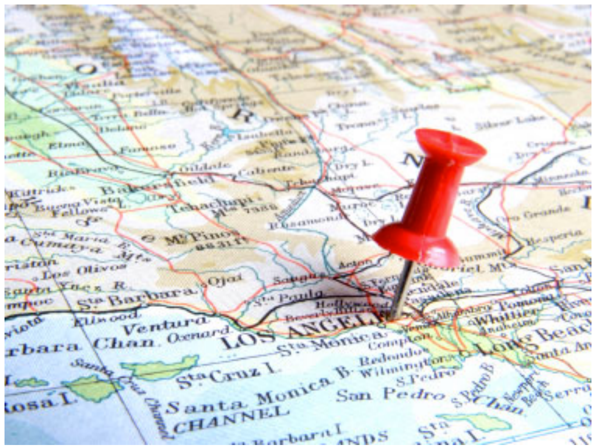
H. Efstathiades, D. Antoniades, G. Pallis, M. D. Dikaiakos

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 u* ***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

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

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

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

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

- Users: ~1 million
- Tweets: ~1.5 billion
- Geo-tagged: ~6%

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

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
  - twitter
  - ~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

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

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

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

TABLE IV. WORKPLACE LOCATION DATASET: DEMOGRAPHIC CHARACTERISATION

Time-Frame Clustering methodology

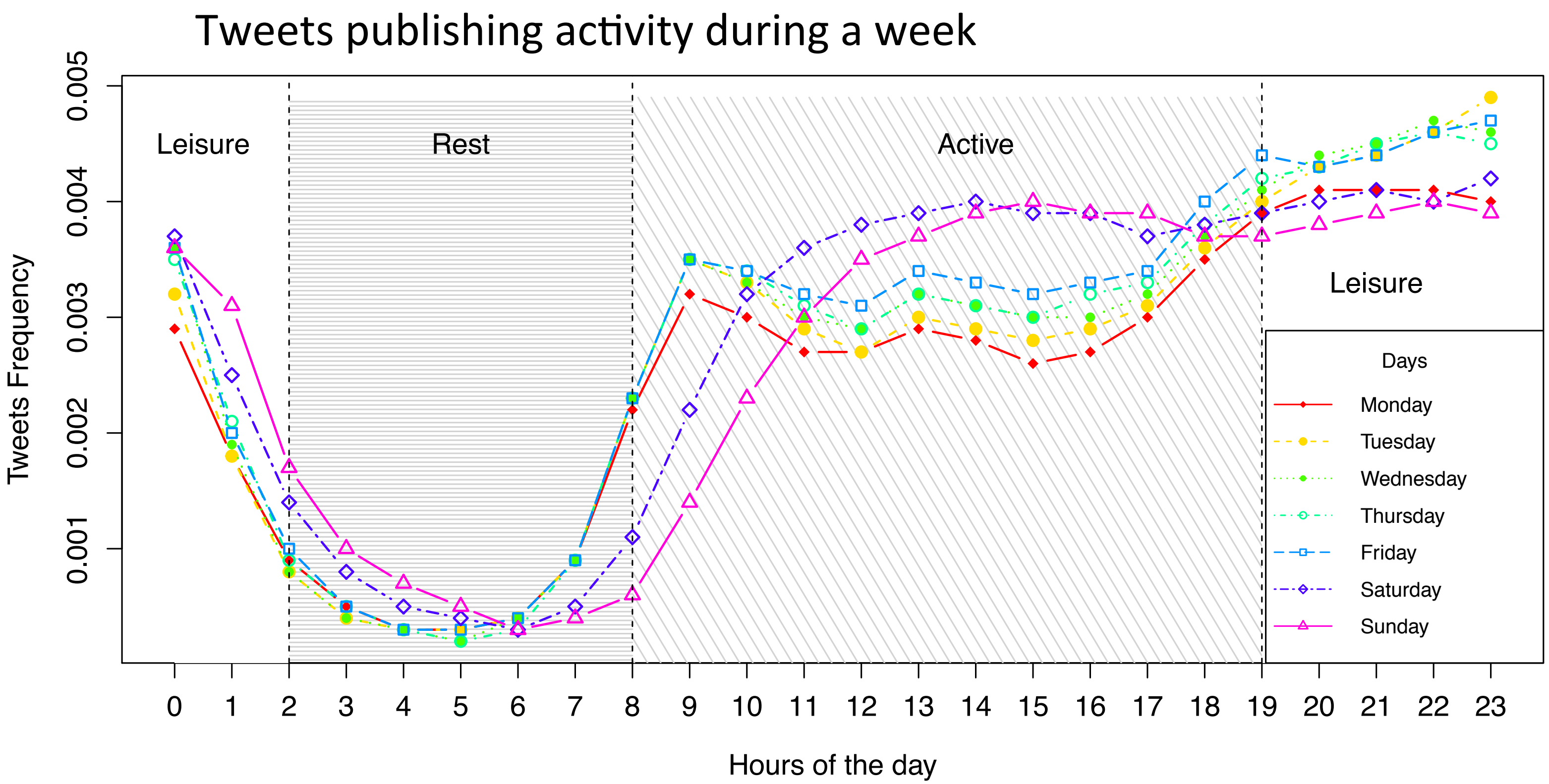
USERS KEY LOCATIONS

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.

