Northeastern University
Network Science Institute

Mining Smartphone and Mobility Data: Algorithms & Applications to LBSNs & Mobile Advertising

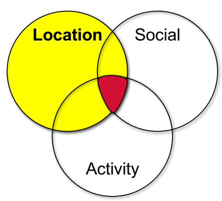
Tina Eliassi-Rad
tina@eliassi.org
@tinaeliassi

Mining Smartphone and Mobility Data

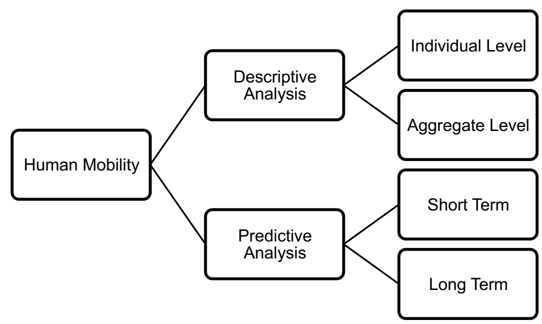
Algorithms: Location & Context

Context includes…

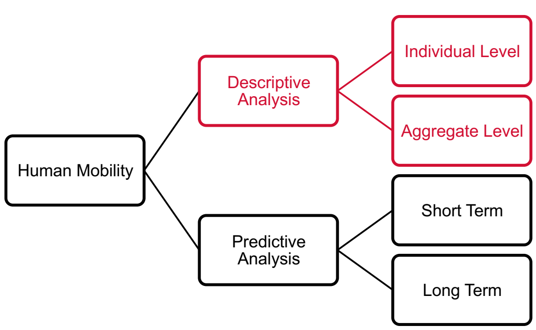
- Location/Local
  - What resources are nearby?
  - Where are you?
- Social
  - Who are you with?
- Activity
  - What are you doing?



Work in human mobility

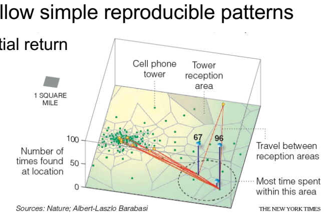


Work in human mobility



Human mobility at the individual level (descriptive)

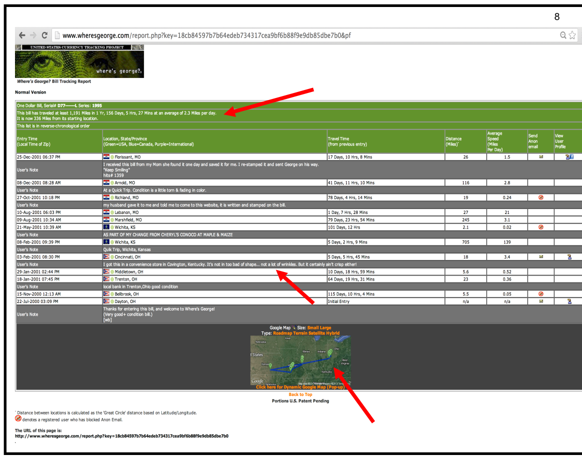
- Human trajectories are not random!
- They have high degree of temporal and spatial regularity
- Individual humans follow simple reproducible patterns
  - Exploration + preferential return
- Impact: epidemic prevention, emergency response, urban planning, ...



Marta Gonzalez, Cesar Hidalgo, Albert-Laszlo Barabasi: Understanding individual human mobility patterns. Nature 453, 779-782, 2008.

7

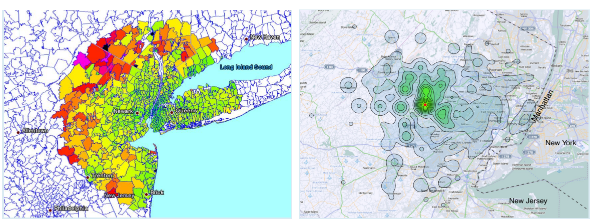
8



9

Human mobility at the aggregate level (descriptive)

- Billions of anonymized Call Detail Records (CDRs) from a cellular network
- Characterized *daily travel*, *carbon emissions*, *number of workers and event goers*, and *traffic volumes* of hundreds of thousands of people



Richard A. Becker, Ramon Caceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, Chris Volinsky: Human mobility characterization from cellular network data. Commun. ACM 56(1): 74-82 (2013).

