Spatial Big Data Research and Applications at UCY

Demetris Zeinalipour

Data Management Systems Laboratory

Department of Computer Science

University of Cyprus

http://www.cs.ucy.ac.cy/~dzeina/

The First Europe-China Workshop on Big Data Management
May 16, 2016, Kumpulan Kampus, University of Helsinki, Finland

Big Data Management

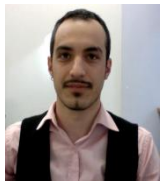
http://udbms.cs.helsinki.fi/BigData2016/

University of Cyprus

- Cyprus:
  - Island in Mediterranean
  - EU Member (2004) / Eurozone (2008)
- University of Cyprus:
  - Founded in 1989, 350 Faculty, 7000 Stud, 21 Departments
  - Ranked 55th in top-200 Young Univ. by the Times.
- Dept. of Computer Science – 22 Faculty Members
- Data Management Systems Laboratory (DMSL) Scope:
  - Mobile and Sensor Data Management
  - Big Data Management (Parallel and Distributed)
  - Spatio-Temporal Data Management
  - Network and Telco Data Management
  - Crowd, Web 2.0 and Indoor Data Management
  - Data Privacy Management

DMSL @ UCY

- **People:**
  - 4 Researchers (Faculty + Postdocs)
  - 2 PhD Students + MSc Students
  - BSc. (USC, UCLA/R, UoT, Oxford, UCL, Imperial)





- **External Collaborators**
  - Panos K. Chrysanthis (U of Pittsburgh, USA)
  - Wang-Chien Lee (Penn State, USA)
  - Yannis Theodoridis (U of Pireaus, Greece)
  - Evaggelia Pitoura (U of Ioannina, Greece)
  - Gerhard Weikum (MPI, Germany)
- **Funding:** EU, National & Industry

Talk Outline

- **Spatial Indoor Data – Anyplace**
  - Proposal, Papers, Software, Users, Community
- Spatial Social Data – Rayzit
  - Proposal, Publications, Software, Users,
- Spatial Telco Data – Spate
  - Proposal, Publications, Software,
- Spatial Green Data – GreenCharge
  - Proposal

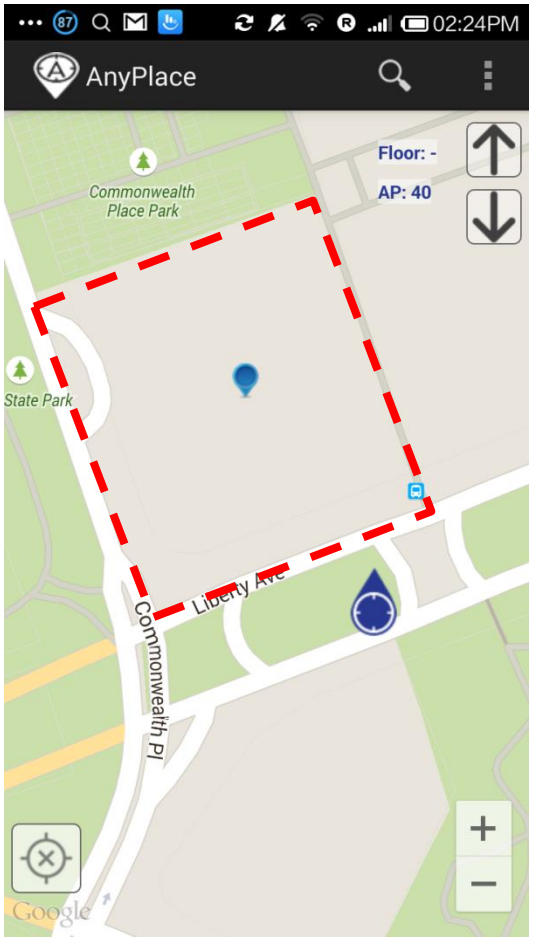
Anyplace

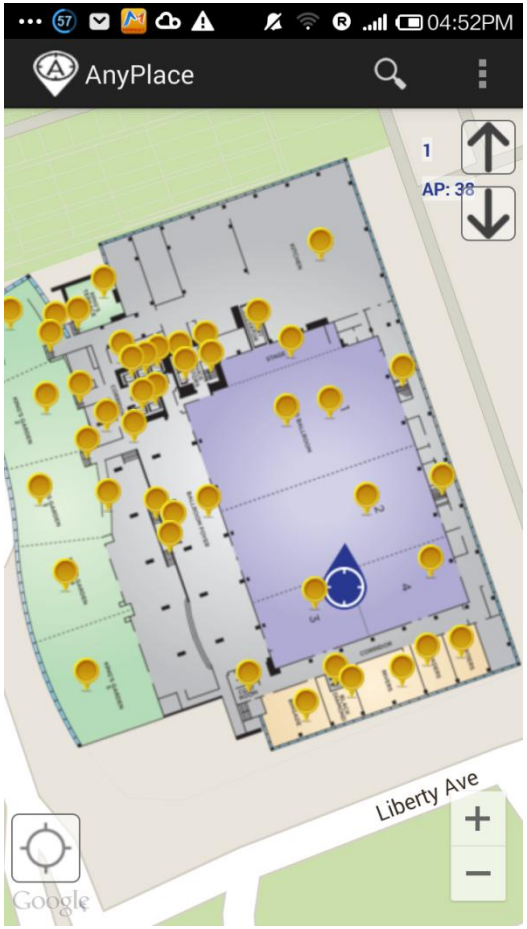
A complete open-source **Internet-based Indoor Navigation (IIN)** Service

- Aims to become **the predominant open-source** Indoor Localization Service.
- **Active community:** Germany, Russia, Australia, Canada, UK, etc. – Join today!
- Android, Windows, iOS, JSON API

http://anyplace.cs.ucy.ac.cy/

Anyplace





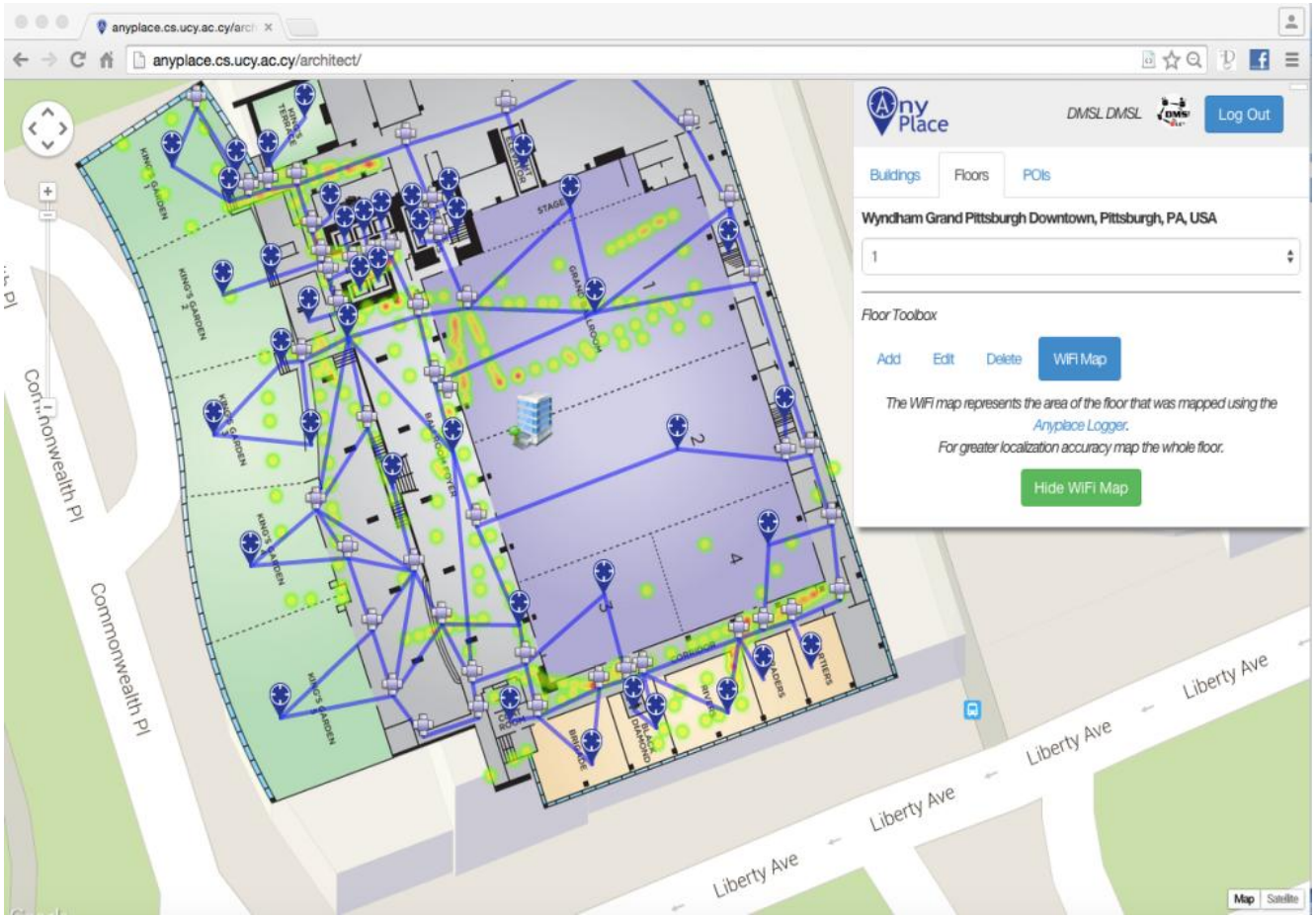
Before ☹

(using Google API Location)