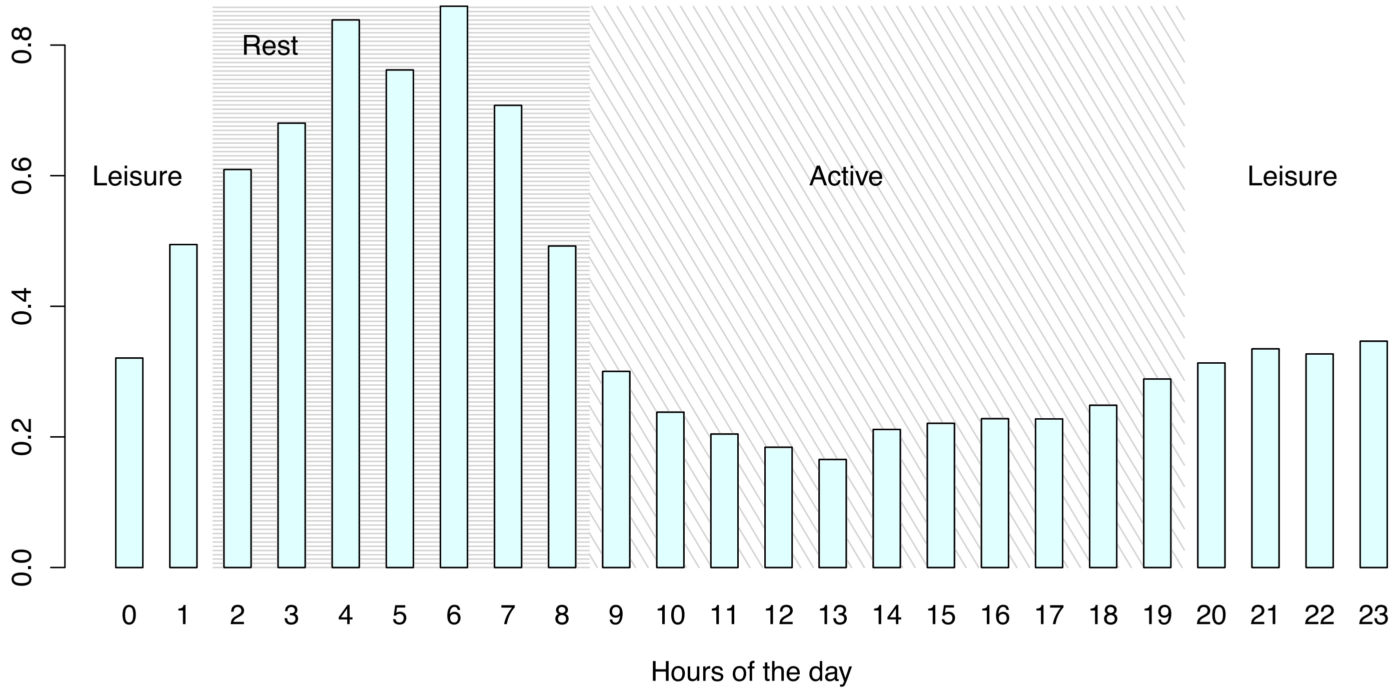
Observations



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

Observations

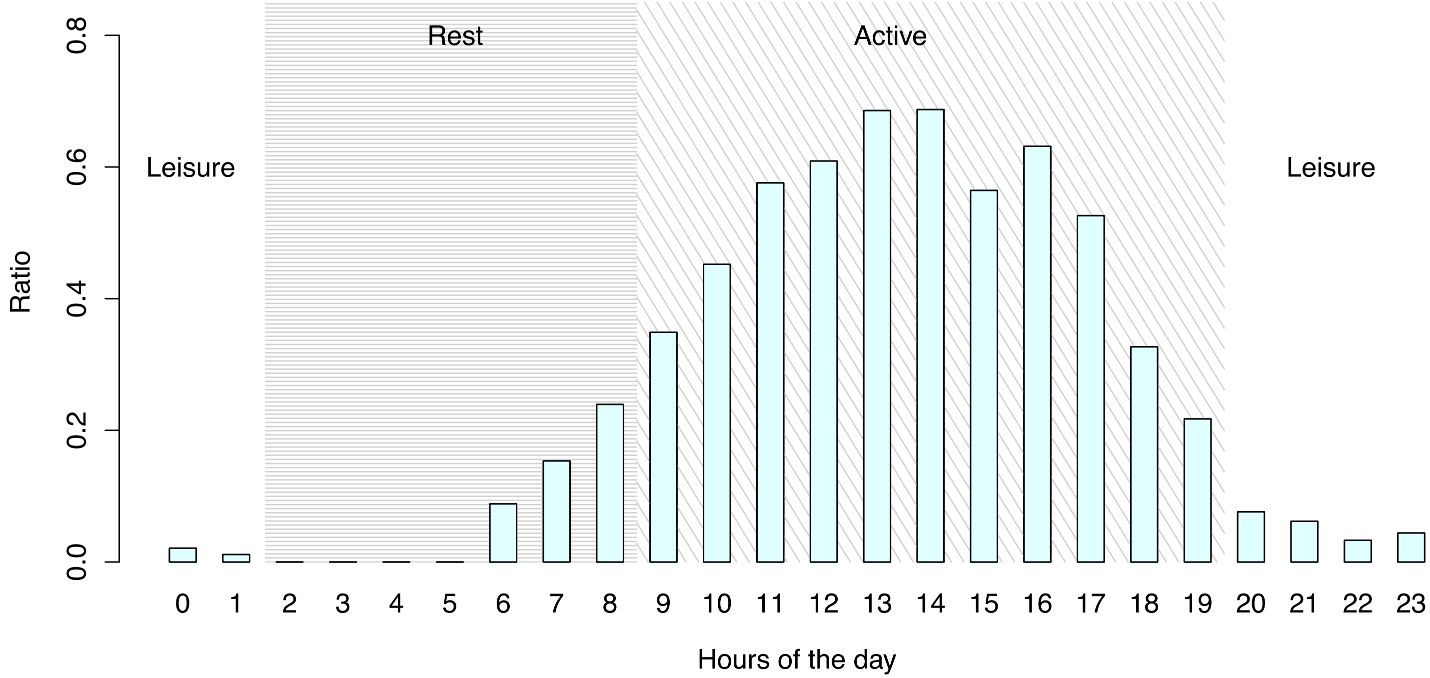
Tweeting rate distribution from **home** on an hourly basis. Y-axis represents the portion of total Tweets that have been produced during a specific hour.