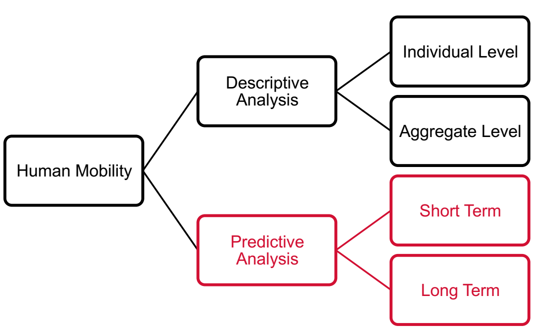
10

Human mobility (descriptive)

- M. Kim, D. Kotz, S. Kim: Extracting a mobility model from real user traces. In InfoCom 2006:1-13
- K. Lee, S. Hong, S. Kim, I. Rhee, S. Chong: Slaw: A new mobility model for human walks. In InfoCom 2009, 855-863
- Z. Li, B. Ding, J. Han, R. Kays, P. Nye: Mining periodic behaviors for moving objects. In KDD 2010, 1099-1108
- M. Kim, D. Kotz, D: Identifying unusual days. Journal of Computing Science and Engineering 5(1), 2011:71-84
- …

11

Work in human mobility



12

Human mobility at the individual level (predictive; short-term; GPS)

- "Where are you going to be in the next hour?"
  - Successful techniques: hidden Markov models, random walk based formalisms
  - Performance around 3-5 km off; classification accuracy low 90%
  - Learning from GPS alone
    - D. Ashbrook, T. Starner: Using GPS to learn significant locations and predict movement across multiple users. Personal Ubiquitous Comput. 7, 2003:275-286.
    - L. Liao, D. Fox, H. Kautz: Location-based activity recognition using relational Markov networks. In IJCAI 2005.
    - J. Krumm, E. Horvitz: Predestination: Inferring destinations from partial trajectories. In UbiComp 2006: 243-260.
    - B. Ziebart, A. Maas, A. Dey, J. Bagnell: Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In UbiComp 2008:322-331.
    - …

Slide 13

Human mobility at the individual level (predictive; long-term; GPS)

- “Where are you going to be 285 days from now at 2PM?”
- FarOut
  - Identifies periodicity via Fourier analysis (mapping time to frequency)
  - Uses PCA for pattern extraction
  - Utilizes PCA-based classification
- Performance continuous rep.: 1 km off; baseline 2.5km off
- Performance discrete rep: 80% accuracy up to 80 weeks into the future; baseline ~60%

Adam Sadilek & John Krumm: Far Out: Predicting Long-Term Human Mobility. AAAI 2012.

Slide 14

Human mobility at the individual level (predictive; long-term; GPS)

- Data: 32K days worth of GPS data across 703 subjects (½ people; ½ cars)
- High variance in area across subjects
  - From 30 to more than 10^8 km^2
  - Surface area of earth = 5.2×10^8 km^2
- Number of contiguous days = 7 to 1247
  - $\mu = 45.9$; $\sigma = 117.8$
- Captures both continuous (raw GPS) and discretized (triangular cells) data
- Each subject has a matrix D, where each row is a day.

Slide 15

Human mobility at the individual level (predictive; long-term; GPS)

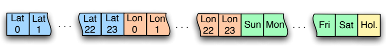


Figure 3: Our continuous vector representation of a day $\boldsymbol{d}$ consists of the median latitude and longitude for each hour of the day (00:00 through 23:59), binary encoding of the day of week, and a binary feature signifying whether a national holiday falls on $\boldsymbol{d}$.

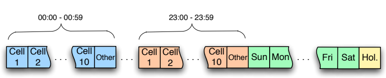


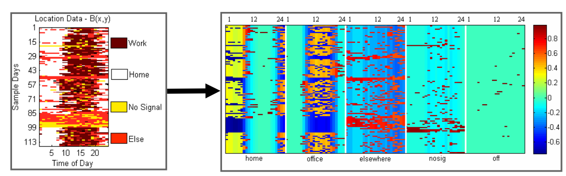
Figure 4: Our cell-based vector representation of a day $\boldsymbol{d}$ encodes the probability distribution over dominant cells conditioned on the time within $\boldsymbol{d}$, and the same day-of-week and holiday information as the continuous representation (last 8 elements).

Adam Sadilek & John Krumm: Far Out: Predicting Long-Term Human Mobility. AAAI 2012.