After ☺

(using Anyplace Location & Indoor Models)

Showcase I: Hotel in Pittsburgh, USA

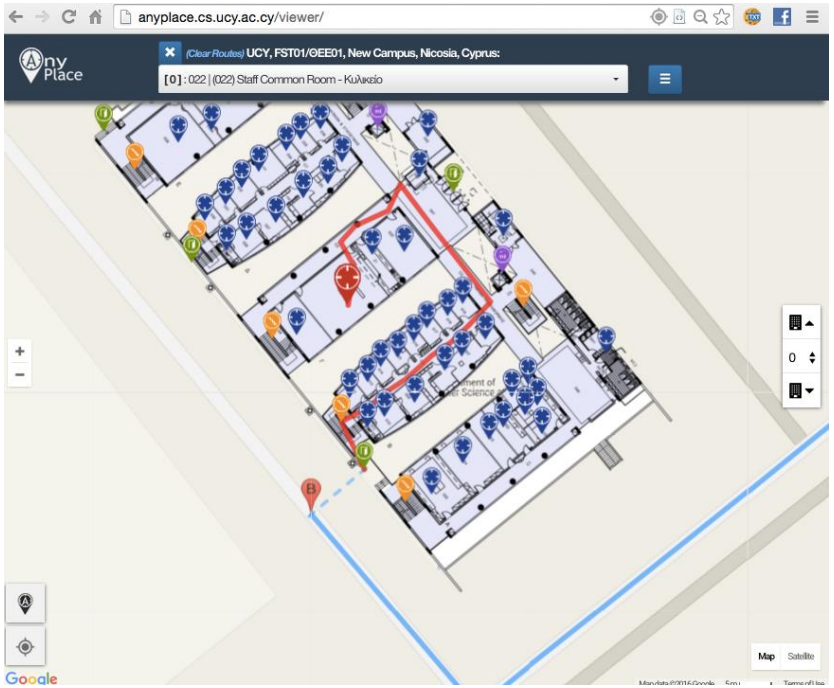


Modeling + Crowdsourcing

University of Cyprus

Showcase II: Univ. of Cyprus

- Office Navigation @ Univ. of Cyprus
  - Outdoor-to-Indoor Navigation through URL.
  - **60 Buildings** mapped, Thousands of POIs (stairways, WC, elevators, equipment, etc.)

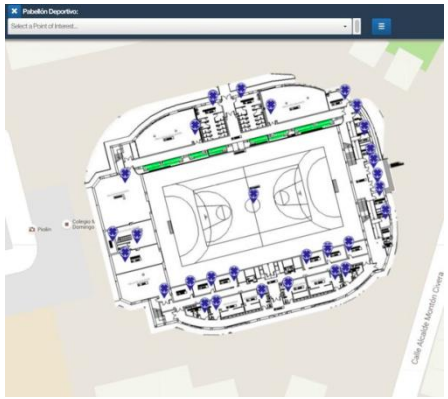


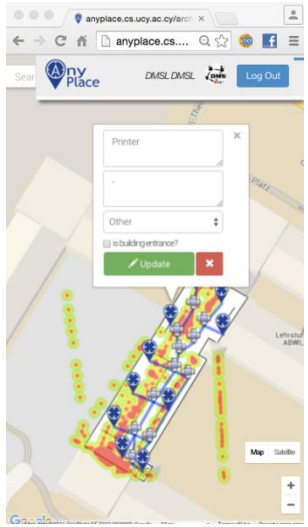
Example:

http://goo.gl/ns3lqN

Other Showcases

- Univ. of Würzburg, Institut für Informatik
  - Mapped in about 1 hour
- Universidad de Jaén, Spain
  - Campus Navigation (9 Buildings)
- Univ. of Mannheim, Library
  - Aims to offer Navigation-to-Shelf





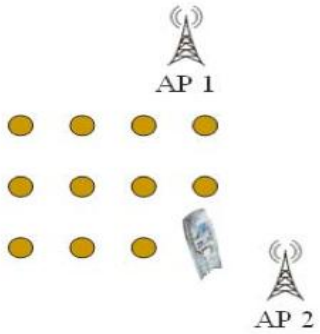
Anyplace Worldwide Deployment

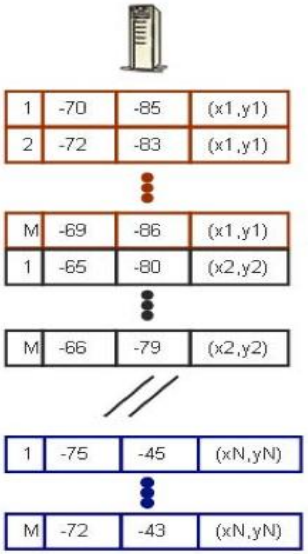
Anyplace Challenges

- Challenge I: Location Accuracy
  - IEEE MDM'12, ACM Mobisys'13, ACM IPSN'14, ACM IPSN'15, IEEE IC'16.
  - Awards:
    - Best Demo Award @ IEEE MDM'12.
    - 2nd Position with 1.96m! @ Microsoft Indoor Localization Competition at ACM IPSN'14, Berlin, Germany, April 13-14, 2014.
    - 1st Position at EVARILOS Open Challenge, European Union (TU Berlin, Germany), 2014.
- Challenge II: Location Privacy
  - IEEE TKDE'15
- Challenge III: Location Prefetching
  - IEEE MDM'15
- Other Challenges: Big Data, Crowdsourcing

Location Accuracy

- **Mapping Area with WiFi Fingerprints**
  - Repeat process for rest points in building. (IEEE MDM'12)
  - Use 4 **direction mapping (NSWE)** to overcome body blocking or reflecting the wireless signals.
  - Collect measurements while **walking** in **straight lines** (IPIN'14)





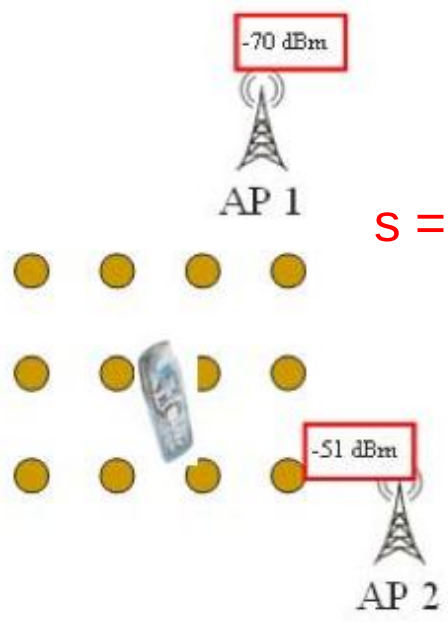
Location Accuracy

Video

"Anyplace: A Crowdsourced Indoor Information Service", Kyriakos Georgiou, Timotheos Constambeys, Christos Laoudias, Lambros Petrou, Georgios Chatzimilioudis and Demetrios Zeinalipour-Yazti, Proceedings of the 16th IEEE International Conference on Mobile Data Management (MDM '15), IEEE Press, Volume 2, Pages: 291-294, 2015

Location Accuracy

- **Positioning with WiFi Fingerprint**
  - Collect Fingerprint $s = [s_1, s_2, \ldots, s_n]$
  - Compute distance $|| r_i - s ||$ and position user at:
    - Nearest Neighbor (NN)
    - K Nearest Neighbors ($w_i = 1 / K$) -convex combination of k loc
    - Weighted K Nearest Neighbors ($w_i = 1 / || r_i - s ||$)



