H. Efstathiades, D. Antoniades, G. Pallis, M. D. Dikaiakos

IDENTIFICATION OF KEY LOCATIONS BASED ON ONLINE SOCIAL NETWORK ACTIVITY
Motivation

• Key Locations information is of **high** importance for various fields

• Potentials for
  – Understanding users’ movement
  – Influence of location in social structure
  – Design social network architectures
  – Transportation patterns analysis
  – Etc.

• Can be used in combination with Open Data
Motivation

• Only a small number of users share such information in OSN profiles

• Majority in relatively high granularities
  • Country level
  • State level
  • City level
• Is it possible to infer a user’s **Home** and **Workplace** locations simply by observing the locations and time the user tweeted from?

  – We present a methodology which infers users key locations at **post-code** level
    • With the use of geo-tagged Twitter data
    • Evaluation on 3 distinct geographical regions
      – Outperforms different studies in cases of granularity and accuracy
      – Compare and validate our results with open-data
Related Work

• Identification of users’ locations from OSN is in high interest for researchers

• Approaches:
  – Content-based
    • Analysis of the text that users publish
    • Their accuracy is at most 57% for 10Km granularity
  – Geo-tagged based
    • Based on geographical info (latitude, longitude)
    • Mainly for “ground truth” construction regarding home locations (Assumed to have 100% accuracy)
DATASETS
Datasets

• We construct two different datasets
  
  – Home Location Identification

  – Workplace Location Identification
HOME LOCATION
Home Location

• We collect the Twitter Stream from 3 different areas:
  – The Netherlands
  – London, UK
  – Los Angeles County, US

• Collected users act as seeders

• Randomly collect users for whom they have reciprocal relationship
  – Filter out non-individual users
Home Location

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Users</th>
<th>Tweets</th>
<th>Geo-tagged Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>TW-NL</td>
<td>Netherlands</td>
<td>702,593</td>
<td>668,684,891</td>
<td>16,445,151</td>
</tr>
<tr>
<td>TW-LA</td>
<td>LA County</td>
<td>350,637</td>
<td>532,738,302</td>
<td>35,645,531</td>
</tr>
<tr>
<td>TW-LO</td>
<td>London</td>
<td>182,272</td>
<td>232,331,077</td>
<td>35,406,092</td>
</tr>
</tbody>
</table>

**TABLE I.** Home location dataset: number of users, number of tweets and geo-tagged tweets, for each of 3 regions of the resulted dataset.

<table>
<thead>
<tr>
<th>Name</th>
<th>Post-code areas</th>
<th>Average area radius (Km)</th>
<th>Ground Truth Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>TW-NL</td>
<td>286</td>
<td>2.68</td>
<td>1414</td>
</tr>
<tr>
<td>TW-LA</td>
<td>62</td>
<td>2.75</td>
<td>370</td>
</tr>
<tr>
<td>TW-LO</td>
<td>151</td>
<td>2.37</td>
<td>760</td>
</tr>
</tbody>
</table>

**TABLE II.** Home location dataset: number of post-code areas and average area radius in Km, for each of 3 regions of the resulted dataset.

**Ground truth users:** Users who report their exact coordinates (latitude, longitude) or post-code location

- Users: ~1 million
- Tweets: ~1.5 billion
- Geo-tagged: ~6%
WORKPLACE LOCATION
Workplace Location

- Work location is not usually clearly stated by a Twitter user in her personal profile
  - Profiles are used for a completely different purpose than career-related tools

- LinkedIn
  - a professional social network
  - users publish career related information
    - Including the place they work
      - City level
Workplace Location

• Listen to the public stream of *Friendfeed* for 1 week
  – Aggregator tool
  – Resulted to ~20,000 users

• Retrieve users who have both
  – [LinkedIn](https://www.linkedin.com)
  – [Twitter](https://twitter.com)
  – ~3000 users
Workplace Location

• Problem: Company’s reported location is the headquarters location

• Pre-processing analysis for aggregated profiles
  • Identify the exact branch of the company/employer at post-code level
  • Identify geo-location information for the workplace of 317 different users from different countries
# Workplace Location

