Privacy-Friendly Mobility Analytics using Aggregate Location Data

ACM SIGSPATIAL 2016

Apostolos Pyrgelis, Emiliano De Cristofaro, Gordon Ross
University College London

November 3, 2016

Motivation

Motivation

- Mobility analytics are useful in modern cities - journey planning, congestion prevention, improving transportation service levels

Motivation

- Mobility analytics are useful in modern cities - journey planning, congestion prevention, improving transportation service levels
- But, large scale collection of individual users' location data raises privacy concerns (life-style, political / religious inclinations)

Motivation

- Mobility analytics are useful in modern cities - journey planning, congestion prevention, improving transportation service levels
- But, large scale collection of individual users' location data raises privacy concerns (life-style, political / religious inclinations)
- Anonymization of location traces is **ineffective**

Our Proposal

Our Proposal

- Our approach : *data aggregation* for gathering location statistics

Our Proposal

- Our approach : *data aggregation* for gathering location statistics
- Our goals :

Our Proposal

- Our approach : *data aggregation* for gathering location statistics
- Our goals :

  1. usefulness of aggregate locations for mobility analytics

Our Proposal

- Our approach : *data aggregation* for gathering location statistics
- Our goals :
  1. usefulness of aggregate locations for mobility analytics
  2. real-world deployability of a system for privacy-friendly location data collection via crowd-sourcing

Roadmap

Roadmap

- Experiment with real-world mobility datasets (TFL, SFC)

Roadmap

- Experiment with real-world mobility datasets (TFL, SFC)
- Methodology for performing mobility analytics over aggregate locations
  1. forecasting traffic volumes in *regions of interest* (ROIs)
  2. detecting mobility anomalies
  3. improving traffic volume predictions in the presence of anomalies

Roadmap

- Experiment with real-world mobility datasets (TFL, SFC)
- Methodology for performing mobility analytics over aggregate locations
  1. forecasting traffic volumes in *regions of interest* (ROIs)
  2. detecting mobility anomalies
  3. improving traffic volume predictions in the presence of anomalies
- Design a privacy-respecting system for crowd-sourcing location data
- Empirical evaluation of computation / communication / energy complexities

Transport for London (TFL)

Transport for London (TFL)

- Logs of anonymized oyster card trips including Underground (LUL), National Rail (NR), Overground (LRC), Docklands Light Railway (DLR)

- Logs of anonymized oyster card trips including Underground (LUL), National Rail (NR), Overground (LRC), Docklands Light Railway (DLR)
- Monday, March 1 to Sunday, March 28, 2010 (4 weeks)

Transport for London (TFL)

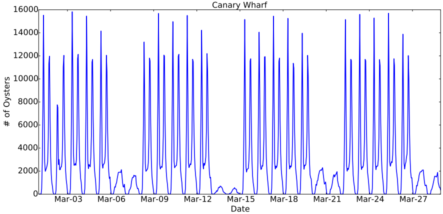
- Logs of anonymized oyster card trips including Underground (LUL), National Rail (NR), Overground (LRC), Docklands Light Railway (DLR)
- Monday, March 1 to Sunday, March 28, 2010 (4 weeks)
- 60 million trips as performed by 4 million unique users, over 582 stations

Transport for London (TFL)

- Logs of anonymized oyster card trips including Underground (LUL), National Rail (NR), Overground (LRC), Docklands Light Railway (DLR)
- Monday, March 1 to Sunday, March 28, 2010 (4 weeks)
- 60 million trips as performed by 4 million unique users, over 582 stations
- We build hourly time series (TS) of stations (Y_t), counting # of users tapping-in/out at each station

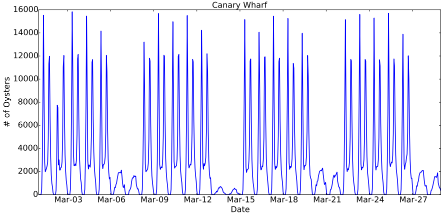
TFL Aggregates

TFL Aggregates

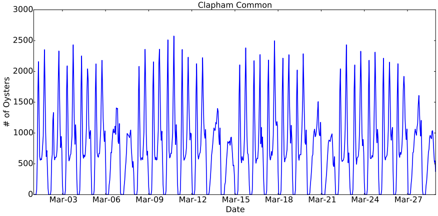


Canary Wharf TS.

