Addressing the Sparsity of Location Information on Twitter

Dimitrios Kotzias
University of Athens
dkotzias@di.uoa.gr

Ted Lappas
Stevens Institute of Technology
tlappas@stevens.edu

Dimitrios Gunopulos
University of Athens
dg@di.uoa.gr

ABSTRACT

Micro-blogging services such as Twitter have gained enormous popularity over the last few years leading to massive volumes of user generated content. In combination with the proliferation of smart-phones, this information is generated *live* from a *multitude* of content contributors. Interestingly, the content and timestamp of tweets is not the only information that can produce useful knowledge. The location information of users is of great significance since it can be utilized in a variety of applications such as emergency identification, tracking the spread of a disease and advertising. Unfortunately, information regarding location is very rare since many users do not accurately specify their location, and fewer posts have geographic coordinates. In this work, we aim to confront this data sparsity issue. Utilizing Twitter's social graph and content, we are able to obtain users from a specific location. We optimize our method to work with minimum amount of queries considering the large volume of data in such settings. We also provide a mechanism for geo-locating a tweet within a city and present the qualitative enrichment in our data, achieved by our method.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Data Mining; J.4 [**Social and Behavioural Sciences**]: Sociology

General Terms

Algorithms

Keywords

Social Networks, Data Sparsity, Location Profiling

1. INTRODUCTION

Over the last decade on-line platforms where individuals generate and contribute content have gained massive popularity. In most of these platforms individuals are connected, establishing social networks. Such networks attracted interest from various scientific fields due to the numerous research challenges and the plethora of potential applications. A typical example is Twitter currently hosting more than 200 million users contributing more than 400 million tweets per day [12]. Twitter is unique among the social networks since it propagates information to an immense amount of people very quickly, in fact many times, faster than conventional news networks [23]. Combined with the increasing popularity of smart phones, and mobile internet availability, users generate content from a variety of locations in real time, creating a very diverse and spatio-temporarily spread-out network of information.

Naturally, with the formation of such a network, there has been an increase in research as well as applications for spatio-temporal data [18, 13, 8]. These applications utilize the content, time of creation as well as location of data. Location information is important as it can transfer knowledge from the online back to the real world, and aid towards personalized and localized information services. It is invaluable for a variety of applications such as discovering the way diseases spread [6], emergency identification and response [25], localized event detection and relevant news propagation [28], analysis of the behavioural patterns and mobility of people within a city [8, 10] as well as online advertising [1]. Furthermore, social networks themselves are also adopting a local-focus philosophy, with examples of Twitter recently starting reporting local trends[1], while search engines consider your location among other factors when returning results[2].

Despite the fact that user contributed content in such platforms is always characterized by a well defined timestamp, unfortunately location information is very sparse. Research in Twitter suggests that a big percentage of users either do not provide their location information in their profiles, or input noisy data [7], with only 48% of users providing an actual location with city or lower level accuracy [11]. Moreover, the number of tweets with geographical coordinates is much lower, in the order of 1% [26]. Reasonably, this sparsity of information constitutes a major issue for all the aforementioned applications that require it.

In this paper, we address the problem of sparsity of location data by providing a framework that utilizes information from the content of a users' tweets, as well as the social graph around her. Considering the large volume of data in such settings and the fact that our system works with the Twitter API, we take into account the limitations set by such systems and optimize our method for the minimum amount of queries. Our contribution is threefold:

[1]https://blog.twitter.com/2010/now-trending-local-trends
[2]http://www.google.com/landing/now/

- We formalise the problem of location identification using both social graph information and text information as a resource utilization optimization problem in the context of the limitation in the Twitter system.
- We analyse the locality of the social graph in Twitter, and the connectivity between users who live in geographical proximity.
- We provide a framework for (i) identifying more users at a location of a city-level granularity, (ii) attaching geographical coordinates to individual tweets within that city.

In addition to the precision of our model, we also perform quantitative and qualitative analysis on how the increase in users, and thus tweets, enriches our data by extracting topic models and analysing them. We chose to apply our method on Twitter due to (i) its increased academic interest, (ii) large volume of active users and tweets and (iii) its unique nature as a fast information and events propagator.

2. RELATED WORK

Related work can be broadly divided into research done for identifying the location of a user, or the location of each individual tweet.