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

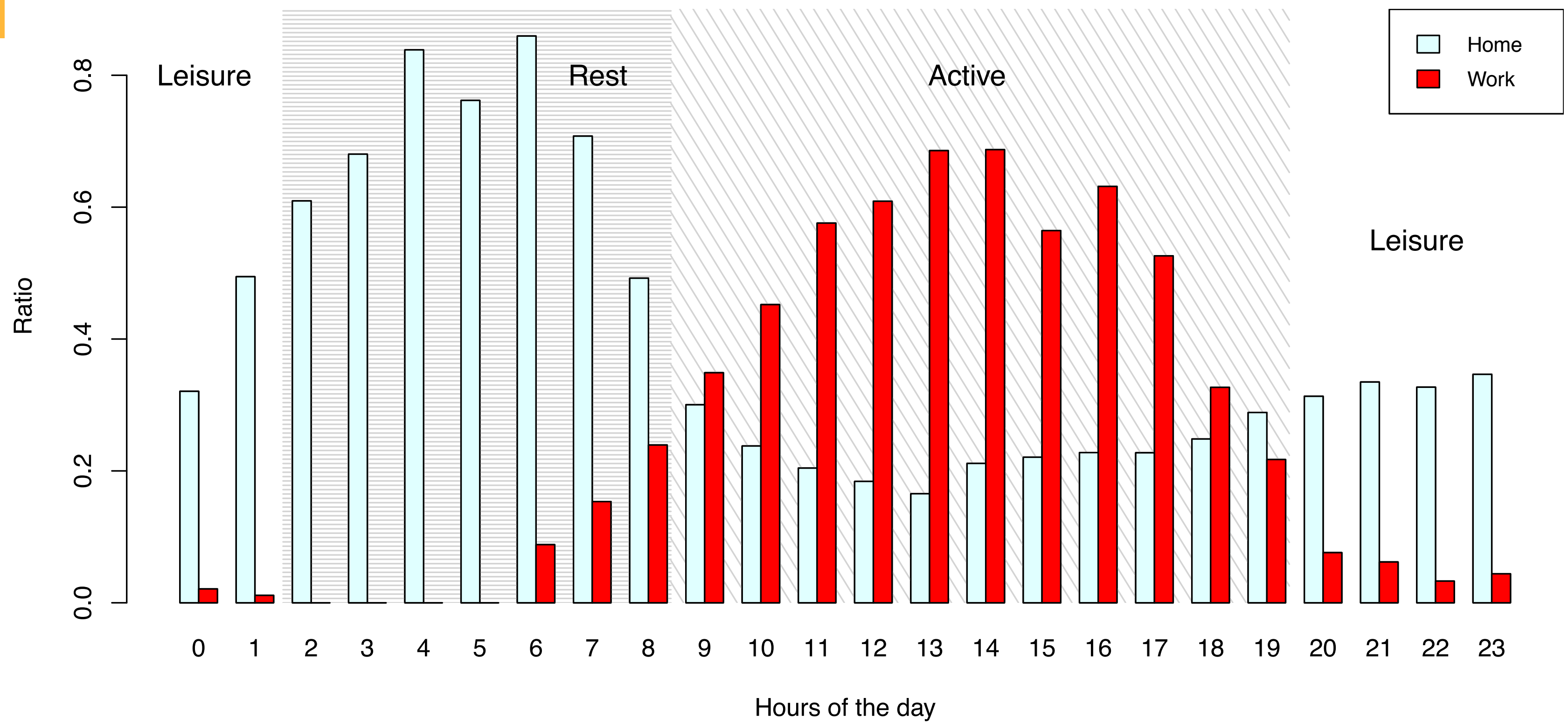
Observations

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.