<table>
<thead>
<tr>
<th>Name</th>
<th>Users</th>
<th>Tweets</th>
<th>Geo-tagged Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>TW-LinkedIn-Work</td>
<td>317</td>
<td>915,933</td>
<td>73,003</td>
</tr>
</tbody>
</table>

**TABLE III.** Workplace location dataset: Number of users, number of tweets and geo-tagged tweets.

<table>
<thead>
<tr>
<th>Country of origin</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>34.7</td>
</tr>
<tr>
<td>Great Britain</td>
<td>11.3</td>
</tr>
<tr>
<td>Italy</td>
<td>5.7</td>
</tr>
<tr>
<td>Spain</td>
<td>5.1</td>
</tr>
<tr>
<td>Canada, France, Turkey</td>
<td>4.7 (each)</td>
</tr>
<tr>
<td>Other(23)</td>
<td>29.1</td>
</tr>
</tbody>
</table>

**TABLE IV.** Workplace location dataset: Demographic characterisation

<table>
<thead>
<tr>
<th>Industry</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>21.8</td>
</tr>
<tr>
<td>Information Technology</td>
<td>16.4</td>
</tr>
<tr>
<td>Marketing and Advertising</td>
<td>11.7</td>
</tr>
<tr>
<td>Computer Software</td>
<td>8.2</td>
</tr>
<tr>
<td>Online Media</td>
<td>7.6</td>
</tr>
<tr>
<td>Other(51)</td>
<td>34.3</td>
</tr>
</tbody>
</table>
Time-Frame Clustering methodology

**USERS KEY LOCATIONS**
Hypothesis

• Users tend to spend a significant, but distinct, amount of their time during an average day in two key locations of interest; Home and Workplace locations.

• These two locations are much more likely to appear in the user’s geo-tagged activity during specific timeframes, than locations that are not so frequent in users routine.
Observations

We expect that the user will mostly be posting tweets from a single location during Rest and Active.
Observations

Tweeting rate distribution from **home** on an hourly basis. Y-axis represents the portion of total Tweets that have been produced during a specific hour.

Probability of tweeting from **Home** tends to increase significantly during (and close to) the **Rest** timeframe.
Observations

Tweeting rate distribution from **workplace** on an hourly basis. Y-axis represents the portion of total Tweets that have been produced during a specific hour.

Probability of tweeting from **Work** tends to increase significantly during (and close to) the **Active** timeframe.
Observations

[Bar chart showing the distribution of leisure, rest, active, and work activities over 24 hours.]
Observations

Number of different locations from which user tweet during **Active** and **Leisure** hours.

90% of the cases the user will post at max from a handful of locations during **Active** timeframe.
Proposed Methodology

– Each Tweet has a different weight based on:
  • Time that has been tweeted
  • Location that we aim to extract

– Each day has a unique weight:
  • To avoid cases of frequently tweeted places
    – Concerts
    – Sports events etc.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Rest</th>
<th>Leisure</th>
</tr>
</thead>
<tbody>
<tr>
<td>TW-NL</td>
<td>0.744</td>
<td>0.362</td>
</tr>
<tr>
<td>TW-LA</td>
<td>0.735</td>
<td>0.357</td>
</tr>
<tr>
<td>TW-LO</td>
<td>0.737</td>
<td>0.354</td>
</tr>
</tbody>
</table>

**TABLE V.** Probability of tweeting from Home during Rest and Leisure timeframes for the 3 different datasets.
EVALUATION
Evaluation Scenario

- Identify users’ **home** and **workplace** locations
  - Granularity: **Post-code**
  - **Weight timeframes** based on observations (e.g. 0.73 rest, 0.35 leisure)

- **Ground truth**
  - Users who **report their exact location** (lat,lon) or post-code in Twitter location field
  - Users whom workplace post-code location **has** been inferred

- Comparison with approaches that are used to construct **ground truth**
Evaluation – Pre-processing

• Home identification
  – Eliminate common well known locations – POI
    • Attractions
    • Hotels, restaurants, bars etc.
    • Landmarks

• Bring all geo-tagged information to a common format
  – post-code granularity
Evaluation – Pre-processing