In the first category, Eisenstein et al. attempt to solve the problem through geographical topic models [9]. They capture the difference in the use of language for a specific topic, between people from distant areas. They are able to predict the location of a user with an error mean distance of 900 km, and achieve a 27% accuracy when predicting the state of a user. Their results indicate that people have different ways of discussing a topic in different areas, however, there has to be a significant distance between these areas. Within a city level granularity, people are more likely to have similar conversational habits. More recently, Ahmed et al. in [2] proposed a tree-like hierarchical structure of topics, at which the lower levels of the tree, represent more specific versions of the general topics at the parent nodes. This way, they are able to extract location specific topics, and place users with an average error of 298 kilometers in the same dataset. Cheng et al. in [7] utilized the locality of phrases rather than topics, and manage to pinpoint the city of 51% of users, within 100 miles of their actual location. Mahmud et. al. in [22] improve upon this method by identifying *named local words*. These terms are very local terms such as the name of a location or the name of places, retrieved from services such as foursquare. They use a Multimonial Naive Bayes classifier and test hierarchical algorithms that first predict the country and then the state of a user, to estimate the city of a user. They achieve an accuracy of 58% for 100 miles radius.

These approaches take advantage of the difference in the use of language about specific topics or words which identify the location of users, however in cases of smaller areas, where there are not many language differences, these methods would not perform well. Our work differentiates at two fundamental levels. Firstly, we take advantage of a users writings as well as the relationships in the social graph, which enables us to predict location with greater geographical detail. Secondly we also propose a way to attach exact geographical coordinates in a tweet level, after we have identified the location of the user who created it.

In the second category, Ikawa et. al in [14] attempt to estimate the location of a tweet by associating expressions with locations. For each query tweet, they find the location with the closest word list and place it there. However, the underlying assumption that people will tweet about the place they are in, and then express their feelings about it, does not always hold, and they achieve an accuracy of 14% for a radius of 5 kilometres. Li et, al achieve a better precision for the same problem as they attempt to identify Places of Interest (POI) a tweet may belong to. [21]. They build a Language Model for each POI, based on tweets that occurred there and information crawled from websites, and then rank the KL-divergences for each query tweet, to identify the candidate POI's. They test their method for the 10 top POIs at a city, and reach an accuracy of more than 60% for their best case. However, their accuracy fluctuates greatly based on the number of tweets about a POI, and their premise is somehow unrealistic, since in a real-world scenario, there are much more than 10 possible locations within a city. Kinsella et. al provide a framework, which is closer to our work since it is used to pinpoint the location of both users and tweets in a variety of granularities in [15]. They build language models for each location and test a Query Likelihood model, in order to predict the location of a tweet. Their best results accomplish an accuracy of 31.9% for users for a town granularity and 13.9% for tweet location in zip-code granularity. In their work, they use the same model to solve both problems, while we discriminate between the two and mostly take advantage of the social graph for user location prediction and term-models for tweet location prediction, which enables us to create a more robust system.

Ren et. al [24] take the social graph into account in order to identify a users' location. They place each user to the location of the majority of his friends. They achieve a precision of 59.3%, however they only test it for 704 users. Working on a much larger dataset from facebook data, Backstrom et. al [4] suggest a correlation between friendship and distance on the map, and build a more elaborate model to find the probability of a users home location, given the location of his contacts. They place 67.7% of users correctly, however, both these methods, assume that we know all the friends of the user we are trying to locate, which in very large graphs such as those created by social media is rarely the case. Our method aims to minimize the number of queries at the social graph, and more importantly operates without the knowledge of the location of all the friends of a user.

To the best of our knowledge, despite the large number of relevant papers which focus on estimating the location of a single user, the problem of identifying more users from a specific location, has yet to be solved. In contrast to most of the previous work done in this area we: (i) attempt to identify the location of a user in a *city level* accuracy, which is much more limited than 100 miles radius, set this far, (ii) combine this information to attach geographical coordinates to a tweet, by taking under consideration *both* the relationships of users in the social graph as well as the content of their writings, and (iii) use a computational model based on the Twitter system, optimized for the number of queries.

3. PROBLEM DEFINITION

Location data in twitter are rare, and given the value of such information we attempt to discover more users at a specific location. We analyse data about a city as whole, and

we set two interconnected goals: (i) Discover more users that live in that city and (ii) create a system that is able to attach exact coordinates on a Tweet level within that city. Moreover, we assume that we incur a *cost* each time we make a query about the connections or the location of a user. This assumption is pragmatic because social networks have limitations on how many times one can inquire about such information within a specific time frame. For instance Twitter only allows for 15 queries per 15 minutes for the friends of user. Taking under consideration the size and increase rate of social networks today, this constitutes a practical constrain as well, since it is computationally expensive to process the entire graph. The problems are formally defined below:

Problem 1 - User Level
Given set of users U who tweeted from a specific city C, identify as many users from the same city as possible, by asking at most $\boldsymbol{k}$-queries about the social graph.