TFL Aggregates

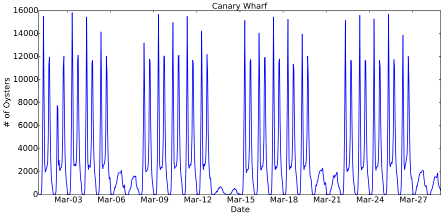


Canary Wharf TS.

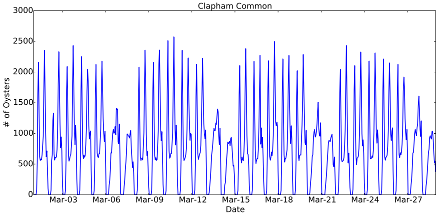


Clapham Common TS.

TFL Aggregates



Canary Wharf TS.



Clapham Common TS.

We observe daily / weekly **seasonality** and stationarity

San Francisco Cab Network (SFC)

- Mobility traces of 536 cabs in San Francisco between May 19 to June 8, 2008 (3 weeks)

San Francisco Cab Network (SFC)

- Mobility traces of 536 cabs in San Francisco between May 19 to June 8, 2008 (3 weeks)
- 11 million GPS coordinates

San Francisco Cab Network (SFC)

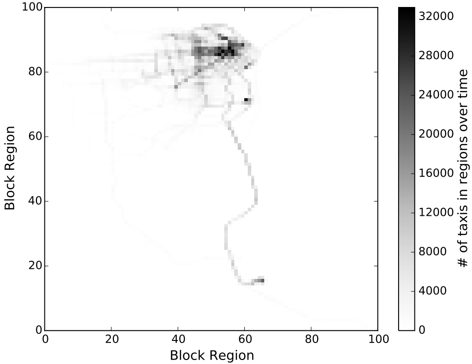
- Mobility traces of 536 cabs in San Francisco between May 19 to June 8, 2008 (3 weeks)
- 11 million GPS coordinates
- San Francisco grid of 100 x 100 regions, each of 0.19×0.14 sq mi

San Francisco Cab Network (SFC)

- Mobility traces of 536 cabs in San Francisco between May 19 to June 8, 2008 (3 weeks)
- 11 million GPS coordinates
- San Francisco grid of 100 x 100 regions, each of 0.19×0.14 sq mi
- We build hourly time series (TS) for ROIs (Y_t), counting # of taxis that have reported presence

San Francisco Cab Network (SFC)

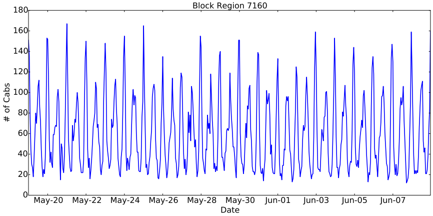
- Mobility traces of 536 cabs in San Francisco between May 19 to June 8, 2008 (3 weeks)
- 11 million GPS coordinates
- San Francisco grid of 100 x 100 regions, each of 0.19×0.14 sq mi
- We build hourly time series (TS) for ROIs (Y_t), counting # of taxis that have reported presence



SFC 100×100 grid.

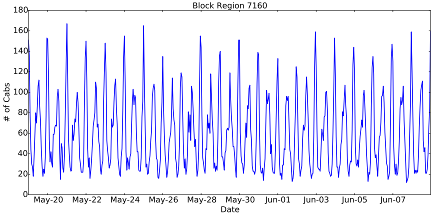
SFC Aggregates

SFC Aggregates

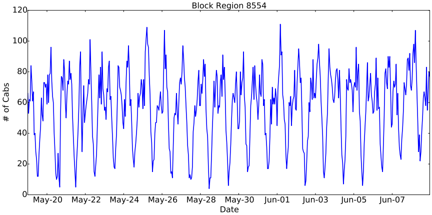


Region 7160 TS.

SFC Aggregates

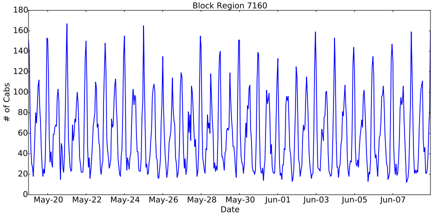


Region 7160 TS.