Slide 16

Human mobility at the individual level (predictive; long-term but < 24 hours)

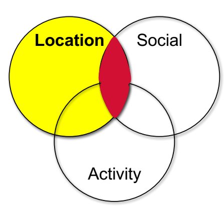
- N. Eagle and A. Pentland: Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology & Sociobiology* 63(7), 2009:1057-1066
  - Predictions up to 12 hours into the future
  - Class labels: {Home, Elsewhere, Work, No Signal, Off}.
  - PCA-based classification
  - 79% accuracy



Slide 17

Context includes…

- Location/Local
  - What resources are nearby?
  - Where are you?
- Social
  - Who are you with?
- Activity
  - What are you doing?



Slide 18

Location & Social

1. Nathan Eagle and Alex (Sandy) Pentland. 2006. Reality mining: sensing complex social systems. Personal Ubiquitous Comput. 10, 4 (March 2006), 255-268.
2. Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. KDD 2011, 1082-1090.
3. Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, Cecilia Mascolo: Socio-spatial properties of online location-based social networks. ICWSM 2011.
4. Salvatore Scellato, Anastasios Noulas, Cecilia Mascolo: Exploiting place features in link prediction on location-based social networks. KDD 2011: 1046-1054.
5. Yu Zheng: Location-based social networks - users. In Computing with Spatial Trajectories, Chapter 8, Eds: Yu Zheng, Xiaofang Zhou, Springer, 2011.
6. Yu Zheng, Xing Xie: Location-based social networks - location. In Computing with Spatial Trajectories, Chapter 9, Eds: Yu Zheng, Xiaofang Zhou, Springer, 2011.
7. Huiji Gao, Jiliang Tang, Huan Liu: gSCorr: Modeling geo-social correlations for new check-ins on location-based social networks. CIKM 2012: 1582-1586.
8. Huiji Gao, Jiliang Tang, Huan Liu: Exploring social-historical ties on location-based social networks. ICWSM 2012.
9. Chloë Brown, Vincenzo Nicosia, Salvatore Scellato, Anastasios Noulas, Cecilia Mascolo: Where online friends meet: Social communities in location-based networks. ICWSM 2012.
10. Miltiadis Allamanis, Salvatore Scellato, Cecilia Mascolo: Evolution of a location-based online social network: analysis and models. Internet Measurement Conference 2012: 145-158.
11. M. Domenico, A. Lima, and M. Musolesi. Interdependence and predictability of human mobility and social interactions. In *Nokia Mobile Data Challenge: http://research.nokia.com/page/12000*, MDC 2012.
12. P.A. Grabowicz, J.J. Ramasco, B. Goncalves, V.M. Eguiluz. Entangling Mobility and Interactions in Social Media. PLOS One 9(3):e92196. March 2014.
13. J. Toole, C. Herrera-Yaque, C.M. Schneider, M.C. Gonzalez. Coupling Human Mobility and Social Ties. In arXiv:1502.00690v1, February 2015.
14. …

19

Location & Social

1. Nathan Eagle and Alex (Sandy) Pentland. 2006. Reality mining: sensing complex social systems. Personal Ubiquitous Comput. 10, 4 (March 2006), 255-268.
2. Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. KDD 2011, 1082-1090.
3. Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, Cecilia Mascolo: Socio-spatial properties of online location-ba…
4. Salvatore S… location-based social netw…
5. Yu Zheng: … Eds: Yu Zheng, Xia…
6. Yu Zheng, … , Chapter 9, Eds: Yu Zh…
7. Huiji Gao, … tion-based social netw…
8. Huiji Gao, … CWSM 2012.
9. Chloë Brow… e friends meet: Social com…
10. Miltiadis Al… twork: analysis and models. Internet Measurement Conference 2012: 145-158.
11. M. Domenico, A. Lima, and M. Musolesi. Interdependence and predictability of human mobility and social interactions. In *Nokia Mobile Data Challenge: http://research.nokia.com/page/12000*, MDC 2012.
12. P.A. Grabowicz, J.J. Ramasco, B. Goncalves, V.M. Eguiluz. Entangling Mobility and Interactions in Social Media. PLOS One 9(3):e92196. March 2014.
13. J. Toole, C. Herrera-Yaque, C.M. Schneider, M.C. Gonzalez. Coupling Human Mobility and Social Ties. In arXiv:1502.00690v1, February 2015.
14. …

- Long distance movements are influenced by ties in the social network.
- This is not true for short-range movements or temporally periodic movements.