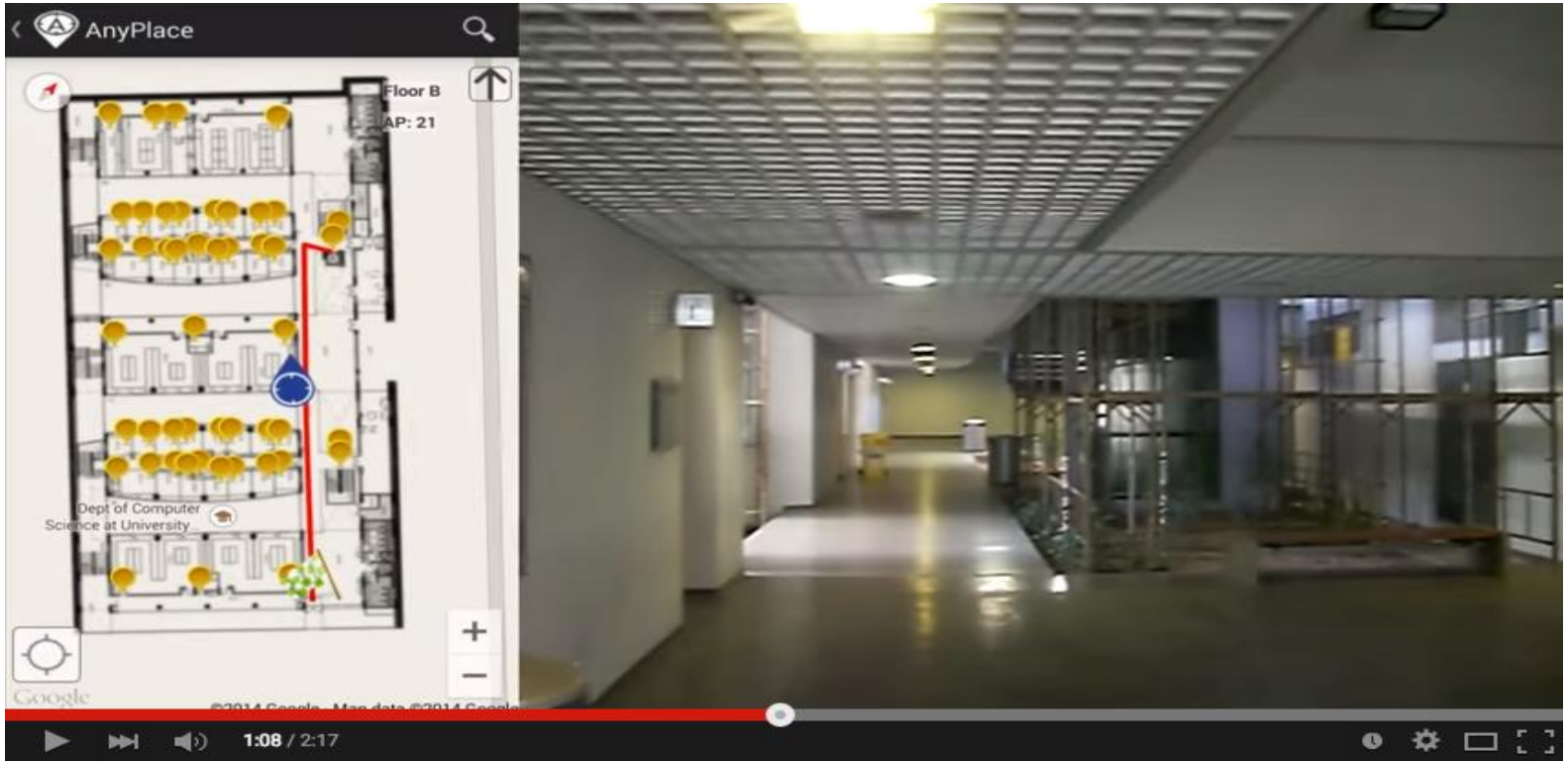
s = [-70, -51]

NN, KNN, WKNN

RadioMap

$r_1 = [-71, -82, (x_1,y_1)]$
$r_2 = [-65, -80, (x_2,y_2)]$
…
$r_N = [-73, -44, (x_N,y_N)]$

Location Accuracy



Video

"Anyplace: A Crowdsourced Indoor Information Service", Kyriakos Georgiou, Timotheos Constambeys, Christos Laoudias, Lambros Petrou, Georgios Chatzimilioudis and Demetrios Zeinalipour-Yazti, Proceedings of the 16th IEEE International Conference on Mobile Data Management **(MDM '15)**, IEEE Press, Volume 2, Pages: 291-294, 2015

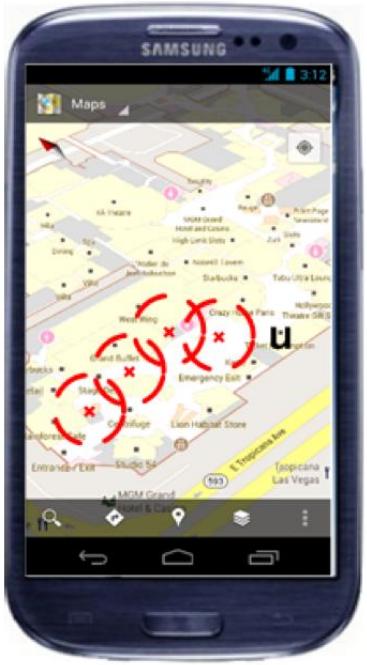
Big-Data Challenges

- Massively **process** RSS log traces to generate a valuable Radiomap
- Processing current logs in Anyplace for a single building takes **several minutes**!
- Challenges in MapReduce:
  - Collect Statistics (count, RSSI mean and standard deviation)
  - Remove Outlier Values.
  - Handle Diversity Issues

Location Privacy

- An IIN Service can continuously **“know” (surveil, track or monitor)** the location of a user while serving them.
- **Location tracking** is unethical and can even be illegal if it is carried out without the explicit user consent.
- **Imminent privacy threat**, with greater impact that other **privacy** concerns, as it can occur at a very **fine granularity**. It reveals:
  - The stores / products of interest in a mall.
  - The book shelves of interest in a library
  - Artifacts observed in a museum, etc.

Location Privacy



User u

I can see these Reference Points, where am I?

(x,y)!

IIN Service

-*Privacy-Preserving Indoor Localization on Smartphones*, Andreas Konstantinidis, Paschalis Mpeis, Demetrios Zeinalipour-Yazti and Yannis Theodoridis, in **IEEE TKDE'15.**

-*Towards planet-scale localization on smartphones with a partial radiomap"*, A. Konstantinidis, G. Chatzimilioudis, C. Laoudias, S. Nicolaou and D. Zeinalipour-Yazti. In ACM HotPlanet'12, in conjunction with **ACM MobiSys '12,** ACM, Pages: 9--14, 2012.

Location Privacy

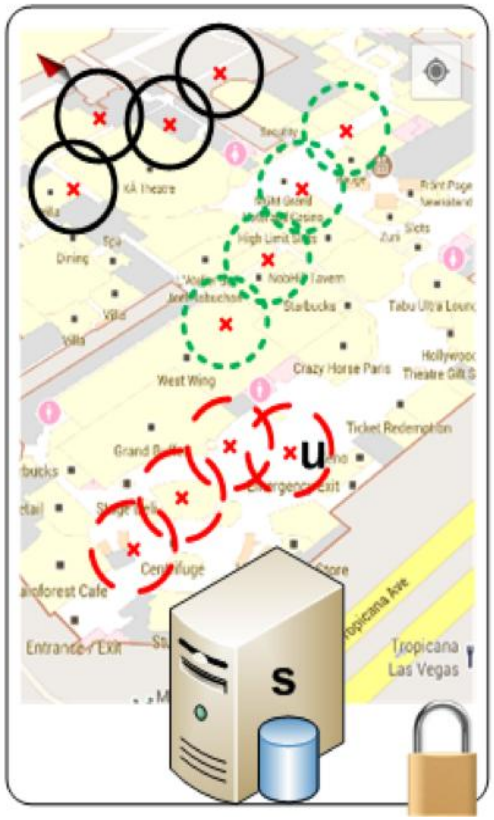
WiFi

WiFi

...

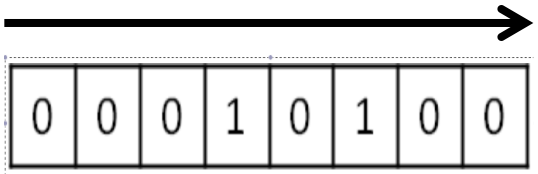
WiFi

Temporal Vector Map (TVM)



User u

kAB Filter (u's APs)



Set Membership Queries

IIN Service

K=3 Positions

k-Anonymity Bloom (kAB)

- Founded on space-efficient probabilistic **Bloom Filters** data structures.
  - allocate a vector of **b bits**, initially **all set to 0**, use **h** independent hash functions to hash every Access Point seen by a user to the vector.

AP1 AP1AP2 AP2 b

0	1	0	0	1	0	0	1	0	0

- **Tradeoff** between **b** and **probability** of a **false positive.**
- Given **h optimal** hash functions, **b bits** for the Bloom filter and the **number M** of elements a **user u** can calculate the amount of false positives produced by the Bloom filter:

False Positive Ratio

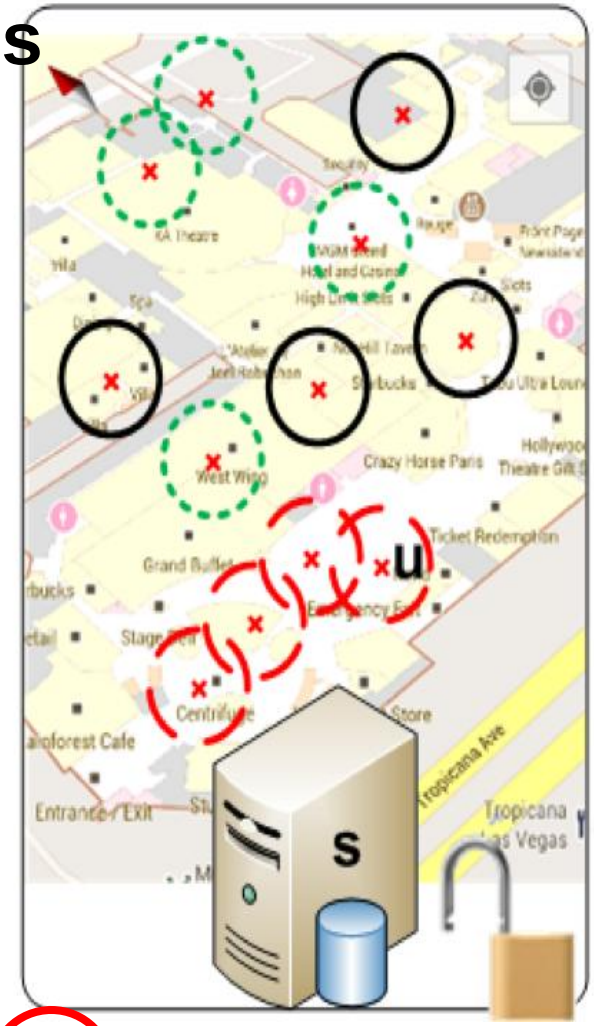
$$fpr \approx (1 - e^{-h/b})^h = k / M$$