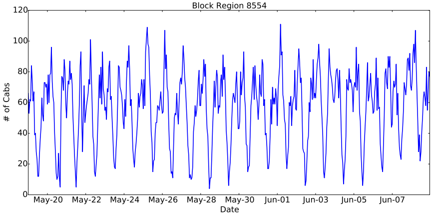


Region 8554 TS.

SFC Aggregates



Region 7160 TS.



Region 8554 TS.

We observe daily, weekly patterns and stationarity

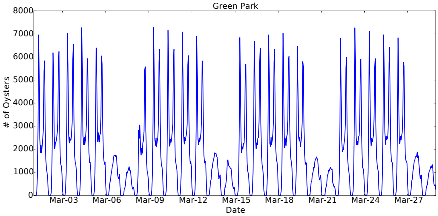
Removing Seasonality

Removing Seasonality

Additive decomposition of TS : $D_t = Y_t - \overline{Y_t}$

Removing Seasonality

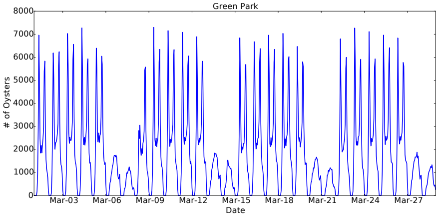
Additive decomposition of TS : $D_t = Y_t - \overline{Y_t}$



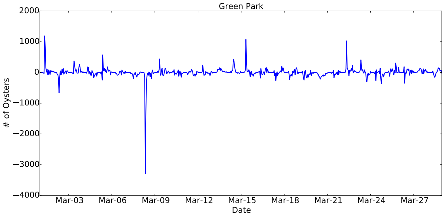
Green Park Aggregate TS.

Removing Seasonality

Additive decomposition of TS : $D_t = Y_t - \overline{Y_t}$



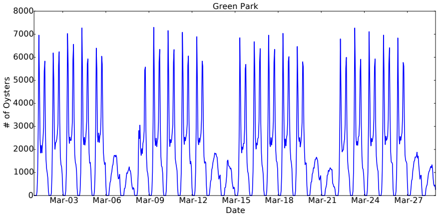
Green Park Aggregate TS.



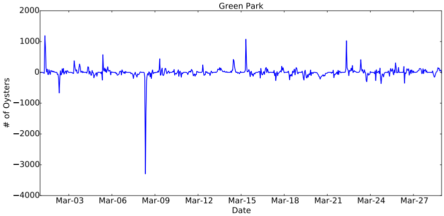
Green Park De-seasonalized TS.

Removing Seasonality

Additive decomposition of TS : $D_t = Y_t - \overline{Y_t}$



Green Park Aggregate TS.



Green Park De-seasonalized TS.

De-seasonalized time series (D_t) show strong **auto-regressive** structure

Forecasting Traffic Volumes in ROIs

Forecasting Traffic Volumes in ROIs

- ARMA_{SEAS} modeling

Forecasting Traffic Volumes in ROIs

- ARMA_{SEAS} modeling
- $\widehat{Y_t} = \widehat{D_t} + \overline{Y_t}$, using a sliding window

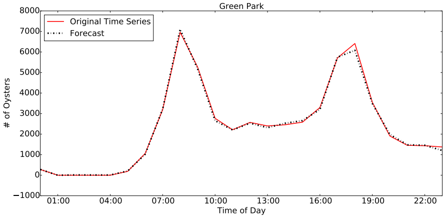
Forecasting Traffic Volumes in ROIs

- ARMA_{SEAS} modeling
- $\widehat{Y_t} = \widehat{D_t} + \overline{Y_t}$, using a sliding window
- Evaluate accuracy via absolute forecast error ($e_t = |Y_t - \widehat{Y_t}|$)

- Experiment with *top* 100 TFL stations and SFC ROIs

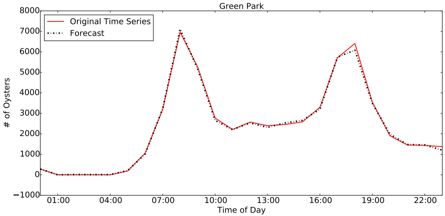
- Experiment with *top* 100 TFL stations and SFC ROIs
- 5 days of data for training (D_t) - 1 day of testing ($\widehat{Y_t}$ vs Y_t)

- Experiment with *top* 100 TFL stations and SFC ROIs
- 5 days of data for training (D_t) - 1 day of testing ($\widehat{Y_t}$ vs Y_t)

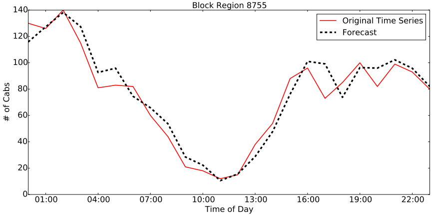


Green Park Predictions, March 25.

