification of users on Twitter by considering the temporal and partition-specific parameters. The solution takes into account the temporal dimension of the users on Twitter to improve the accuracy of the predictions based only on social network partitioning. The principal assumption is that a user's social network friendship structure is influenced by her spatio-temporal behavior. Based on that, we first partition the users into homogeneous friendship groups using

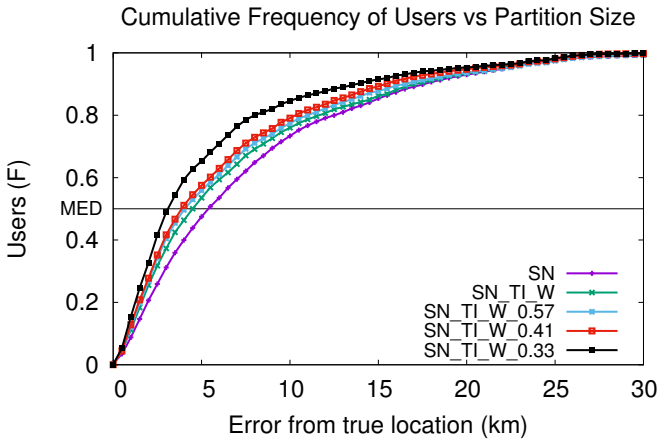


Fig. 7: The effect of *Partition Pruning* on geo-location predictability. SN is the baseline method, SN_TI_W is our method with temporal categorization with no pruning and $SN\_TI\_W\_X$ represents experiments based on temporal categorization with different pruning values on recall, where X shows the recall. The best prediction accuracy $MED = 3.13km$ is achieved over 33% of the users $Recall = 0.33$ by pruning the smaller partitions.

an off-the-shelf partitioning algorithm. Then, we apply a temporal categorization over each group to extract more fine-grain spatio-temporal sub-partitions among which the users tend to be static with respect to their geographical location. Finally, we apply prediction over each partition using standard cross-validation method. We examine the accuracy of our algorithm on a large-scale Twitter dataset and compare the results with one of the approaches with the best-reported results. Our solution outperforms the state-of-the-art by only taking the temporal dimension into account while maintaining the same recall. In addition, we show that the prediction's accuracy is strongly influenced by the size of the underlying social network partitions and we can significantly improve the accuracy by limiting the prediction to larger social network partitions.

One way to improve our model is by considering dynamic temporal components in the model such that the algorithm can automatically decide on the temporal categorization. Another approach is to consider an overlapping partitioning that assigns each user to multiple social community groups. Thus, the algorithm can choose the prediction from different partitions during different time-slots. Moreover, the analysis can be performed to discover and apply stronger geo-location distinguishing factors on social network partitions that can improve the quality of the underlying social network partitioning with respect to geo-location predictability.

REFERENCES

[1] N. B. Lassen, R. Madsen, and R. Vatrapu, "Predicting iphone sales from iphone tweets," in *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*, Sept 2014, pp. 81–90.

[2] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, "Chatter on the red: What hazards threat reveals about the social life of microblogged information," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '10. New York, NY, USA: ACM, 2010, pp. 241–250. [Online]. Available: http://doi.acm.org/10.1145/1718918.1718965

[3] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: A content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 759–768. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871535

[4] R. Compton, D. Jurgens, and D. Allen, "Geotagging one hundred million twitter accounts with total variation minimization," *CoRR*, vol. abs/1404.7152, 2014. [Online]. Available: http://arxiv.org/abs/1404.7152

[5] 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*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1082–1090. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020579

[6] J. McGee, J. Caverlee, and Z. Cheng, "Location prediction in social media based on tie strength," in *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, ser. CIKM '13. New York, NY, USA: ACM, 2013, pp. 459–468. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505544

[7] Y. Takhteyev, A. Gruzd, and B. Wellman, "Geography of twitter networks," *Social Networks*, vol. 34, no. 1, pp. 73 – 81, 2012, capturing Context: Integrating Spatial and Social Network Analyses. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378873311000359

[8] J. Goldenberg and M. Levy, "Distance is not dead: Social interaction and geographical distance in the internet era," *CoRR*, vol. abs/0906.3202, 2009.