Size of Bloom Filter

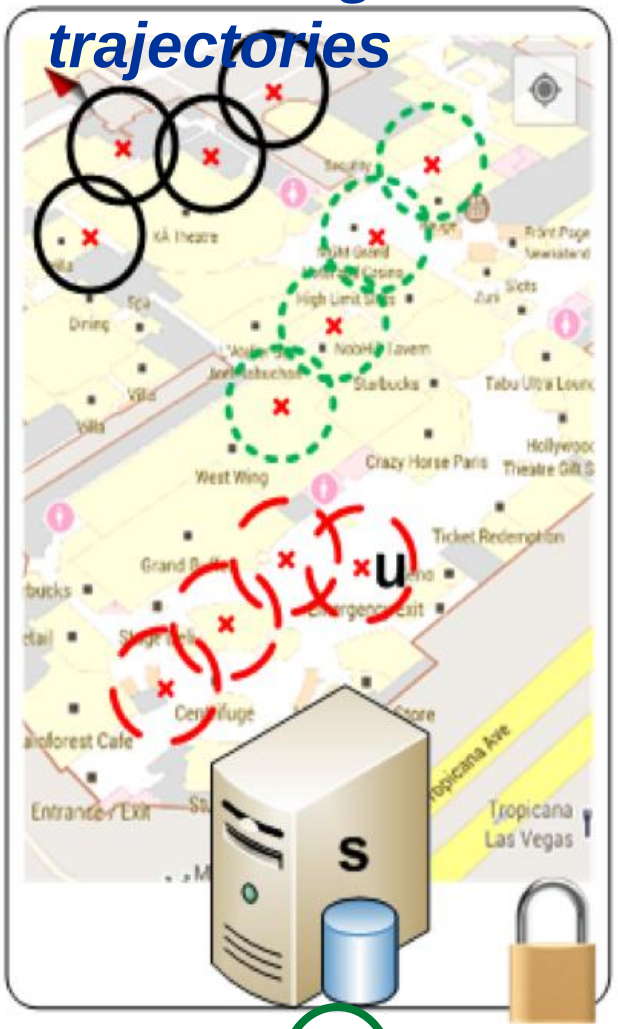
$$b = \left\lfloor \frac{-h}{\ln(1 - \sqrt[h]{k/M})} \right\rfloor$$

Location Privacy

TVM Continuous



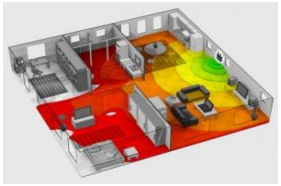
Camouflage trajectories



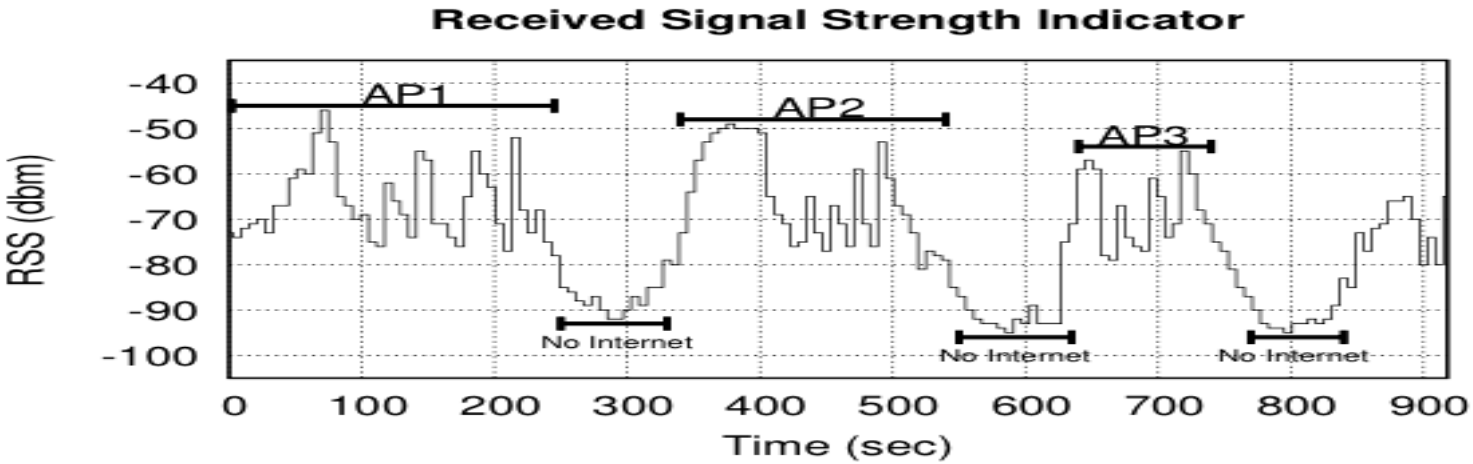
IIN determines u's location by exclusion

Location Prefetching

- **Problem*:** Wi-Fi coverage might be **irregularly available** inside buildings **due to poor WLAN planning** or due to **budget constraints.**



- A user **walking inside** a **Mall in Cyprus**
  - Whenever the user **enters a store** the RSSI indicator falls **below a connectivity** threshold **-85dBm. (-30dbM to -90dbM)**
  - When **disconnected IIN can't** offer navigation anymore ☹

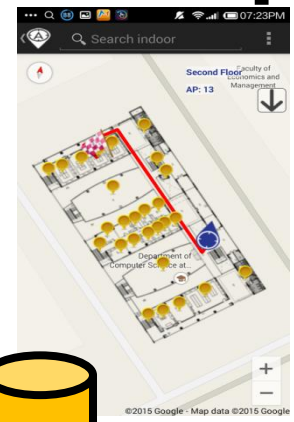


* "Radiomap Prefetching for Indoor Navigation in Intermittently Connected Wi-Fi Networks", A. Konstantinidis, G. Nikolaides, G. Chatzimilioudis, G. Evagorou, D. Zeinalipour-Yazti and P.K. Chrysanthis, In IEEE MDM'15.

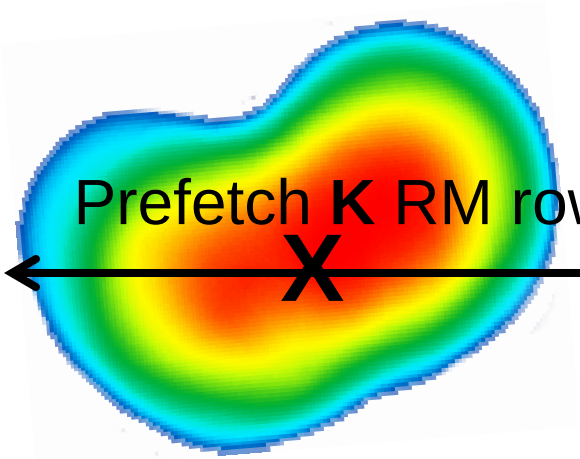
Location Prefetching

PreLoc Navigation

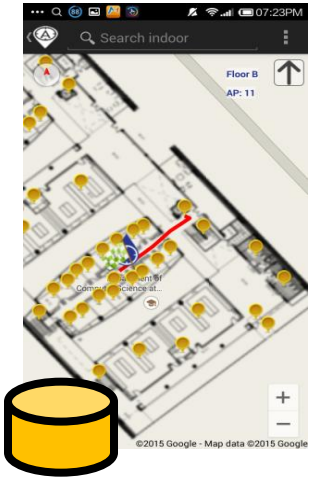
IIN Service



Prefetch **K** RM rows



Localize from Cache



Prefetch **K** RM rows

☺ Intermittent Connectivity

Time

RadioMap (RM) on s

Location	ap_1	ap_2	…	ap_M
l_1	90	85	…	-1
l_2	-1	-1	…	-1
l_3	-1	40	…	-1
l_4	-1	-1	…	-1
l_5	-1	-1	…	80
l_6	80	-1	…	-1
l_N	…	…	…	…

z

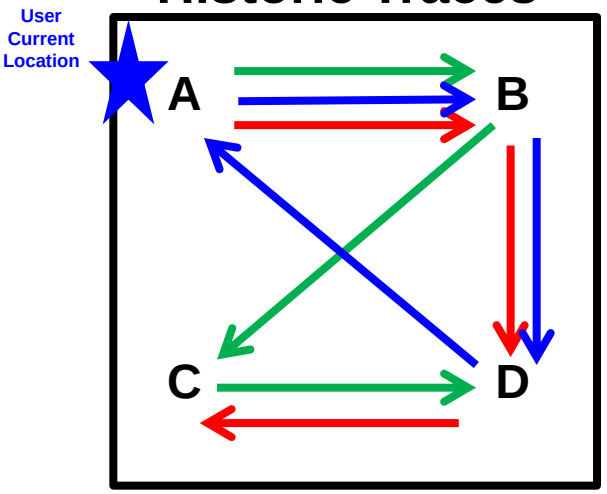
PreLoc Selection Step

- PreLoc aims to **sequence** the **retrieval** of fingerprint **clusters**, as generated by BFR clustering, such that the **most important clusters** are downloaded first.
- **Question:** Which **clusters** should a user **download** at a **certain position** if Wi-Fi **not available** next?
  - PreLoc **prioritizes** the download of **RM** entries using **historic traces** of user inside the building !!!