- Experiment with *top* 100 TFL stations and SFC ROIs
- 5 days of data for training (D_t) - 1 day of testing ($\widehat{Y_t}$ vs Y_t)

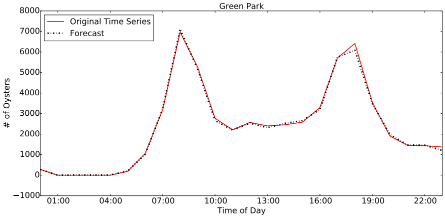


Green Park Predictions, March 25.

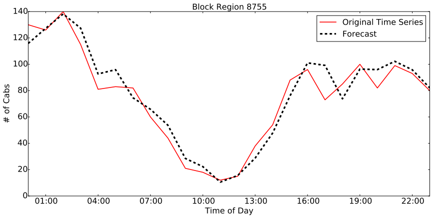


Region 8755 Predictions, June 5.

- Experiment with *top* 100 TFL stations and SFC ROIs
- 5 days of data for training (D_t) - 1 day of testing ($\widehat{Y_t}$ vs Y_t)



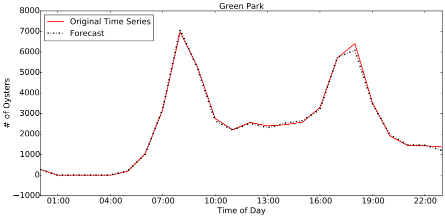
Green Park Predictions, March 25.



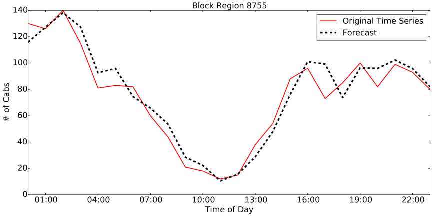
Region 8755 Predictions, June 5.

- Comparison to a baseline *black-box* ARMA model on Y_t

- Experiment with *top* 100 TFL stations and SFC ROIs
- 5 days of data for training (D_t) - 1 day of testing ($\widehat{Y_t}$ vs Y_t)



Green Park Predictions, March 25.



Region 8755 Predictions, June 5.

- Comparison to a baseline *black-box* ARMA model on Y_t
- **Improved** predictions when considering seasonal effects (e.g. TFL average error : 19% vs 600%)

Detecting Traffic Anomalies in ROIs

Detecting Traffic Anomalies in ROIs

- ARMA_{SEAS} modeling - rely on absolute forecast error (e_t)

Detecting Traffic Anomalies in ROIs

- ARMA_{SEAS} modeling - rely on absolute forecast error (e_t)
- Apply the 3σ rule, with confidence interval : $\lambda = \mu + 3\sigma$

Detecting Traffic Anomalies in ROIs

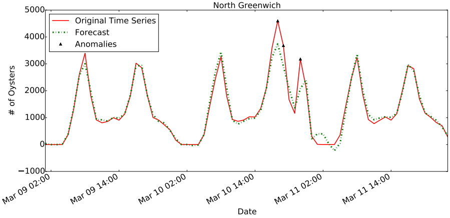
- ARMA$_{SEAS}$ modeling - rely on absolute forecast error (e_t)
- Apply the 3σ rule, with confidence interval : $\lambda = \mu + 3\sigma$
- Detect an anomaly at time t if : $e_t > \lambda$

- Train the ARMA_{SEAS} model with 1 week data, test it against the rest of weeks

- Train the ARMA_{SEAS} model with 1 week data, test it against the rest of weeks
- Top 100 TFL stations : 896 anomalies

- Train the ARMA_{SEAS} model with 1 week data, test it against the rest of weeks
- Top 100 TFL stations : 896 anomalies
- Top 100 SFC blocks: 366 anomalies