Problem 2 - Tweet Level
Given a set of tweets T_1 from a specific city C for which we know their exact geographical location, and a set T_2 for which we do not have such information, attach geographical coordinates to T_2.

We consider the two problems to be interconnected, because by following a hierarchical scheme, the tweets of users identified in C, can be considered as T_2 and be geo-located on the map, with the solution to problem 2.

4. PRELIMINARIES

4.1 Twitter

Twitter consists of messages which have a maximum length of 140 characters, which may or may not include location information. With the proliferation of mobiles phones, many users *tweet* from a variety of locations. However, there is a wide array of topics and uses for a tweet, which in combination with the limited text size, render most of them impossible to analyse and categorize. Users have the ability to have a static location in their profiles, however, according to [11] the location field of Twitter users many times is empty or contains inaccurate or in-comprehensive information.

4.2 Graph Analysis and Motivation

Previous work has indicated that there is a correlation between proximity on the real world and proximity on the online social graph[16, 28, 20, 4]. Furthermore the small world effect is even greater in local communities [3], hence we question whether these effects continue to intensify in a more local level; the area of a city.

We collected tweets from a specific area using the Twitter streaming API, which provides, two types of tweets; a set which has exact geographical coordinates attached to each tweet and one which occurred within the bounds we set, but only has an approximate location. However, not all users who post a tweet from a city, live there. Some may have declared non-existing locations or nothing in their profiles, while others could be just visiting the city. In order to identify the location of a user we used a gazetteer provided by the GeoNames (www.geonames.org), from which we collected locations names from within our specified area, and checked if they matched a users' location. Understandably our evaluation method is not perfect, since location names to not correspond to unique places on the map, however this is an insignificant percentage in our dataset.

In order to test whether locality in the real world is correlated to locality in the network, we apply a set of filters in our data. Initially we re-create a part of the social graph, only considering edges that are bidirectional. Most real world relationships by far exceed the energy of a mere follow back, and thus people that are connected in real life, are more likely to have a bidirectional edge than a unidirectional. The latter are more likely to occur when the two parties do not really know each other in the real world. Furthermore, we discarded very popular users which we defined as having more than $f = 30000$ followers or friends, to avoid celebrities, who tend to be very central nodes, with very little location information.
Our first filter, was then to extract the largest connected component of the sample graph, formed from users who tweeted within C.
Secondly, we applied a measure in an attempt to remove users who are just visiting a city, we only kept users who tweeted more than n times in a period greater than h hours. For this experiment, we set $n = 4$ tweets and $h = 100$ hours.

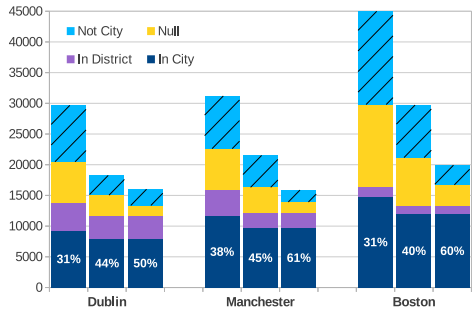


Figure 1: User Analytics for Each City.

The first column of Figure 1 indicates statistics for users, for three different cities, namely *Dublin*, *Manchester*, and *Boston*. As expected in all three cases, more than 50% (the two top boxes) have the field empty or in a location which does not correspond to an area in the city. In contrast only a percentage in the order of 30% has a city level accuracy in their profiles, while the rest belong to a much larger region which includes out target city. We consider as regions, Ireland for Dublin, England for Manchester and the state of Massachusetts for Boston.
Interestingly, these numbers change significantly when we apply our first filter with all location categories being reduced, except for the percentage of people in the city. Since this is only a sample of the graph, being connected in this case, translates to stronger connectivity in the complete graph, which reaffirms that there is a correlation between locality in the real world and the online, and local communities seem to have a high clustering coefficient.
The third column illustrates the data after our third filter. Interestingly the users who pass this filter, and users in the largest connected component have many common nodes, who are mostly users who indeed live in the city. This fact holds true across all three cities which indicates that sample connectivity is a good pruning filter for local users.

Our intuition, based on our analysis, is that local communities on the real graph will form clusters with a high clustering coefficient in the social graph since there is a suggested correlation between location and friendship. From the data given above we can conclude that (i) the majority of users who live and tweet from the same city, are part of a strongly connected graph, with users that live there, and (ii) that connectivity from a sample is a good measure to prune users that do not belong in a city.