- Bring all geo-tagged information to a common format
  - Use a geo-coding API to retrieve boundaries of each post-code area
  - Map user’s who report exact location in corresponding area
Evaluation - Metrics

- **ACC - Accuracy**: gives the percentage of correctly inferred users’ key locations over the total sample size [1, 2, 3]

- **ACC@R - Accuracy within radius (R)**: gives the percentage of correctly inferred users’ key locations identified within R Km from users reported locations [1, 2, 3]

- **AED - Average Error Distance**: defines the distance, in Km, between the inferred location (center of the post-code in our case) and user’s reported location [1, 3]

1. S. Katragadda, M. Jin, and V. Raghavan. An unsupervised approach to identify location based on the content of user’s tweet history. In *Active Media Technology* 2014
3. K. Ryoo and S. Moon. Inferring twitter user locations with 10 km accuracy. *WWW’14*
Evaluation - Methods

- **MP - Most Popular** marks as home location the most popular location, in number of geo-tagged tweets, visited by the user. [4]

- **MC - Median Clustering** marks the user’s home location by calculating the median value of locations the user tweeted from. [5]

- **TF-C – Time-Frame Clustering** is the method presented in our paper.

---

5. K. Ryoo and S. Moon. Inferring twitter user locations with 10 km accuracy. *WWW’14*
Evaluation - Results

• On ground truth data

<table>
<thead>
<tr>
<th>Method</th>
<th>TW-NL</th>
<th>TW-LO</th>
<th>TW-LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MP</td>
<td>0.69</td>
<td>0.47</td>
<td>0.55</td>
</tr>
<tr>
<td>MC</td>
<td>0.67</td>
<td>0.19</td>
<td>0.39</td>
</tr>
<tr>
<td>TF-C</td>
<td>0.81</td>
<td>0.68</td>
<td>0.701</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>AED</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>3.21</td>
</tr>
<tr>
<td>MC</td>
<td>3.93</td>
</tr>
<tr>
<td>TF-C</td>
<td>2.77</td>
</tr>
</tbody>
</table>

**TABLE VI.** Home-Location identification performance measured in accuracy (ACC) and Average Error Distance (AED) in Km, for 3 different approaches in 3 different areas.
Evaluation - Results

• On ground truth data
• How many Tweets does TF-C require?

![Graph showing performance vs. number of tweets]

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

- Workplace identification

Proposed methodology is able to identify the exact workplace location at post-code granularity with 63% accuracy and ~80% at a 10km granularity.
Evaluation - Results

- Home and Workplace: On open-data

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

- 84% of the areas the predicted and real post-code population rate differ only by 0.005.
- 85% of post-code areas the predicted and real employees rate differ by less than 0.005, while only 5% differs by more than 0.01.
Evaluation - Discussion

• We can detect a user’s home location in a radius smaller than 10Km in most of the cases

• **MP** and **MC**
  – both methods used to provide ground truth data
  – low detection accuracy, between 20 and 70%  

• We can provide a more accurate ground truth
  – Help improve the methods themselves
  – and their detection accuracy
Evaluation - Discussion

• Workplace location identification

  – 80% for identification of user workplace in a 10Km proximity.

  – First study which constructs a dataset and performs analysis on workplace locations using Twitter
FUTURE WORK
Future Work

• Link users’ location with open data

• Investigate research questions:
  – How the socio-economic characteristics of an area influence the social graph?
  – How the locations visited by the user affect her social network connections?
  – How the user transports derived by Twitter data can be used to support city planning procedures?
Future Work

• We construct weighted graphs of areas
  – Mobility graphs

• Each link denotes a mobility relation between habitants of an area
  – Mobility could be defined: Habitants moved from area A to area B

• Weight: percentage of source vertex habitants who travel to destination vertex
Future Work

• Are we able to identify events based on anomalies detection on mobility graphs?
CONCLUSIONS
Conclusions

• Present problem of users location identification from OSN
• Present a study on Tweeting activity and users key locations
• Propose a methodology for inferring users key locations
  – uses geo-tagged twitter data
• Evaluation on 3 distinct geographical regions
  – Outperforms different studies in cases of granularity and accuracy
Thank You!
RELATED WORK
Geo-tagged based