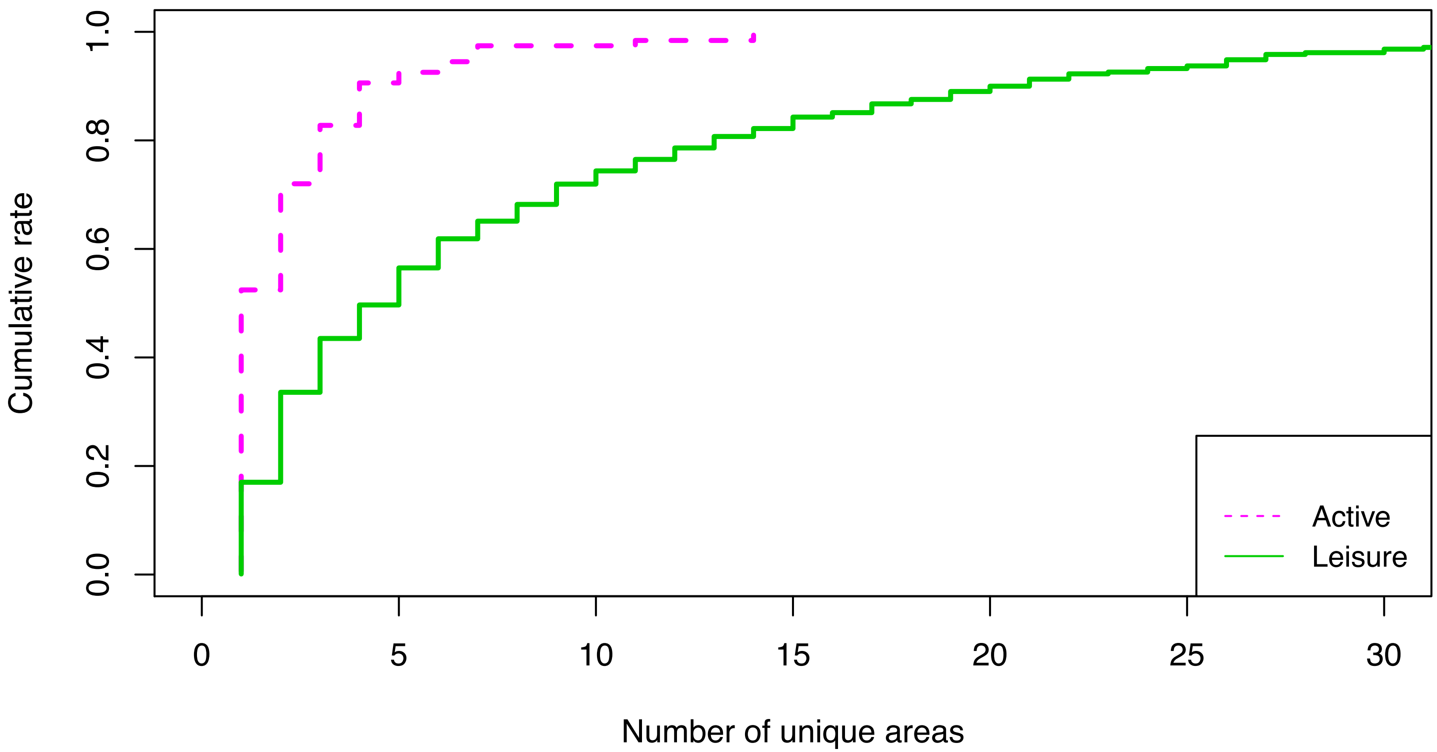
Probability of tweeting from **Work** tends to increase significantly during (and close to) the *Active* timeframe.

Observations



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.

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

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

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
2. J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *ACM Trans. Intell. Syst. Technol. 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.

4. P. Georgiev, A. Noulas, and C. Mascolo. The call of the crowd: Event participation in location-based social services. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 2014
5. K. Ryoo and S. Moon. Inferring twitter user locations with 10 km accuracy. WWW'14