PreLoc Selection Step

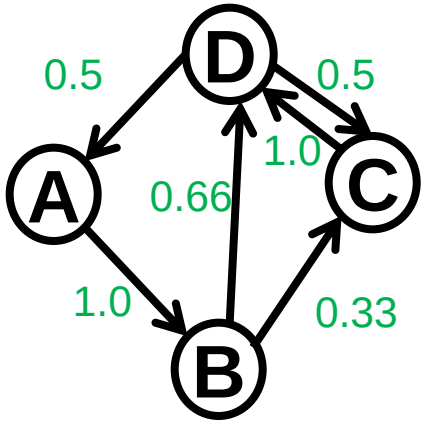
- PreLoc relies on the **Probabilistic Group Selection (PGS)** heuristic to determine the **RM entries** to **prefetch** next.

Historic Traces



(statistically independent transitions between vertices + no stationary transitions in historic traces)

Dependency Graph (DG)



Probabilistic k=3 Group Selection

Do **Best First Search Traversal of DG from A:** follow the most promising option using priority queue.

P(**A,B**)=**1.0**
P(A,B,**D**)=**0.66**
P(A,B,D,**C**)=0.66*0.5=**0.33**
~~P(A,B,D,A) => cycle~~
~~P(A,B,D,C,D)=> cycle~~
P(A,B,C)=**0.33**
Empty queue – finished!

Early stop!

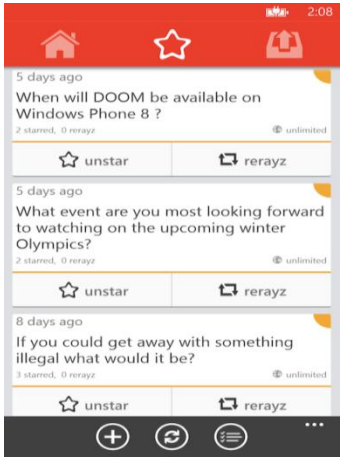
Talk Outline

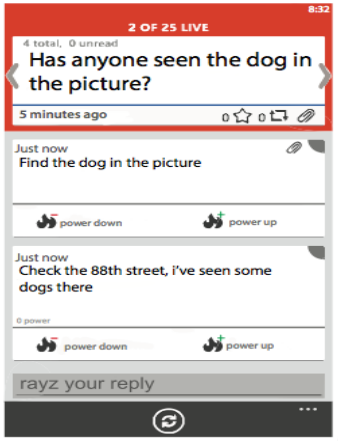
- **Spatial Indoor Data – Anyplace**
  - Proposal, Papers, Software, Users, Community
- **Spatial Social Data – Rayzit**
  - **Proposal, Publications, Software, Users,**
- Spatial Telco Data – Spate
  - Proposal, Publications, Software,
- Spatial Green Data – GreenCharge
  - Proposal

Rayzit

- Rayzit is a **crowd messaging technology** that delivers **questions, inquiries and ideas** to the **KNN users.**
  - Not location-based app (e.g., 5 km) but kNN-based!
  - Addresses: Proximity, Privacy, Bootstrapping, etc.
- Rayzit was funded by the Appcampus Program (Microsoft, Nokia & Aalto, Finland).
  - Ranked among 5 best apps of the program from 3500 apps.
  - Few thousand downloads and active users after their marketing!