• Ground truth construction:
  – MP: Most popular location regarding geo-tagged tweets marked as user’s location [2] [4]
  – MC: Pair of (median(latitude),median(longitude)) marked as user’s home [1][3]
  – **Accuracy**: Hypothesized to be 100%

• A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM ’12*
  – use geo-tagged information of their ego network
  – need for at least 2 geo-active friends
  – needs at least 100 geo-tagged tweets for a one month period, from the user’s friends
  – **Accuracy**: 62%
  – Ground truth: MP

---

4. R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. Understanding Human Mobility from Twitter. 2014
Content-based

  – Use a location dictionary for places all over the United States.
  – Accuracy: 57% at city level
  – Ground truth: MP

  – Probabilistic model to assign location data to popular words in Twitter
  – Use words’ popularity to identify the location of the users that tweet them
  – Accuracy: 57% at 10Km radius.
  – Ground truth: MC
Dataset Collector

- Collects data from Twitter
- **Given as input a list of user_ids or screen_names:**
  - **Global Workload** is distributed based on the number of Local Distributors
  - **Local Workload** is distributed in different instances based on availability of local resources
  - **Each instance** is able to run forever as monitoring service adds or removes resources based on instance needs
    - **Throughput:** 3000 – 3200 users/hour per Local Distributor
# Dataset Description

<table>
<thead>
<tr>
<th>Location</th>
<th>Users</th>
<th>Tweets</th>
<th>Geo-tagged Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands</td>
<td>702,593</td>
<td>668,684,891</td>
<td>16,445,151</td>
</tr>
<tr>
<td>LA County</td>
<td>350,637</td>
<td>532,738,302</td>
<td>35,645,531</td>
</tr>
<tr>
<td>London</td>
<td>182,272</td>
<td>232,331,077</td>
<td>35,406,092</td>
</tr>
</tbody>
</table>

Table 1: Number of users, number of Tweets and geo-tagged Tweets, for each of 3 regions of the resulted dataset.

<table>
<thead>
<tr>
<th>Location</th>
<th>Post-code areas</th>
<th>Average area radius (Km)</th>
<th>Ground Truth Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands</td>
<td>286</td>
<td>2.68</td>
<td>1414</td>
</tr>
<tr>
<td>LA County</td>
<td>62</td>
<td>2.75</td>
<td>370</td>
</tr>
<tr>
<td>London</td>
<td>151</td>
<td>2.37</td>
<td>760</td>
</tr>
</tbody>
</table>

Table 2: Number of post-code areas and average area radius in Km, for each of 3 regions of the resulted dataset.
Early results

• **Zwolle** is a municipality and the capital city of the province of Overijssel, Netherlands [Wikipedia]
  - Population: about 125,000
  - Its habitants are mostly locals

• **Amstelveen** is a municipality in the province of North Holland, Netherlands [Wikipedia]
  - Population: about 85,000
  - A large percentage of its habitants are students, as VU Amsterdam is located in this area
Early results

<table>
<thead>
<tr>
<th>Habitants Leisure Areas</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zwolle</strong></td>
</tr>
<tr>
<td>ABROAD, LEISURE AREAS IN ABROAD, SCHIPHOL INTL</td>
</tr>
<tr>
<td>UTRECT, LEISURE AREAS IN AIRPORT, HOLLAND SPORT AMSTERDAM</td>
</tr>
<tr>
<td><strong>Amstelveen</strong></td>
</tr>
<tr>
<td>BOAT CENTER</td>
</tr>
</tbody>
</table>

Hariton Efstathiades, h.efstathiades@cs.ucy.ac.cy
Early results

- Leisure weekdays
- Leisure weekends
- Work place

• People tend to
  • live close to their leisure places or vice versa. (Similar behavior identified by P. Georgiev and A. Noulas, 2014)
  • Not so close to their workplace

*Netherlands
KING’S DAY
AMSTERDAM
...the craziest day of the year!!!
Mobility Graphs

• Are we able to identify events based on anomalies detection on mobility graphs?

• So far:
  – Constructed mobility graphs daily snap shots for users from Netherlands
  – Collect Facebook events for the same period