5. OUR APPROACH

Our approach formalizes the algorithm which considers the limited resources one has for the Twitter API. For this purpose, we built our methods considering our analysis of the social graph as well as the restrictions one can have in similar settings. Based on these observations, we built a method that is able to identify users from a specific city, by asking the minimum number of queries about the social graph.

5.1 Problem 1: User Discovery

For our first problem we use the `MaxEdge Algorithm`: From our set of tweets T that have geo-location within the area of a city C, we extract the users who created them. We then perform some *enhacement assessments*, based on our Analysis in 4.2 on them and consider this group of users to be our ground truth, from now on referred as *seed*. We then perform our graph discovery with the following algorithm: We create edges with weights from the *seed* to their friends, and create the rest of the known graph, referred to as $frontier$. In our method, initially all the edge weights are set to 1.0. Each of the nodes in the $frontier$ has a score equal to the sum of the weights of the seed connected to it $f_j = \sum w_{ij}$. We then start to crawl by discovering the node with the maximum score. We query for his location and if this person is located in C we add him to our *seed*, query about his connections and update our social graph. We proceed in the same way until we exhaust our limit of k-queries.

Essentially our method is an enhanced first step of a BFS algorithm. Given that the graph in Twitter is mostly connected, a DFS is bound to escape the users living in area quite soon, simulating a random walk. Following this crawling method, at each step we create the most strongly connected graph possible, by maximizing the clustering coefficient of the seed. Given our analysis in Section 4.2, this increases the likelihood of finding users in the same city. In addition we minimize conductance, which is a measure indicating the quality of the community structure a part of a graph has [19]. In our case it is defined as the number of edges between the seed and the frontier over the number of edges inside the seed, which implies that a good community has low conductance. Our method's efficiency is based on the fact that it only queries the nodes that maximize the seed connectivity and minimize the conductance.

A representation of the graph formulated by the users we crawled, is in Figure 2. In this figure, the nodes on the left side, within the bounding box, represent the *seed* nodes denoted as s, which are users who tweeted from the city. As mentioned in section 4.2 a big part of these nodes are connected, while there are also smaller connected parts and nodes with few or no internal edges at all. The red nodes on the right side, represent our current $frontier$. These are

Algorithm 1 MaxEdge Algorithm

Input: A set of users U who tweeted from within C, k
Output: New users in C

```
seed ← ∅
for all <U_i in T> do
  if Assess(U_i) is true then
    seed += U_i
    updateFrontier(FriendsOfU_i)
  end if
end for
while k > 0 do
  NewUser ← MaxWeight(frontier)
  if NewUser in C then
    updateFrontier(FriendsOfNewUser)
  end if
  k ← k − 1
end while
```

nodes we have not queried yet, however they have at least one link with one node in the *seed*, hence we know of their existence, from the edge list of the s. Currently the node with the most edges to s has 4 edges, and if we discover and accept this node as being in C, there is a node with 3 edges who will then have 4 and will be our next query. Our method does not require us to maintain any internal edges within the seed, or nodes without external edges, thus limiting the space requirements and making it plausible to maintain in memory the information needed for a single city.

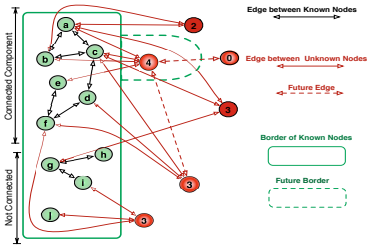


Figure 2: Seed and Frontier Example

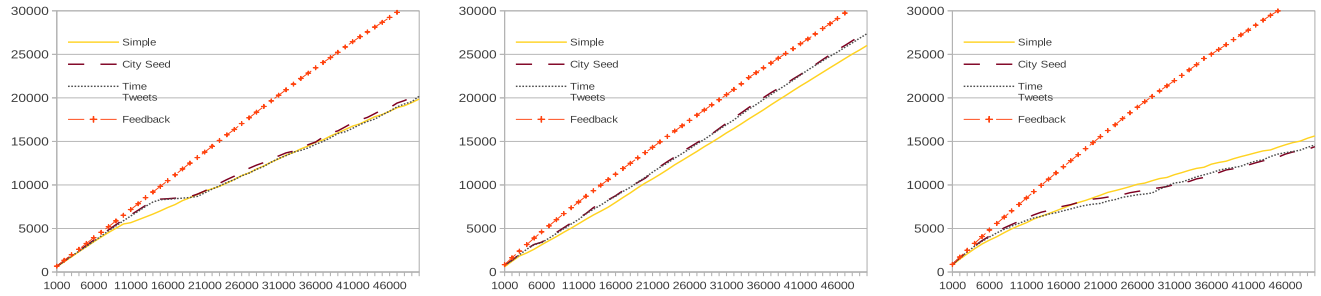
5.2 Problem 2: Tweet Location Discovery

Our goal in this problem is to identify the geographical coordinates of a tweet. We evaluate two methodologies, one assuming we only have knowledge regarding the text of a tweet, and the other assuming we have information about the user who created it as well. For both methods, initially we segment the area, into squares of equal length a.