20

Location & Social

1. Nathan Eagle and Alex (Sandy) Pentland. 2006. Reality mining: sensing complex social systems. Personal Ubiquitous Comput. 10, 4 (March 2006), 255-268.
2. Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. KDD 2011, 1082-1090.
3. Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, Cecilia Mascolo: Socio-spatial properties of online location-based social networks. ICWSM 2011.
4. Salvatore Scellato, Anastasios Noulas, Cecilia Mascolo: Exploiting place features in link prediction on location-based social networks. KDD 2011: 1046-1054.
5. Yu Zheng: Location-based social networks - users. In Computing with Spatial Trajectories, Chapter 8, Eds: Yu Zheng, Xiaofang Zhou, Springer, 2011.
6. Yu Zheng, Xing Xie: Location-based social networks - location. In Computing with Spatial Trajectories, Chapter 9, Eds: Yu Zheng, Xiaofang Zhou, Springer, 2011.
7. Huiji Gao, Jilian… based social networks…
8. Huiji Gao, Jilia… M 2012.
9. Chloë Brown, V… ends meet: Social commun…
10. Miltiadis Allamanis, Salvatore Scellato, Cecilia Mascolo: Evolution of a location-based online social network: analysis and models. Internet Measurement Conference 2012: 145-158.
11. M. Domenico, A. Lima, and M. Musolesi. Interdependence and predictability of human mobility and social interactions. In *Nokia Mobile Data Challenge: http://research.nokia.com/page/12000*, MDC 2012.
12. P.A. Grabowicz, J.J. Ramasco, B. Goncalves, V.M. Eguiluz. Entangling Mobility and Interactions in Social Media. PLOS One 9(3):e92196. March 2014.
13. J. Toole, C. Herrera-Yaque, C.M. Schneider, M.C. Gonzalez. Coupling Human Mobility and Social Ties. In arXiv:1502.00690v1, February 2015.
14. …

- Social ties affect human mobility.
- Human mobility affects social ties.

21

Predicting the next check-in on a LBSN

- Predicting the next FourSquare check-in
- A supervised approach
  - M5 tree and linear regression
- Tried out a bunch of features
  - User mobility features
  - Global mobility features
  - Temporal features
- How well?
  - Much lower accuracies compared to GPS
  - Much harder task

Anastasios Noulas, Salvatore Scellato, Neal Lathia, Cecilia Mascolo: Mining User Mobility Features for Next Place Prediction in Location-Based Services. ICDM 2012: 1038-1043.

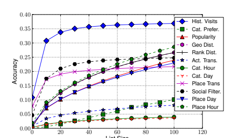


Fig. 2. Feature Predictability: Mean Accuracy for all features when they are being tested on an individual basis for different prediction list sizes N.

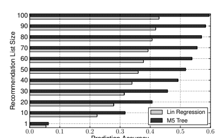


Fig. 4. Average accuracy for the supervised learning algorithms (linear regression and M5 model trees) for different recommendation list sizes.

22

Spatial search on a LBSN (using all available data, GPS, check-ins, *etc*)

- Best models of venue search are spatiotemporal models using the mixture of Gaussians, the timeliness feature, as well as popularity combined as a linear sum of log-likelihoods

Model	P@1
Baseline (nearest by distance)	0.130
Spatial Gaussian mixture model	0.193
Spatiotemporal models	0.277

- B. Shaw, J. Shea, S. Sinha, A. Hogue. Learning to Rank for Spatiotemporal Search. WSDM 2013: 717-726.
- R. Kumar, M. Mahdian, B. Pang, A. Tomkins, S. Vassilvitskii. Driven by Food: Modeling Geographic Choice. WSDM 2015: 213-222.

23

Spatial search on a LBSN (using all available data, GPS, check-ins, *etc*)

- Best models of social venue search: LambdaMart with a laundry list of features:
  - Spatial score, timeliness, popularity, here now, personal history, creator, mayor, friends here now, personal history w/ time of day

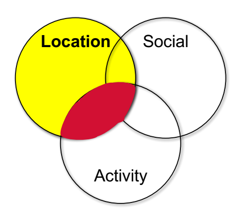
Model	P@1
Random	0.009
Spatial only	0.201
User history only	0.358
Popularity only	0.143
Linear regression: spatial + temporal	0.230
Linear regression: spatial + temporal + popularity	0.251
Linear regression: all features	0.434
Coordinate ascent: all features w/ nonlinear pairs	0.493
LambdaMART: all features	0.531

- B. Shaw, J. Shea, S. Sinha, A. Hogue. Learning to Rank for Spatiotemporal Search. WSDM 2013: 717-726.
- Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. Information Retrieval, 13(3):254–270, June 2010.