Evaluation - Results

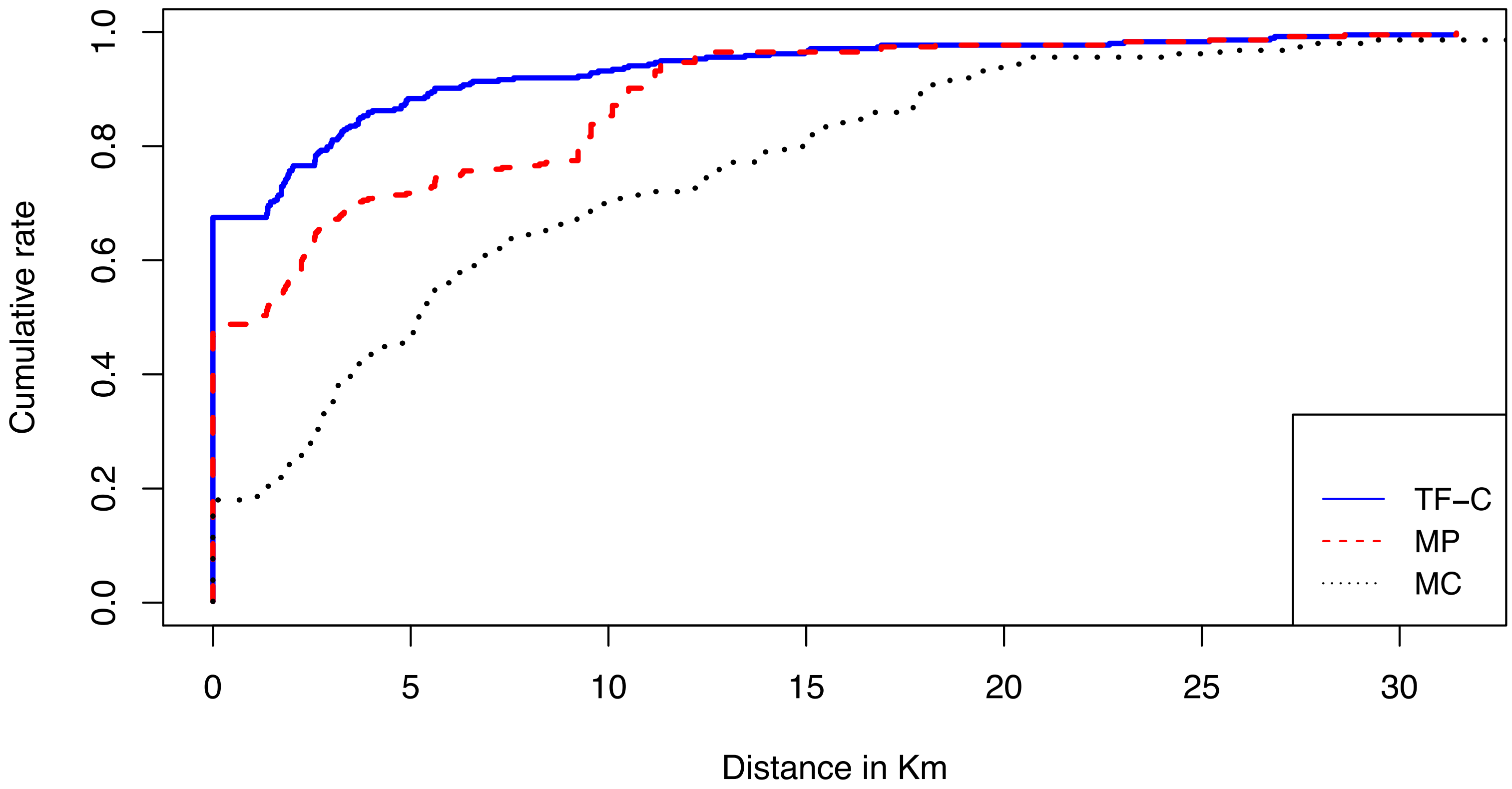
- On ground truth data

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

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

Evaluation - Results

- On ground truth data



- How many Tweets does TF-C require?