For the first approach, which we denote as **QL** we consider the Query Likelihood Model as defined in [17], since it provided the best results in [15]. However, instead of having different locations as different documents, we considered all the tweets that occurred in the same box as a document, and assign a query tweet to the box that has the maximum likelihood of having produced that tweet.

For our second method we exploit information about the users and his tweeting habits, to create the method **QLU**. In this method we assign to each square a probability that the user u tweeted from the box b as

$$p(User_u|Box_b) = \frac{|Tweets_u \in b|}{\sum Tweets_u}$$



(a) Dublin (b) Manchester (c) Boston

Figure 3: Users in C v Number of Queries in Twitter

and choose the box with the maximum probability p. If information regarding the user u is not available, we return to the **QL** method to locate it.

We consider this method as the next step of the User Discovery problem in a hierarchical model in order to identify the originating location of tweets. Adding geographical coordinates to a text works better if the possible area is limited, especially in the case of tweets, which contain mostly common words and very few location information.

6. EXPERIMENTS

6.1 Datasets

We run our experiments on geolocated tweets from 3 different cities. Our data were collected between April and August 2013, and their details are presented in Table 1. New Users, refers to the number of extra users we crawled with our method, that were from the same city. The Dublin All dataset, refers to all the tweets from users we were able to crawl that live in Dublin, which was used for the topic model experiments described in section 6.4

	Tweets	Users	Area	New Users
Dublin	1.9 M	43 k	1224 km^2	179 k
Manchester	1.3 M	40 k	462 km^2	70 k
Boston	1.5 M	55 k	1521 km^2	73 k
Dublin All	71 M	220 k	1224 km^2	-

Table 1: Dataset Details

6.2 User Discovery Results

In this experiment we measured how well our method, performs in discovering users from the a city, and illustrate our results for Boston, Dublin and Manchester, for $k = 50000$.

Initially we define as seed s all the users who tweeted from within C with an exact geo-location in their tweets.

`Simple` refers to the algorithm described in 5.1 which takes all these users as its seed.

`City Seed` differentiates by considering as seed, the largest connected component of users who *declare* their in their profiles that they live in C.

`TimeTweets`(n, t) considers as seed the city users who tweeted at least $n = 4$ times with a difference of $t = 100$ hours between the first and last tweet.

`Feedback` refers to a more complex version of our algorithm, which rewards the nodes that pointed to a correct user, and penalises those that pointed to an incorrect one. More specifically, when updating the frontier, we increase or reduce the weights of the edges, of such nodes, by multiplying their current weight with a coefficient $1 \pm c$. For our experiments we set $c = 0.01$. This algorithm uses the same *seed* as `City Seed`.

6.2.1 City Precision

Figure 3 illustrates that our method works well for the first 10 - 15 thousand queries, while filters in our initial seed outperform it slightly. After this point however, the part of the graph which we are trying to maximize its clustering coefficient, becomes too large and too connected with non-local users, and thus our method slowly begins to crawl non-city users. The exact number of dilution in the users, depends heavily on the size of the city, as smaller cities are more interconnected locally while larger cities tend to have more connections to outside communities [4].

This effect can be tackled by using our `Feedback` algorithm, which gave us the highest most stable precision in all cases, finishing with a precision of 62.9% for Dublin, 63.0% for Manchester and 65.6% for Boston. The results indicate that our intuition is correct, users in proximity can be discovered by maximizing connectivity, however after a number of queries, the seed becomes too diluted. `Feedback` essentially crawls the user, who is most likely to belong to a city, by the "majority vote" of all the users who live there.

Precision inevitably drops after a many queries, because the most certain users are already in the seed, and many users will have common connections from the outside world. The `Feedback` algorithm however, learns which users are more inclined to have connections with non local users and assigns less weight to their "votes".