http://rayzit.com/





Windows Store

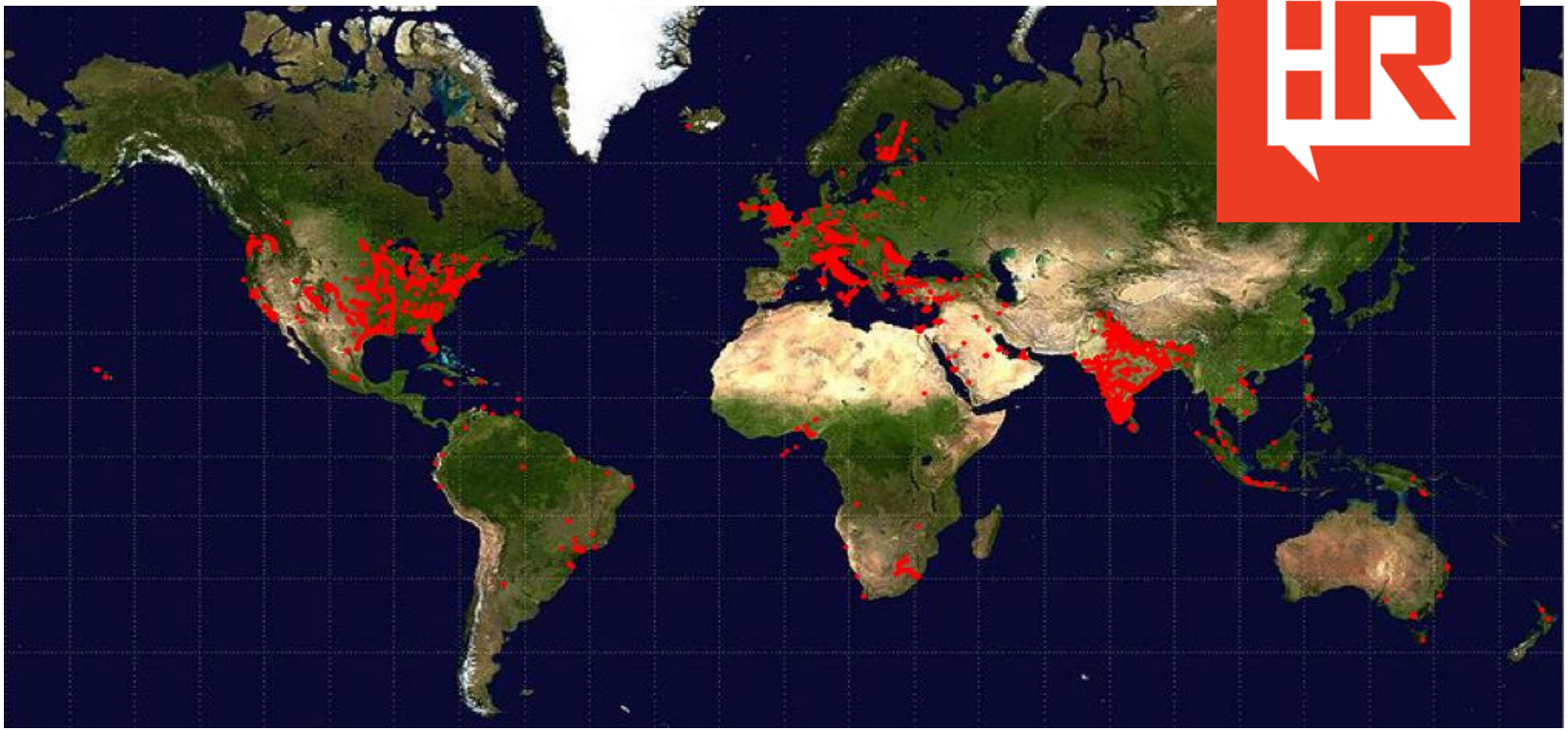
Download on the App Store

GET IT ON Google play

University of Cyprus

Rayzit Concept

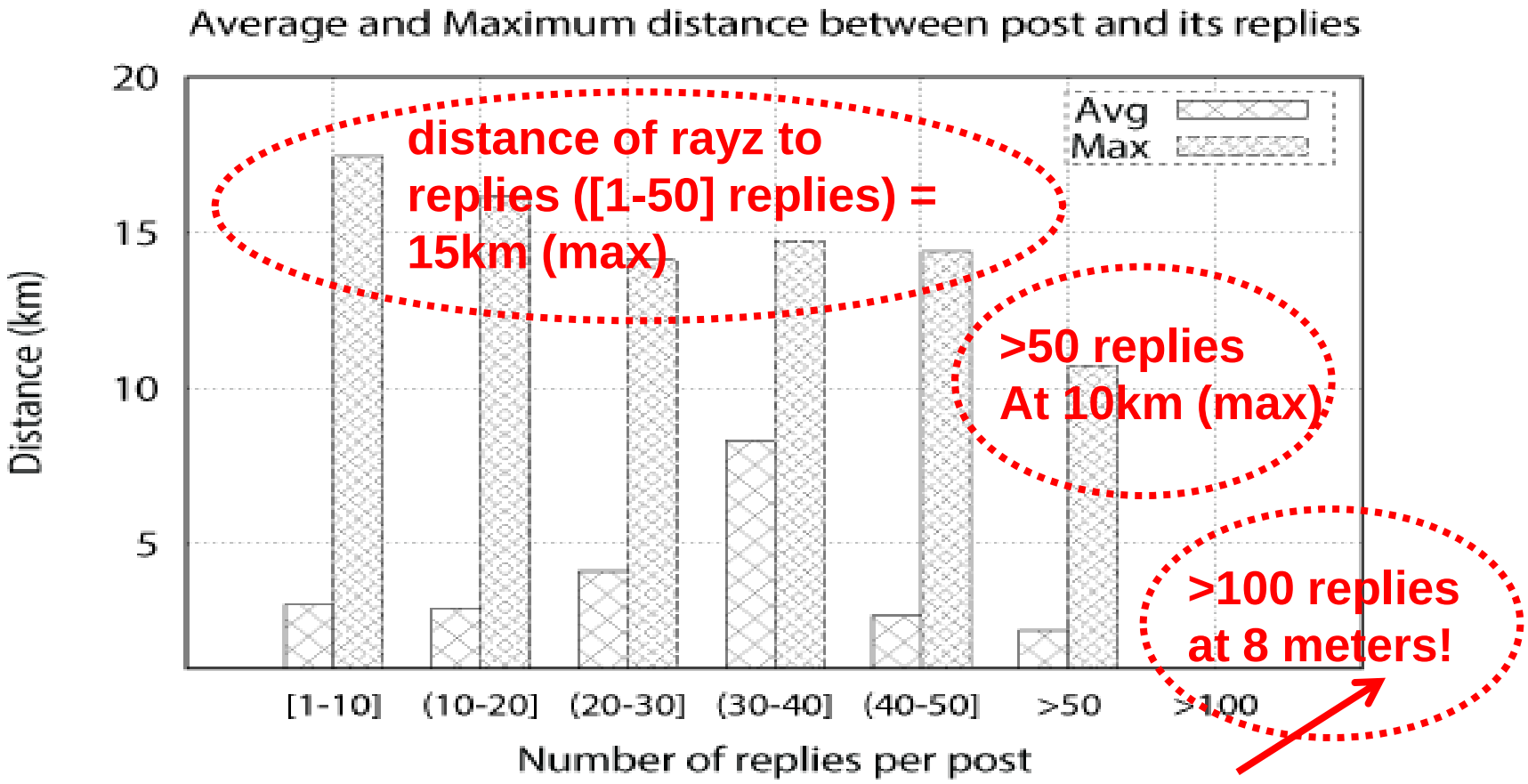
- Rayzit User Map



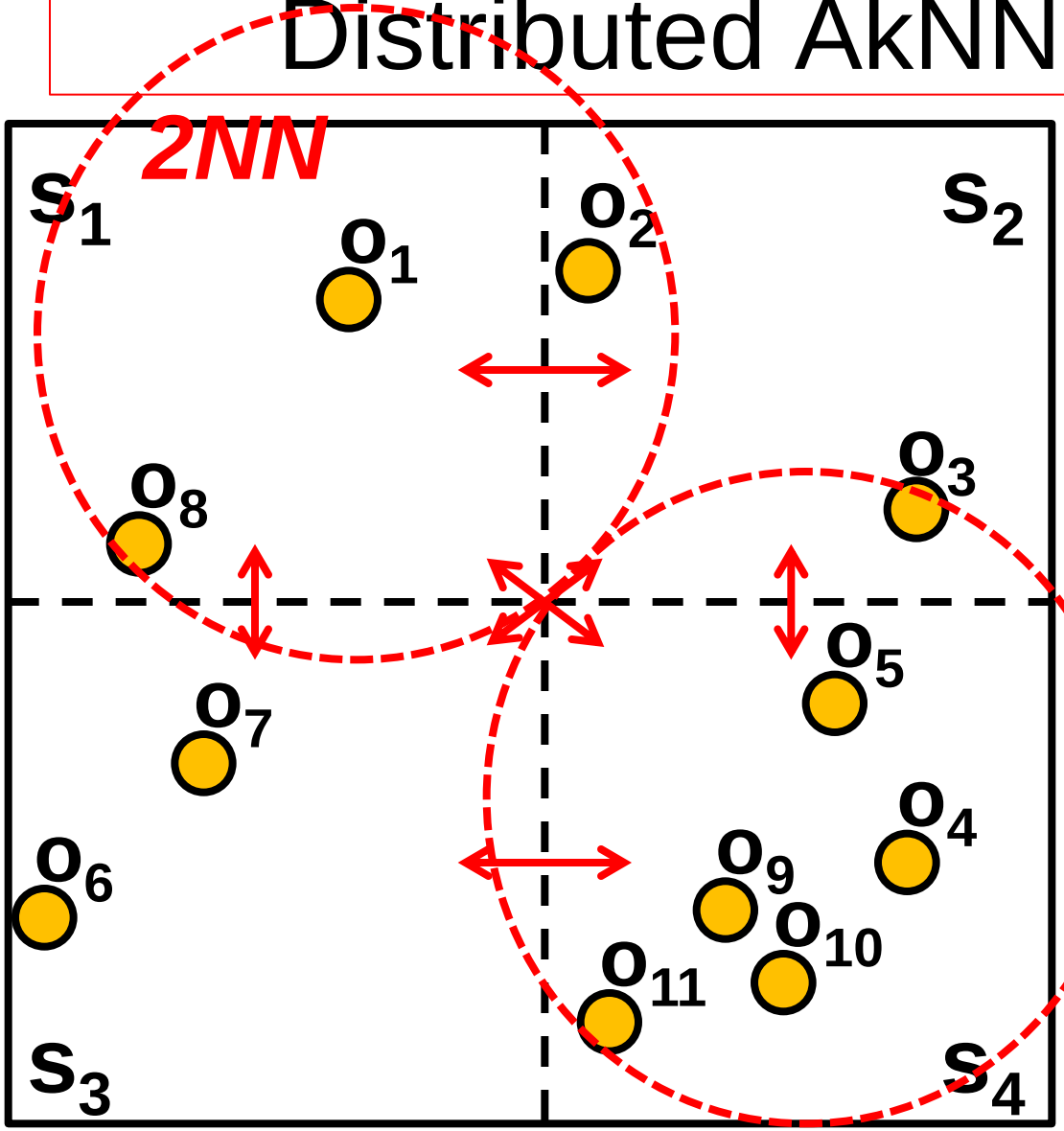
"Rayzit: An Anonymous and Dynamic Crowd Messaging Architecture", Constantinos Costa, Chrysovalantis Anastasiou, Georgios Chatzimilioudis and Demetrios Zeinalipour-Yazti, In IEEE Mobisocial 2015.

Average and Maximum Distance (Rayz, Reply)

- **Hypothesis: Closer Geographic Users means More Replies**



Distributed AkNN Problem



Let us partition the **11 objects** over **4 servers** $\{s_1,\ldots,s_4\}$

Problems:

A) Communication Overhead

- ***o_1:*** *2NN(o_1) are located on adjacent nodes s1 and s2*

=> We need to minimize the replication (f) among nodes!

B) Load Balancing

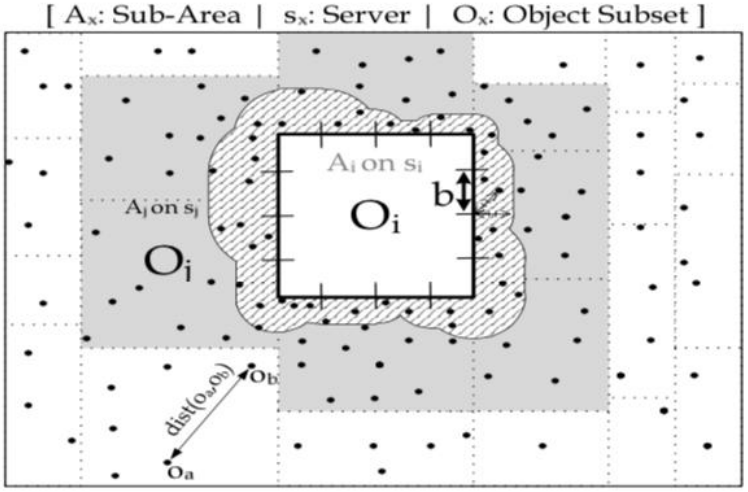
- ***s_4:*** *15 distances among 6 objects (i.e., n(n-1)/2)*
- ***s_3:*** *only 3 distances!*

=> We need to yield a fair space partitioning!

"Distributed In-Memory Processing of All k Nearest Neighbor Queries", G. Chatzimilioudis, Constantinos Costa, Demetrios Zeinalipour-Yazti, Wang-Chien Lee and Evaggelia Pitoura**, IEEE TKDE'16,** IEEE Computer Society, Vol. 28, Iss. 4, pp. 925-938, 2016.

The Spitfire Algorithm

- **Spitfire Execution Phases**
  - Partition problem space (each cell at least k)
  - Replicate border cases minimally.
  - Perform local AkNN computation



Distributed Algorithm implemented in MPI

"Distributed In-Memory Processing of All k Nearest Neighbor Queries", G. Chatzimilioudis, Constantinos Costa, Demetrios Zeinalipour-Yazti, Wang-Chien Lee and Evaggelia Pitoura, IEEE TKDE'16, IEEE Computer Society, Vol. 28, Iss. 4, pp. 925-938, 2016.

Other Hadoop AkNN

- Bibliography included a few new distributed algorithms founded on the MR programming model.