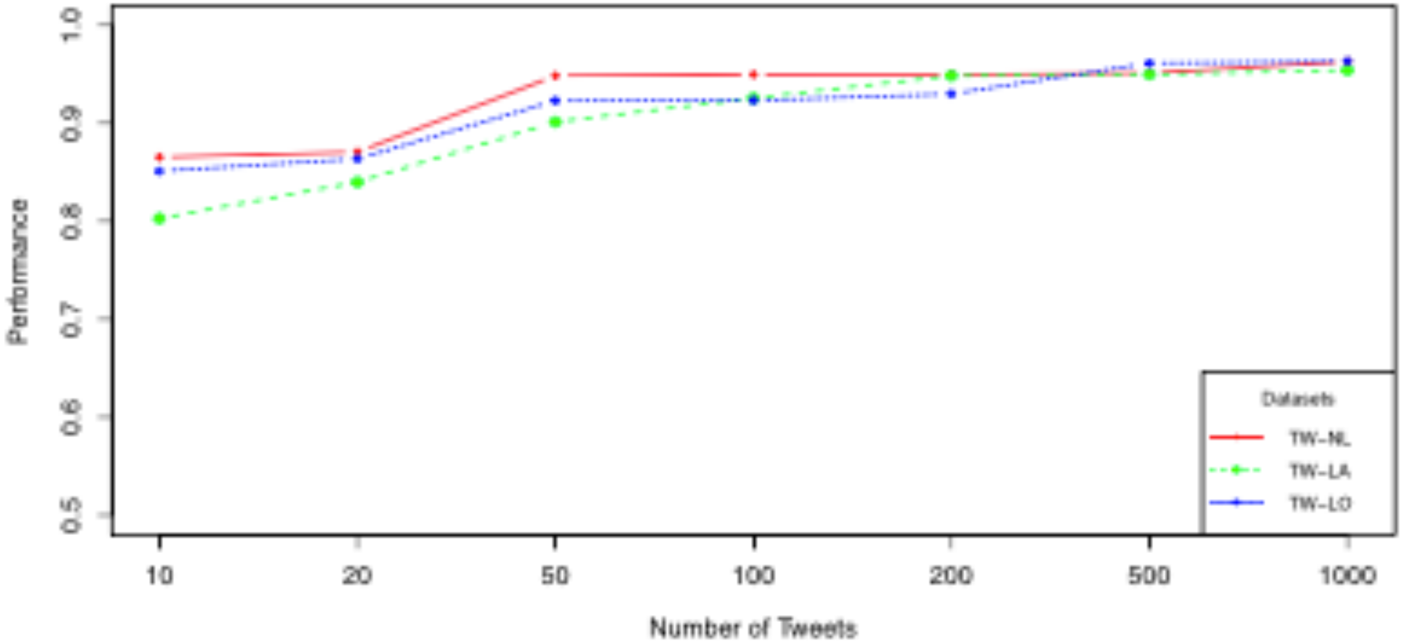
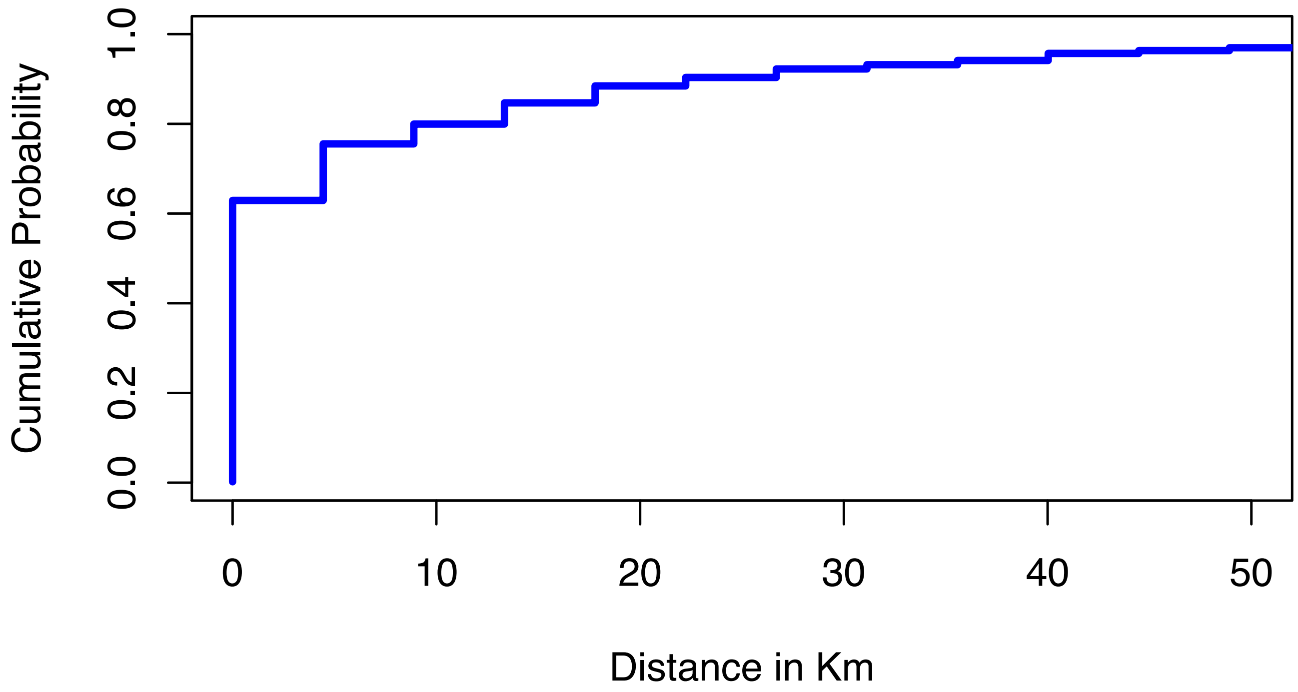


Fig. 5. Performance of proposed method in contrast to the number of recent tweets for the 3 datasets.