6.2.2 Region Precision

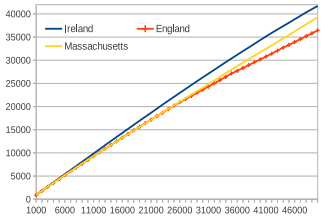
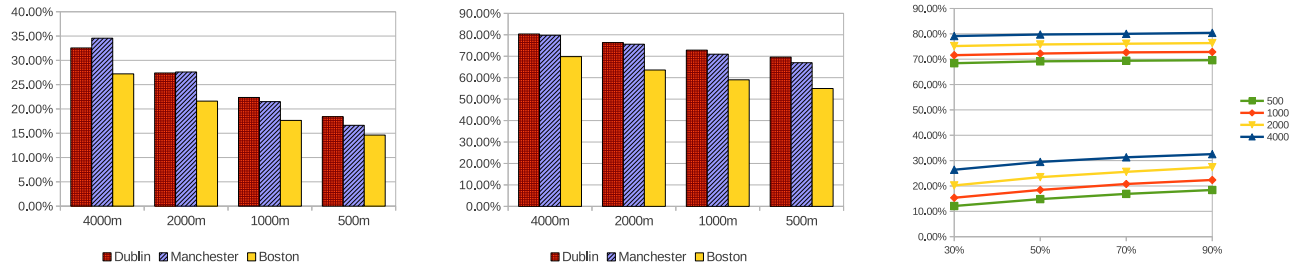


Figure 5: Users in Region v Number of Queries

Figure 5 represents the accuracy for regional results. The



(a) QL method Accuracy (b) QLU method Accuracy (c) Sensitivity to Training Size

Figure 4: Size of Area - Boston: 39x39km, Manchester: 21x22km, Dublin: 34x36km

question we asked for this experiment, is whether focusing in a local area, will produce good precision, when accounting for users living in a greater region. We tested the `Feedback` algorithm, for a greater region surrounding our city. The precision here is higher, in the order of 80% however not proportionally to the increase in the real world size. This indicates that users are indeed more connected with others in the greater area of their location, however they are also connected globally.

6.3 Tweet Location Discovery Results

QL Method

This experiment evaluates how topical *tweets* are, and whether we can utilize their content in order to identify their exact geographical locations. For the **QL** method we preprocessed the data, by removing stopwords and performing a stemming function (through lucene K-Stemmer stemmer), which slightly improved our accuracy across all configurations. We tested for linear interpolation, however because of the small size of the query (maximum 140 characters) the best results were yielded when $\lambda = 0$. For each experiment we performed a 10-fold cross validation, by splitting the dataset in 10 subsets randomly, and for each experiment training with the 9 parts and tested with the other. The reported figure is the average precision, which had a trivial deviation.

Results of the **QL** method, for $a = 4000$ reach a precision of 33% for Dublin, 34% for the city of Manchester and 27% for Boston, while they drop slightly as we reduce the size of the boxes. Figure 4(a) illustrates the accuracy for each city, for the tested granularities. We can observe, that although the number of possible boxes *quadruples* from each level of detail to the next, the accuracy of our method decreases in much slower fashion, even for the very specific granularity of 500m. It is also noticeable that Boston is always clearly lower than the other two cities, which can be attributed to the fact that the area we chose for Boston is almost four times the size of the area we chose for Manchester, as it is a much larger city. In addition, despite the fact that the area for Manchester is much smaller than the one for Dublin, they have very similar accuracies. This occurs due to the landscape of each city, since some of the area we selected for Dublin, inevitably contains a portion of sea, without any tweets. Table 2 contains more detailed information about the total number of tweets, the number of empty and reported squares, as well as the average Error Distance in kilometres for each city; for $a = 1000m$. Interestingly, the average distance is only slightly reduced, when precision is increased which indicates that our method often chooses neighbouring boxes for smaller granularities. Our method reports the majority of squares, which indicates that it is not affected by the skewed distribution. However this biased distribution cannot be exploited to locate tweets. In order to illustrate that, we created a method that assigns a tweet to a location with a probability proportional to the number of tweets in that box, which yielded an accuracy of less than 1%.

	Grid Size	Tweets	Empty	Reported	Error
Dublin	34 x 36	1099904	309	868	11.604
Manchester	21 x 22	798779	20	442	8.301
Boston	39 x 39	1014232	117	1521	14.822

Table 2: Details for QL experiments

6.3.1 *QLU Method*

In this experiment we evaluate whether *users* are topical, and tend to tweet from the same places, when they do so by their phones. Figure 4(b) presents the same information as the previous section, for the method **QLU** which yielded a precision of 80% for boxes of side 4000m in Dublin, which was reduced only by 10% when reducing the size of the reported area to 1/64 of the original. This result indicates that users, across cities, tweet in a spatial routine pattern, with a very big portion of their tweets from the same specified area.