Algorithms for Distributed Main-Memory AkNN Queries

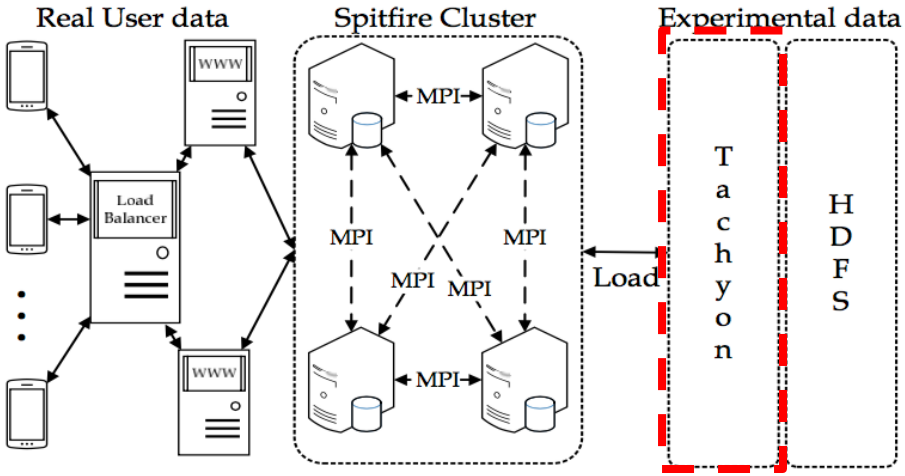
[n: objects | m: servers | f: replication factor | $f << m < n$]

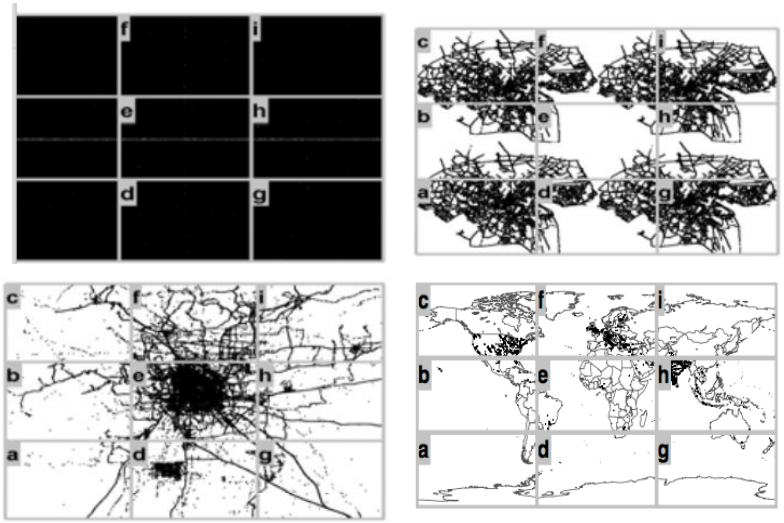
Algorithm	Preproc.	Part. & Repl.	Refinement	Communic.
H-NJ [16]	-	$O(n)$	$O(\frac{n^2}{m})$	$O(mn)$
H-BNLJ [28]	-	$O(n)$	$O(\frac{n^2}{m})$	$O(\sqrt{m}n)$
H-BRJ [28]	-	$O(n)$	$O(\frac{n}{\sqrt{m}}\log\frac{n}{\sqrt{m}})$	$O(\sqrt{m}n)$
PGBJ [16]	$O(\sqrt{n})$	$O(n^{1.5}/m)$	$O(f_{\text{PGBJ}}\frac{n^2}{m^2})$	$O(f_{\text{PGBJ}}n)$
Spitfire	-	$O(n)$	$O(f_{Spitfire}\frac{n^2}{m^2})$	$O(f_{Spitfire}n)$

"Distributed In-Memory Processing of All k Nearest Neighbor Queries", G. Chatzimilioudis, Constantinos Costa, Demetrios Zeinalipour-Yazti, Wang-Chien Lee and Evaggelia Pitoura, IEEE TKDE'16, IEEE Computer Society, Vol. 28, Iss. 4, pp. 925-938, 2016.

Evaluation Testbed

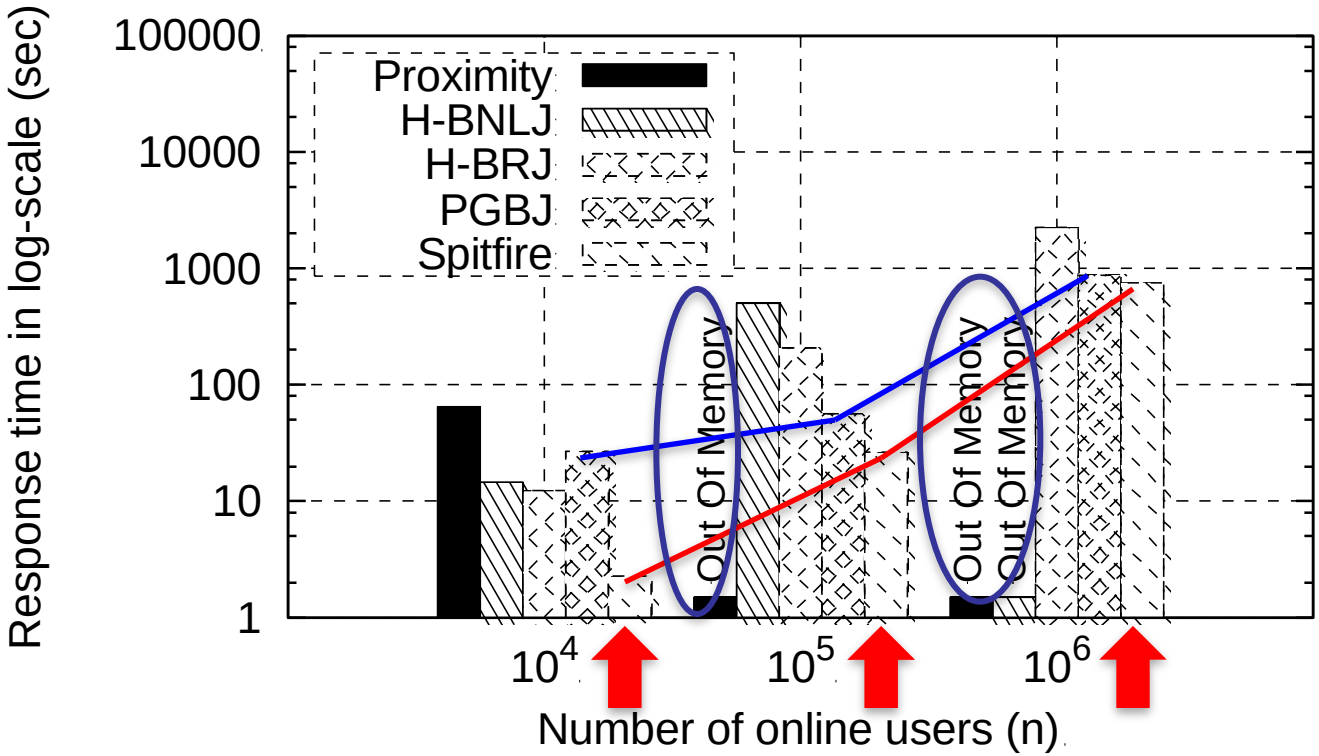
- 9 computing nodes
- 8 GB RAM
- 2 CPUs@2.40GHz
- MapReduce overTachyon (memory-centric file system)
- Datasets:
  - Random (synthetic) – uniform distribution
  - Oldenburg (realistic) – skewed distribution
  - Geolife (realistic) – very skewed distribution
  - Rayzit (real) – skewed distribution up to 2*104 users





Experimental Evaluation

GEOLIFE: Total Computation for varying number of users (**k=64, m=9**)

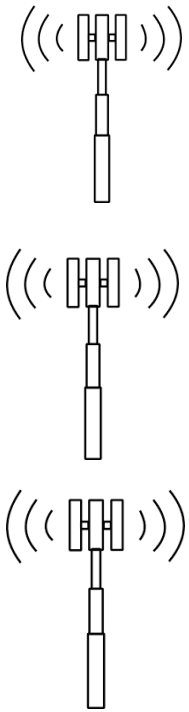


- Spitfire outperforms all other algorithms in all cases.
- Plot for most skewed dataset. Others have better results for Spitfire (67%, 75%, 14% better than PGBJ on Random, Oldenburg, Geolife).
- Spitfire and PGBJ scale better than H-BNLJ and H-BRJ that run out of memory

Talk Outline

- Spatial Indoor Data – Anyplace
  - Proposal, Papers, Software, Users, Community
- Spatial Social Data – Rayzit
  - Proposal, Publications, Software, Users,
- Spatial Telco Data – Spate
  - Proposal, Publications, Software,
- Spatial Green Data – GreenCharge
  - Proposal

Motivation

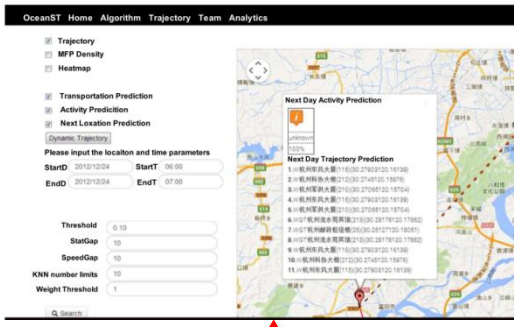


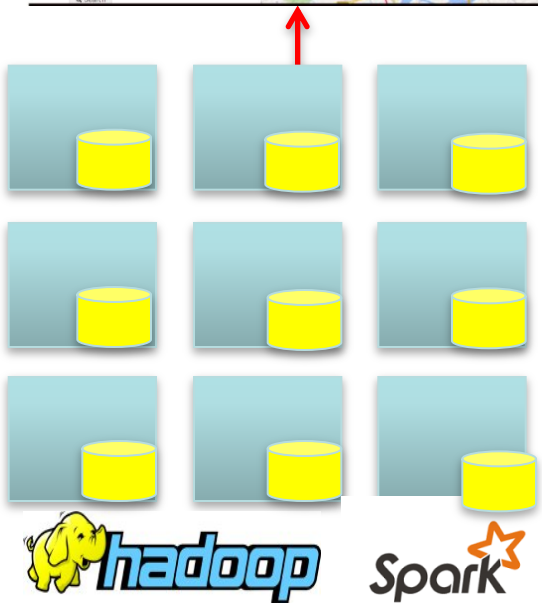
Mobile Broadband Data (MBB)

`(userID, productID, fromNo, deviceID, call drops, bandwidth …)`

5TB MBB / day (Shenzhen, China 10M Huawei customers)

Visual Analytics, Predictions, etc.