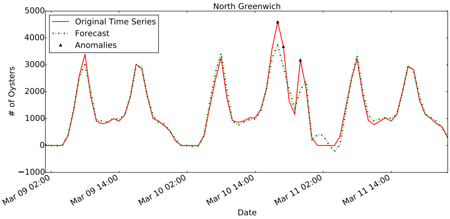
- Train the ARMA_{SEAS} model with 1 week data, test it against the rest of weeks
- Top 100 TFL stations : 896 anomalies
- Top 100 SFC blocks: 366 anomalies



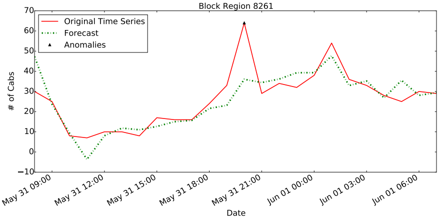
North Greenwich, March 10.

Experiments
Detecting Traffic Anomalies in ROIs

- Train the ARMA_{SEAS} model with 1 week data, test it against the rest of weeks
- Top 100 TFL stations : 896 anomalies
- Top 100 SFC blocks: 366 anomalies

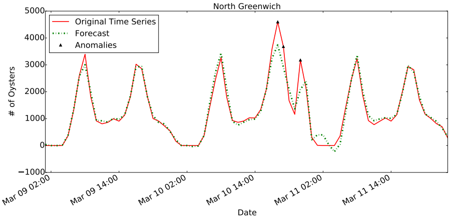


North Greenwich, March 10.

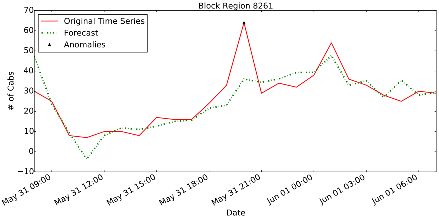


Region 8261, May 31.

- Train the $ARMA_{SEAS}$ model with 1 week data, test it against the rest of weeks
- Top 100 TFL stations : 896 anomalies
- Top 100 SFC blocks: 366 anomalies



North Greenwich, March 10.



Region 8261, May 31.

Note : no ground truth for *anomalies*

Predicting Traffic Volumes during Anomalies

Predicting Traffic Volumes during Anomalies

- Can we improve our predictions in the presence of an anomaly?

Predicting Traffic Volumes during Anomalies

- Can we improve our predictions in the presence of an anomaly?
- Discover *correlated* ROIs by sliding their time series - (Spearman correlation)

Predicting Traffic Volumes during Anomalies

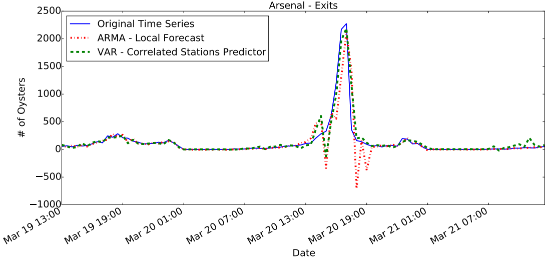
- Can we improve our predictions in the presence of an anomaly?
- Discover *correlated* ROIs by sliding their time series - (Spearman correlation)
- Use a VAR model to capture linear inter-dependencies between time series

- Experiment with 10% of anomalies of TFL (90 anoms) and SFC (30 anoms)

- Experiment with 10% of anomalies of TFL (90 anoms) and SFC (30 anoms)
- Train a VAR model including information from 10 correlated ROIs

- Experiment with 10% of anomalies of TFL (90 anoms) and SFC (30 anoms)
- Train a VAR model including information from 10 correlated ROIs
- Compare against a *baseline* : ARMA_{SEAS} model trained on local data

- Experiment with 10% of anomalies of TFL (90 anoms) and SFC (30 anoms)
- Train a VAR model including information from 10 correlated ROIs
- Compare against a *baseline* : ARMA$_{SEAS}$ model trained on local data



Arsenal Exit Traffic Predictions during an Anomaly, March 20.

- ARMA$_{SEAS}$ Error 93% vs VAR Error 59%

- Overall, significant improvement in predictions when considering information from correlated ROIs

- Overall, significant improvement in predictions when considering information from correlated ROIs
- TFL : 29% improvement in predictions

- Overall, significant improvement in predictions when considering information from correlated ROIs
- TFL : 29% improvement in predictions
- SFC : 18% improvement in predictions

What next?

What next?

- Analytics on aggregate locations offer interesting insights

What next?