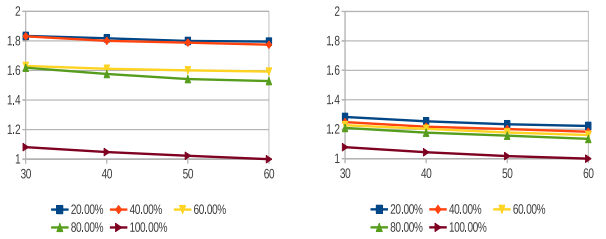
Table 4(c) shows the sensitivity of our method when using various percentages of the dataset as training. The upper lines represent the accuracy **QLU** method while the bottom ones represent **QL**. The latter is sensitive for tweet number, however **QLU** is not, which indicates the strength of people's habits to tweet from nearby locations.

6.4 Evaluation through Topic Models

This experiment illustrates how well we can summarize the information from all the tweets we crawled, as well as the effectiveness and necessity of our methods since we can extract more meaningful and accurate topic models through LDA [5]. Summarizing tweets through topics is important, especially in the case of emergency identification. Modelling the topics of discussion in a city, can aid towards defining which are usual topics of discussion in a city, and thus the ability to spot abnormal ones. Furthermore it can be used to identify the rate with which people change topics of dis-

cussion, and identify a city-wide emergency in case of an abrupt change, directions we are eager to investigate in the future.

For the city of Dublin, we initially used the geo-Located tweets we received from Twitter, and then those we were able to download from the new users we discovered with our method. Assuming that one wants to know what the major topics of discussion in a city C are, during a time interval t, we identify the difference between the two datasets. As a measure of performance of LDA, we use perplexity as defined in [27]. Perplexity essentially tells us how well our probability distributions over the topics, represent the testing set. Our results indicate that the set of GeoLocated Tweets is much more perplexed than the complete dataset, for various training percentages, across different total number of topics. Except for the case of the complete dataset being a training set, the rest of the samples are non inclusive.



(a) GeoLocated Tweets (b) Tweets from all Users

Figure 6: Perplexity v Number of Topics, for various percentages of the Dataset (lower is better)

Figure 6(a) shows the perplexity of the geo-Tweets when divided with the perplexity of the complete set and most topics which is the lowest. We can observe that the perplexity drops significantly when using a bigger dataset, which is of-course expected. However in the second case (Figure 6(b), which is the dataset with all the tweets, the perplexity is at all levels much lower and less sensitive. This indicates two things: Firstly, LDA creates better topic models given our more holistic dataset, and secondly that even with a small percentage of that data, we are able to create and retrieve a more coherent picture about the topics discussed in a city than just with geo-located tweets. In other words, the tweets which are geo-located are not enough to accurately depict what the topics of conversation in a city are.

We also performed a qualitative test, by manually labelling the topics and determining how many can potentially be annotated with a coherent topic of discussion. We labelled the 20% and 100% of both datasets. After considering the results from the largest dataset as ground truth, since it contains all other datasets, we evaluated how many topics from each dataset are relevant to the this. The results are in Table 3 . We present some indicative topics and annotations for each case in Table 4.

	Relevant	Annotated
20% Geo	10	32
100% Geo	18	37
20% All	27	42
100% All	45	45

Table 3: Relevant and Annotated Topics

7. CONCLUSIONS

Identifying information for a specific location is an important problem. In this regard we analysed the structure of social networks for users that live in proximity and concluded that there is a correlation between strong connectivity in the social graph and proximity in the real world. We created a method that captures the dynamic relations on such a graph, and can locate users who live in a specific area, optimized for minimum number of queries in the graph. Furthermore we created a method for precise geo-location of tweets within a city with high accuracy and provided extensive experimentation on a real social network, regarding the effectiveness of our method as well as the quantitative and qualitative benefit from the newly found data.

8. ACKNOWLEDGMENTS

This work has been co-financed by EU and Greek National funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Programs: THALIS - GeomComp, THALIS - DISFER, ARISTEIA - MMD" and the EU FP7 funded project INSIGHT (www.insight-ict.eu).

The authors would like to thank Myrto Vlazaki and Ioannis Katakis for their their valuable insights and recommendations

9. REFERENCES

[1] A. Agarwal, K. Hosanagar, and M. D. Smith. Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets. *Journal of Marketing Research (JMR)*, 48(6):1057–1073, 2011.

[2] A. Ahmed, L. Hong, and A. J. Smola. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 25–36, 2013.

[3] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *WebSci*, pages 33–42, 2012.