[9] C. Wang, J. Wang, X. Xie, and W.-Y. Ma, "Mining geographic knowledge using location aware topic model," in *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, ser. GIR '07. New York, NY, USA: ACM, 2007, pp. 65–70. [Online]. Available: http://doi.acm.org/10.1145/1316948.1316967

[10] 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*, ser. EMNLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1277–1287. [Online]. Available: http://dl.acm.org/citation.cfm?id=1870658.1870782

[11] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. New York, NY, USA: ACM, 2012, pp. 769–778. [Online]. Available: http://doi.acm.org/10.1145/2187836.2187940

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[13] B. P. Wing and J. Baldridge, "Simple supervised document geolocation with geodesic grids," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 955–964. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002593

[14] P. Serdyukov, V. Murdock, and R. van Zwol, "Placing flickr photos on a map," in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 484–491. [Online]. Available: http://doi.acm.org/10.1145/1571941.1572025

[15] S. Kinsella, V. Murdock, and N. O'Hare, ""i'm eating a sandwich in glasgow": Modeling locations with tweets," in *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, ser. SMUC '11. New York, NY, USA: ACM, 2011, pp. 61–68. [Online]. Available: http://doi.acm.org/10.1145/2065023.2065039

[16] G. Li, J. Hu, J. Feng, and K. l. Tan, "Effective location identification from microblogs," in *2014 IEEE 30th International Conference on Data Engineering*, March 2014, pp. 880–891.

[17] R. Li, S. Wang, and K. C.-C. Chang, "Multiple location profiling for users and relationships from social network and content," *Proc. VLDB*

Endow., vol. 5, no. 11, pp. 1603–1614, Jul. 2012. [Online]. Available: http://dx.doi.org/10.14778/2350229.2350273

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, mar 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[19] D. Rout, K. Bontcheva, D. Preoţiuc-Pietro, and T. Cohn, "Where's @wally?: A classification approach to geolocating users based on their social ties," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, ser. HT '13. New York, NY, USA: ACM, 2013, pp. 11–20. [Online]. Available: http://doi.acm.org/10.1145/2481492.2481494

[20] D. Jurgens, "That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships," in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013, pp. 273–282. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6067

[21] Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," *School Comput Sci Carnegie Mellon Univ Pittsburgh PA Tech Rep CMUCALD02107*, vol. 54, no. CMU-CALD-02-107, pp. 1–19, 2002. [Online]. Available: http://discovery.ucl.ac.uk/185718/

[22] L. Kong, Z. Liu, and Y. Huang, "Spot: Locating social media users based on social network context," *Proc. VLDB Endow.*, vol. 7, no. 13, pp. 1681–1684, Aug. 2014. [Online]. Available: http://dx.doi.org/10.14778/2733004.2733060

[23] 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*, ser. WSDM '12. New York, NY, USA: ACM, 2012, pp. 723–732. [Online]. Available: http://doi.acm.org/10.1145/2124295.2124380

[24] C. C. Robusto, "The Cosine-Haversine Formula," *Source: The American Mathematical Monthly*, vol. 64, no. 1, pp. 38–40, 1957. [Online]. Available: http://www.jstor.org/stable/2309088{%}5Cnhttp://about.jstor.org/terms

[25] A. Clauset, M. E. J. Newman, , and C. Moore, "Finding community structure in very large networks," *Physical Review E*, pp. 1– 6, 2004. [Online]. Available: www.ece.unm.edu/ifis/papers/community-moore.pdf

[26] P. Meier. Crowdsourcing crisis information from syria: Twitter firehose vs api. [Online]. Available: https://irevolutions.org/2013/05/30/twitter-api-vs-firehose/#comments

[27] D. Jurgens, T. Finnethy, J. McCorriston, Y. T. Xu, and D. Ruths, "Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice," *The 9th International Conference on Weblogs and Social Media (ICWSM)*, pp. 1–10, 2015.

[28] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, "Supervised text-based geolocation using language models on an adaptive grid," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ser. EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 1500–1510. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390948.2391120

[29] A. Lancichinetti. A not so small collection of clustering methods. [Online]. Available: https://sites.google.com/site/andrealancichinetti/clustering_programs.tar.gz?attredirects=0&d=1

[30] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: Unified and discriminative influence model for inferring home locations," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 1023–1031. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339692