Evaluation - Results

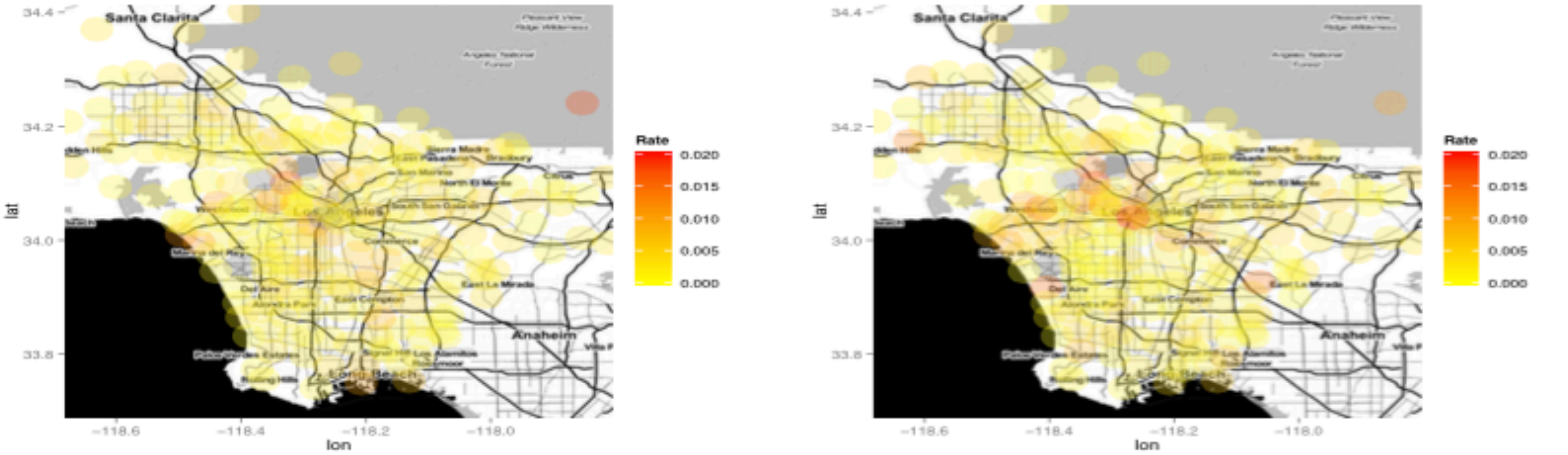
- Workplace identification



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



(a) Differences between real and predicted population rate.

(b) Differences between real and predicted employees rate.

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

- 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%

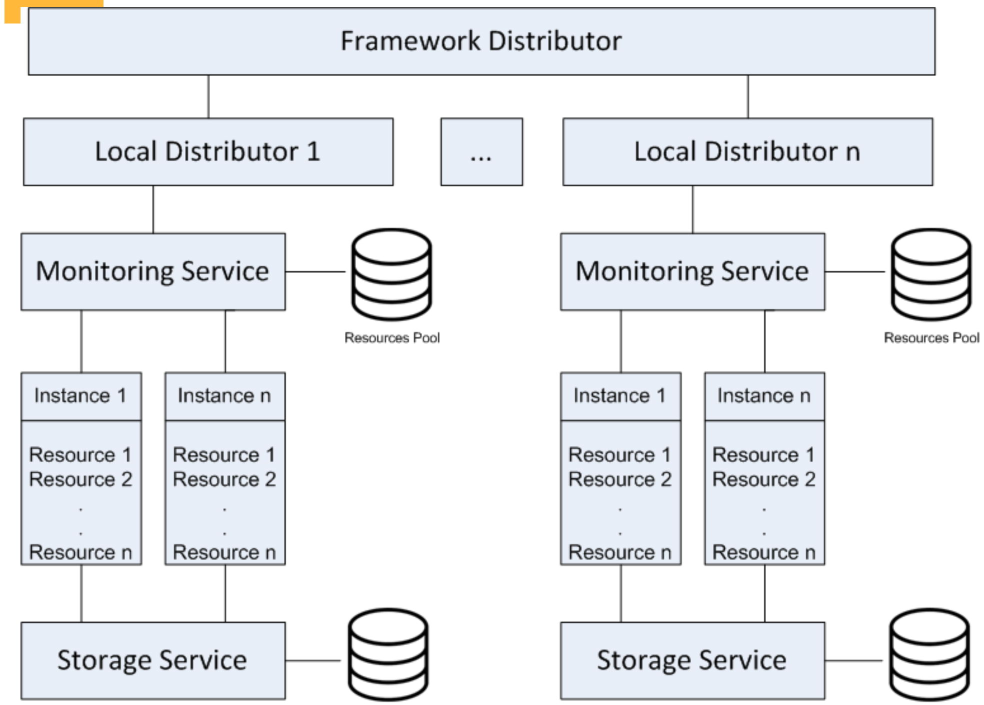
- A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12
  - use geo- tagged information of their ego network
  - need for at least 2 geo-active friends
  - needs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

1. E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. KDD'11
2. P. Georgiev, A. Noulas, and C. Mascolo. The call of the crowd: Event participation in location-based social services. ICWSM 2014.
3. B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky,P. Kazakopoulos, and C. Ratti. Geo-located twitter as proxy for global mobility patterns.
4. R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. Understanding Human Mobility from Twitter. 2014

Content-based

- J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *ACM Trans. Intell. Syst. Technol.*, 5(3):47:1–47:21, July 2014.
  - Use a location dictionary for places all over the United States.
  - Accuracy: 57% at city level
  - Ground truth: MP
- K. Ryoo and S. Moon. Inferring twitter user locations with 10 km accuracy. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web* WWW'14
  - 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