[4] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 61–70, 2010.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] H. S. Burton, W. K. Tanner, G. C. Giraud-Carrier, H. J. West, and D. M. Barnes. Right time, right place health communication on twitter: Value and accuracy of location information. *J Med Internet Res*, 14(6):156, November 2012.

[7] 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*, CIKM '10, pages 759–768, 2010.

[8] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social

20% Geo	abortion	ireland women abortion vote france vinb law issue politics country bill pro europe reading point government tax people bank
20% Geo	tourism	dublin ireland bar park photo guin dub street post tha cliath howth aerfort st bhaile square hotel temple green
20% Geo	football	united mufc great castle de enjoy day pal weekend love mate massive yea matchday stanakatic caskett hun
20% Geo	rugby	lion leinster rugby ihl fitness lions coybib gatland trx bod fitfam pm ym seaofre final match anglotape fella heinekencup
20% Geo	?	donegal christina tom bloomsday roscommon fe casa kenny patrick theapprentice hill floor princess rackard ding alley min
100% Geo	places	dublin ireland bar guin dub tha cliath aerfort bhaile pic hotel airport temple storehouse pub st pint trinity college
100% Geo	football: uk	player united season game play arsenal football sign team haha league fan goal win suarez man mate mufc chelsea
100% Geo	music show	check play rt music gig album video film listen show eventsindublin live song awesome watch band festival game cat
100% Geo	rugby	lion game great win match play final gaa leinster team rugby golf ireland today wimbledon player murray congrats dub
100% Geo	?	watch man life show call god car made hit kid years laugh mr tom men face jesus dream break
20% all	irish jobs	business ireland job dublin social jobfairy marketing digital media hire great tip irishjob startup network tech sales online company
20% all	Politics:EC	ireland bank tax eu people europe pay report government uk protest minister year state home court news service job
20% all	food	food restaurant wine lunch taste special beer coffee free cocktail dinner menu delicious chef recipe bar eat lovely yum
20% all	abortion	abortion vinb life women ireland bill vote seanad pro people baby politics law party dail fg debate prolife labour
20% all	?	ur ya ye im goin yea dont wat tho pal ha gettin il wit ill ah nite jus bout
100% all	Politics: EC	ireland bank tax news eu britain report uk government police protest europe belfast human gold syria attack year minister
100% all	research	ireland great today health support student research eu event conference school day people children education week work launch europe
100% all	abortion	abortion ireland vinb women vote life people bill seanad law party support politics pro dail debate woman gay marriage
100% all	irish jobs	job dublin ireland business jobfairy irishbizparty manager hire sales recruit cork company bizhub engineer developer senior service client
100% all	?	ur ya ye ha goin haha wat tho gettin il wit im dat de jus day yea nite

Table 4: Topics from the LDA for each dataset, and their annotation

networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1082–1090, 2011.

[9] 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, pages 1277–1287, 2010.

[10] G. Fuchs, G. Andrienko, N. Andrienko, and P. Jankowski. Extracting personal behavioral patterns from geo-referenced tweets. 2013.

[11] 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, 2011.

[12] R. Holt. Twitter in numbers, Mar. 2013. http://www.telegraph.co.uk.

[13] 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*, WWW '12, pages 769–778, New York, NY, USA, 2012. ACM.

[14] Y. Ikawa, M. Enoki, and M. Tatsubori. Location inference using microblog messages. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 687–690, 2012.

[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*, SMUC '11, pages 61–68, 2011.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[17] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR*, SIGIR '01, pages 111–119, 2001.

[18] T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras. On the spatiotemporal burstiness of terms. *Proc. VLDB Endow.*, 5(9):836–847, May 2012.

[19] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW*, pages 695–704, 2008.

[20] 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*, KDD '12, pages 1023–1031, 2012.

[21] W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson. The where in the tweet. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2473–2476, 2011.

[22] J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? inferring home locations of twitter users. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[23] S. Murray. Twitter breaks news of whitney houston death 27 minutes before press, Feb. 2012. http://mashable.com.

[24] K. Ren, S. Zhang, and H. Lin. Where are you settling down: Geo-locating twitter users based on tweets and social networks. In *Information Retrieval Technology*, volume 7675 of *Lecture Notes in Computer Science*, pages 150–161. 2012.

[25] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, 2010.

[26] G. Valkanas and D. Gunopulos. Location extraction from social networks with commodity software and online data. In *ICDM Workshops*, pages 827–834, 2012.

[27] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, 2009.

[28] S. Yardi and D. Boyd. Tweeting from the town square: Measuring geographic local networks. In *ICWSM*, 2010.