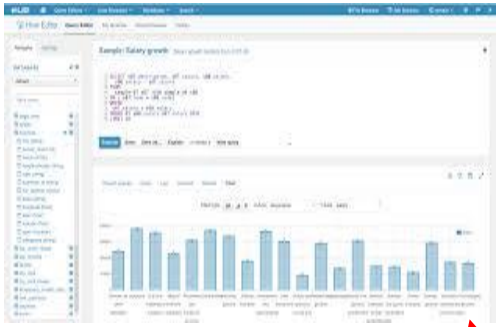
Queries

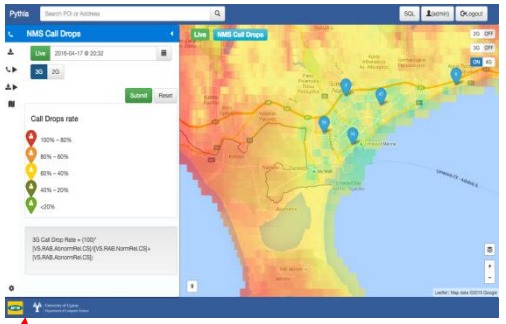
- **Telco:**
  - Where should we install our next Mobile Telecom equipment to increase coverage (e.g., identifying dropped lines from raw logs, etc.)?
  - How satisfied are my customers?
  - Which customers will churn away?
- **Government:** Which are the most congested routes over the last 1 year?
- **Emergency:** Where are our customers after an earthquake?
- **Advertising:** From which advertising panel are most customers passing by in a certain time period.

Some Papers

- **Analytics:**
  - "Telco Churn Prediction with Big Data", Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, Jia Zeng. In **ACM SIGMOD'15**.
- **Privacy:**
  - "Differential privacy in telco big data platform", Xueyang Hu, Mingxuan Yuan, Jianguo Yao, Yu Deng, Lei Chen, Qiang Yang, Haibing Guan, and Jia Zeng**. PVLDB'15.**
- **General on Spatial Big Data:**
  - Ahmed Eldawy and Mohamed F. Mokbel " The Era of Big Spatial Data". In the IEEE International Conference on Data Engineering, **ICDE 2016,** Helsinki, Finland, May, 2016.
  - A. Eldawy, M. F. Mokbel, S. Alharthi, A. Alzaidy, K. Tarek and S. Ghani, "SHAHED: A MapReduce-based system for querying and visualizing spatio-temporal satellite data," 2015 IEEE 31st International Conference on Data Engineering, Seoul, 2015, pp. 1585-1596.

SPATE Architecture





HUE SQL CLI

Time/Space GIS App

Coverage Model

Cell Towers

Once

Apache HIVE

Apache SPARK Jobs

Apache HIVE Warehouse

Apache HDFS Filesystem

NMS

CDR

Network

FTP every 30 minutes (Cron job)

Spate

Research Proposals

- Potential Consortium
  - University of Cyprus, MTN Cyprus (Telco)
  - Huawei Ireland (Outdoor Localization)
  - Beijing University of Posts and Telecommunications (BUPT)
  - …
- Confucius Institute at the University of Cyprus
  - Mobility (e.g., Internships at Huawei, China Scholarship Council)
  - Interested to facilitate EU/China Proposals - Zhenxian Wang
- Potential Calls: EU-China S&T cooperation
  - ICT-07-2017 - 5G PPP Research and Validation of critical
  - technologies and systems - 8/11/2016
  - MG-3.5-2016 Behavioral aspects for safer transport – 29/9/2016

University of Cyprus

CONFUCIUS INSTITUTE

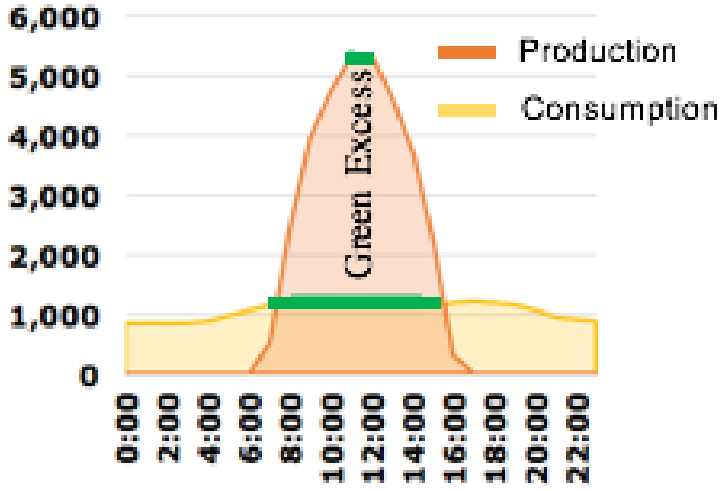
of Cyprus

Talk Outline

- Spatial Indoor Data – Anyplace
  - Proposal, Papers, Software, Users, Community
- Spatial Social Data – Rayzit
  - Proposal, Publications, Software, Users,
- Spatial Telco Data – Spate
  - Proposal, Publications, Software,
- **Spatial Green Data – GreenCharge**
  - **Proposal**

Motivation

- A predominant pillar for a **sustainable** future is energy generated from ***Renewable Energy Sources (RES),*** *such as solar PhotoVoltaic (PV), wind, hydroelectric and biomass.*
- Unfortunately, **production** and **consumption** cycles of **Prosumers** don't match leading to **"Green Excess".**
  - Green Excess Is straining regional power grids that are having a difficult time distributing fluctuating amounts of electricity.



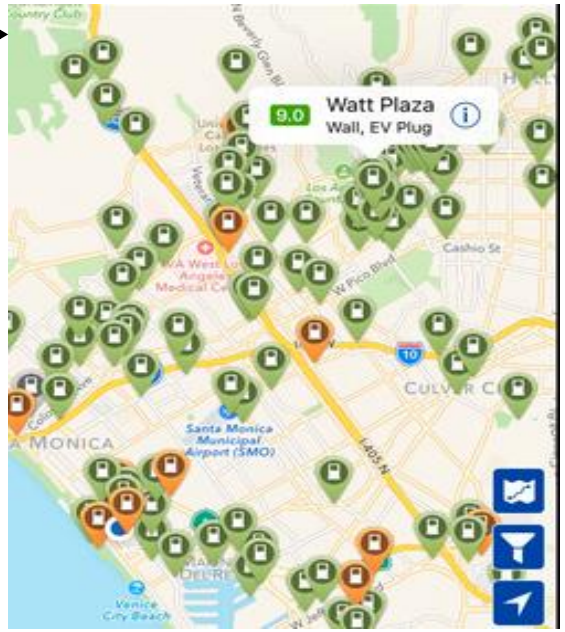
UCY Case Study

- Aims to establish itself as a **leading University** solely **relying** on **RES** to meet its electricity demands with the **10MW PV Apollo solar park.**
  - **Production:** 17GWh-per-annum electricity.
  - **Consumption:** 11GWh-per-annum electricity.
  - **Green Excess:** 11GWh-per-annum!
    - Export Renumeration: 7 cents-per-kwh
    - Import Cost: 15 cents-per-kwh.

- **Problem:** How to maximize self consumption?
- **Solution:** Leading Electric Vehicles to Green Excess with GIS.

Proposed Solution

Creation of a **Big Spatio-Temporal Information Service** that will lead electric vehicles to green energy excess.



End User Apps: Mobile, Web, etc

Intelligence: Analytics & ML, Forecasting, Smart Pairing

Big Data Storage & Processing

Mobility Data

Data at rest

Weather Data

Research Proposals

- Potential Consortium
  - Universities: UCY, Aristotle, Piraeus
  - Municipalities: Thera, etc.
  - Renewable Energy Companies
  - Renewable Energy Producers
  - Electric Car Companies
  - Car Port Companies
  - …
- Potential Calls:
  - CALL: 2016-2017 GREEN VEHICLES
  - Call identifier: H2020-GV-2016-2017

Spatial Big Data Research and Applications at UCY

Demetris Zeinalipour

Thank you – Questions?

Data Management Systems Laboratory
Department of Computer Science
University of Cyprus
http://www.cs.ucy.ac.cy/~dzeina/

The First Europe-China Workshop on Big Data Management
May 16, 2016, Kumpulan Kampus, University of Helsinki, Finland

Big Data Management

http://udbms.cs.helsinki.fi/BigData2016/