- Analytics on aggregate locations offer interesting insights
- Can we collect aggregate locations directly from users, with **privacy**?

What next?

- Analytics on aggregate locations offer interesting insights
- Can we collect aggregate locations directly from users, with **privacy**?
- Challenges : Efficiency, scalability, fault-tolerance

What next?

- Analytics on aggregate locations offer interesting insights
- Can we collect aggregate locations directly from users, with **privacy**?
- Challenges : Efficiency, scalability, fault-tolerance
- Good news: promising results by Melis et al. (NDSS 2016) [1]

[1] Luca Melis, George Danezis, and Emiliano De Cristofaro : Efficient private statistics with succinct sketches, NDSS (2016).

Mobility Data Donors (MDD) Framework

Mobility Data Donors (MDD) Framework

- Design a collaborative framework for aggregate location data collection (users vs aggregator)

Mobility Data Donors (MDD) Framework

- Design a collaborative framework for aggregate location data collection (users vs aggregator)
- **Client Side**:

Mobility Data Donors (MDD) Framework

- Design a collaborative framework for aggregate location data collection (users vs aggregator)
- **Client Side**:
  - Users install MDD app

Mobility Data Donors (MDD) Framework

- Design a collaborative framework for aggregate location data collection (users vs aggregator)
- **Client Side**:
  - Users install MDD app
  - MDD runs on the background, collecting GPS coordinates

Mobility Data Donors (MDD) Framework

- Design a collaborative framework for aggregate location data collection (users vs aggregator)
- **Client Side**:
  - Users install MDD app
  - MDD runs on the background, collecting GPS coordinates
  - Aggregator periodically triggers privacy-preserving aggregation, assigning users to groups

Mobility Data Donors (MDD) Framework

- Design a collaborative framework for aggregate location data collection (users vs aggregator)
- **Client Side**:
  - Users install MDD app
  - MDD runs on the background, collecting GPS coordinates
  - Aggregator periodically triggers privacy-preserving aggregation, assigning users to groups
  - MDD encrypts entries in the matrix that represents user locations

Mobility Data Donors (MDD) Framework

- Design a collaborative framework for aggregate location data collection (users vs aggregator)
- **Client Side**:
  - Users install MDD app
  - MDD runs on the background, collecting GPS coordinates
  - Aggregator periodically triggers privacy-preserving aggregation, assigning users to groups
  - MDD encrypts entries in the matrix that represents user locations
- **Server side**:

Mobility Data Donors (MDD) Framework

- Design a collaborative framework for aggregate location data collection (users vs aggregator)
- **Client Side**:
  - Users install MDD app
  - MDD runs on the background, collecting GPS coordinates
  - Aggregator periodically triggers privacy-preserving aggregation, assigning users to groups
  - MDD encrypts entries in the matrix that represents user locations
- **Server side**:
  - Aggregator collects the encrypted matrices and decrypts **ONLY** aggregate location counts - combines aggregates if collected from multiple groups

MDD Experimental Evaluation

MDD Experimental Evaluation

- Javascript/Node.js implementation of the secure aggregation protocol by Melis et al.
- Port of client side to run on Android, via Apache Cordova

MDD Experimental Evaluation

- Javascript/Node.js implementation of the secure aggregation protocol by Melis et al.
- Port of client side to run on Android, via Apache Cordova
- Cryptographic operations : Edc25519 elliptic curve - 128 bit security

MDD Experimental Evaluation

- Javascript/Node.js implementation of the secure aggregation protocol by Melis et al.
- Port of client side to run on Android, via Apache Cordova
- Cryptographic operations : Edc25519 elliptic curve - 128 bit security
- Android device : Samsung Galaxy A3, 1.2 GHz quad-core Snapdragon 410, 1.5 GB RAM, Lollipop v5.0.2

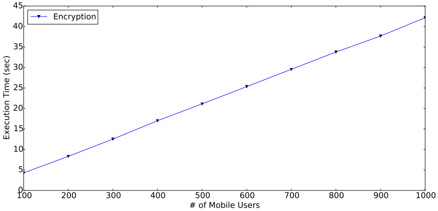
MDD Experimental Evaluation

- Javascript/Node.js implementation of the secure aggregation protocol by Melis et al.
- Port of client side to run on Android, via Apache Cordova
- Cryptographic operations : Edc25519 elliptic curve - 128 bit security
- Android device : Samsung Galaxy A3, 1.2 GHz quad-core Snapdragon 410, 1.5 GB RAM, Lollipop v5.0.2
- PowerTutor app for power monitoring