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

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

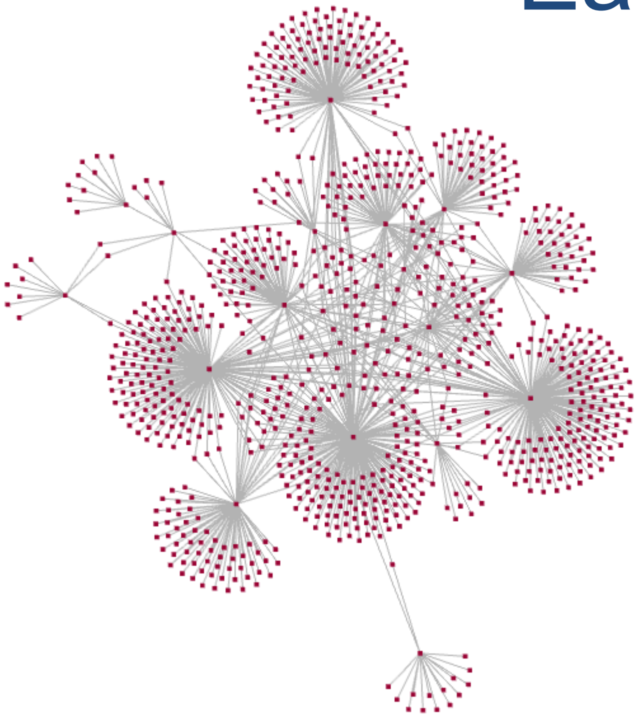
Location	Post-code areas	Average area radius (Km)	Ground Truth Users
Netherlands	286	2,68	1414
LA County	62	2,75	370
London	151	2,37	760

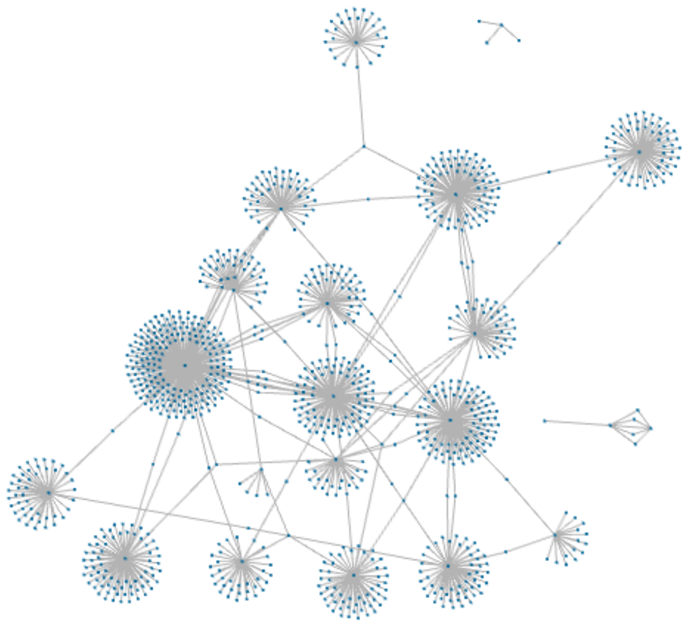
Table 2: Number of post-code areas and average area radius in Km, for each of 3 regions of the resulted dataset.

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

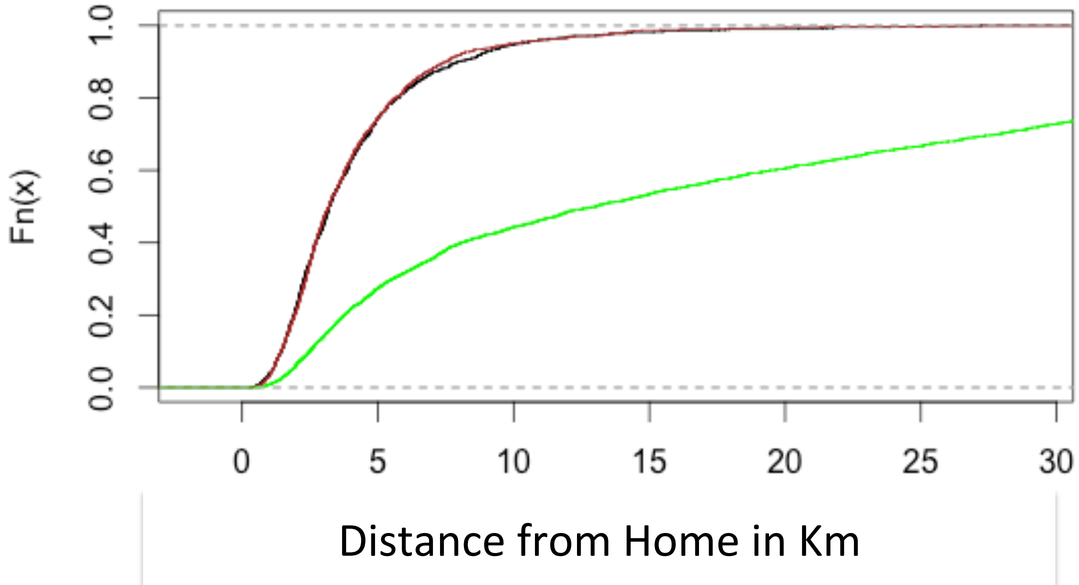
Early results





Habitants Leisure Areas	
Zwolle	Amstelveen
ABROAD, LEISURE AREAS IN UTRECT, LEISURE AREAS IN AMSTERDAM	ABROAD, SCHIPHOL INTL AIRPORT, HOLLAND SPORT BOAT CENTER

Early results



*Netherlands

- 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


KING'S DAY
AMSTERDAM
...the craziest day of the year!!!
KING'S DAY AMSTERDAM
Working Area
Amsterdam-Oost
Oud-Zuid
Watergraafsmeer
Zeeburg
De Pijp

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