Results - TFL

Results - TFL

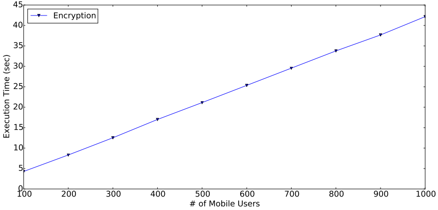
- ROI matrix of size (582, 2)
- $\sim$ 7s encryption for groups of 200 mobile users



TFL Execution Time - Encryption Phase.

Results - TFL

- ROI matrix of size (582, 2)
- $\sim$ 7s encryption for groups of 200 mobile users

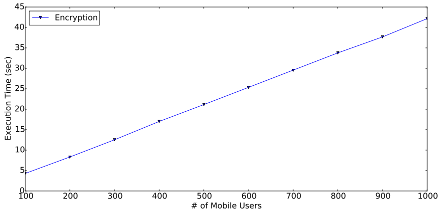


TFL Execution Time - Encryption Phase.

- 10.7KB public keys, 4.54KB encrypted ROI matrix

Results - TFL

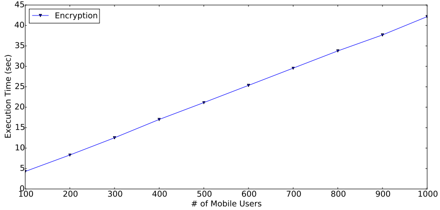
- ROI matrix of size (582, 2)
- $\sim$ 7s encryption for groups of 200 mobile users



TFL Execution Time - Encryption Phase.

- 10.7KB public keys, 4.54KB encrypted ROI matrix
- 826mJ encryption, 609mJ / 322mJ download / upload (via Wi-Fi)

- ROI matrix of size (582, 2)
- $\sim$ 7s encryption for groups of 200 mobile users



TFL Execution Time - Encryption Phase.

- 10.7KB public keys, 4.54KB encrypted ROI matrix
- 826mJ encryption, 609mJ / 322mJ download / upload (via Wi-Fi)

Note : Succinct data representation can be used, if more fine grained data need to be collected (e.g. O-D matrices)

Conclusions

Conclusions

- Mobility analytics over *crowd-sourced* aggregate location data

Conclusions

- Mobility analytics over *crowd-sourced* aggregate location data
- Time series modeling with seasonality for:
  1. forecasting traffic volumes in ROIs
  2. detecting anomalies
  3. improving traffic volume predictions in the presence of anomalies

- Mobility analytics over *crowd-sourced* aggregate location data
- Time series modeling with seasonality for:
    1. forecasting traffic volumes in ROIs
    2. detecting anomalies
    3. improving traffic volume predictions in the presence of anomalies
- Experiments on real-world mobility datasets (TFL, SFC)

Conclusions

- Mobility analytics over *crowd-sourced* aggregate location data
- Time series modeling with seasonality for:
  1. forecasting traffic volumes in ROIs
  2. detecting anomalies
  3. improving traffic volume predictions in the presence of anomalies
- Experiments on real-world mobility datasets (TFL, SFC)
- Privacy-respecting system for data collection
- Mobile application framework (MDD) and empirical evaluation in terms of computation / communication / energy overhead

Future Work

Future Work

- Evaluate our methodology on different mobility datasets

Future Work

- Evaluate our methodology on different mobility datasets
- Privacy quantification and analysis of aggregate location data

Future Work

- Evaluate our methodology on different mobility datasets
- Privacy quantification and analysis of aggregate location data
  - group sizes

Future Work

- Evaluate our methodology on different mobility datasets
- Privacy quantification and analysis of aggregate location data
    - group sizes
    - characteristics of ROIs (density, size, time)

Future Work

- Evaluate our methodology on different mobility datasets
- Privacy quantification and analysis of aggregate location data
    - group sizes
    - characteristics of ROIs (density, size, time)
    - semantics of ROIs

The end...

The end...

Thanks for your attention! Any questions?

The end...

Thanks for your attention! Any questions?

Contact Details: **apostolos.pyrgelis.14@ucl.ac.uk**