24

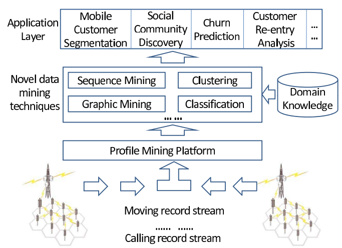
Context includes…

- Location/Local
  - What resources are nearby?
  - Where are you?
- Social
  - Who are you with?
- Activity
  - What are you doing?



Location & Activity

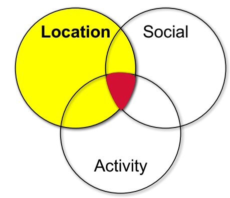
- MobileMiner [Wang et al. SIGMOD 2009]



Tengjiao Wang, Bishan Yang, Jun Gao, Dongqing Yang, Shiwei Tang, Haoyu Wu, Kedong Liu, Jian Pei: MobileMiner: A real world case study of data mining in mobile communication. SIGMOD 2009: 1083-1086

Context includes…

- Location/Local
  - What resources are nearby?
  - Where are you?
- Social
  - Who are you with?
- Activity
  - What are you doing?



Location & Social & Activity

- Lots of work analyzing Twitter data in this space
- A 2012 best paper winner is from U. of Rochester
- *Flap* uses a dynamic Bayesian network per user to predict his/her locations given location of friends, time of day, type of day
- Experiments on over 4M tweets from users in LA and NYC
- It can correctly place a user within a 100m radius with up to 85% accuracy

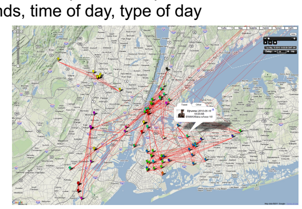
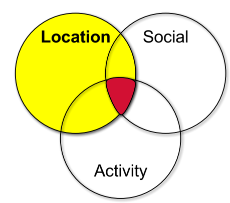


Figure 2: Flaps's visualization of a sample of geo-active friends in NYC. Red links between users represent friendships.

- Adam Sadilek, Henry A. Kautz, Jeffrey P. Bigham: Finding your friends and following them to where you are. WSDM 2012:723-732 (**best paper**)
- Adam Sadilek, Henry Kautz, Jeffrey P. Bigham: Modeling The Interplay of People's Location, Interactions, and Social Ties. IJCAI 2013.

Context includes…

- Location/Local
  - What resources are nearby?
  - Where are you?
- Social
  - Who are you with?
- Activity
  - What are you doing?



MINING SMARTPHONE AND MOBILITY DATA

Applications: Location, Ads, Privacy

Mobile Real-Time Bidding (RTB) Ad Ecosystem

businessinsider.com

31

Ad Targeting on Mobile Devices

- Need reliable location information
  - We collected 21.6M RTB requests on Wed 2/6/2013
  - The majority of them (57%) did **not** have location information
- Reasons behind the missing location information
  1. The RTB system did not forward it
  2. The SSP did not forward it
  3. The device did not capture it
  4. The user did not enable location-based services

- RTB: Real-Time Bid
- SSP: Supply Side Provider

32

Problem Definition

- *Accurately infer location information for IPs in RTB requests*
- What type of location information should we infer?
  - Latitude, longitude
  - Census Block Groups (CBGs)
  - Zip codes
  - …

privacy-friendly location Information

Long T. Le, Tina Eliassi-Rad, Foster Provost, Lauren Moores: Hyperlocal: Inferring location of IP addresses in real-time bid requests for mobile ads. *ACM SIGSPATIAL LBSN* 2013: 24-33.

33

Census Block Groups (CBGs)

- Assumption
  - *CBGs comprise location information fine-grained enough for useful hyper-local ad targeting, yet coarse-grained enough to avoid major privacy concerns.*
- Why is this reasonable?
  - Covers a contiguous area
  - Never crosses state or county boundaries
  - Contains between 600 and 3,000 people
  - US is divided into ~212K CBGs

- US is divided into ~8.2M CBs
- US is divided into 43K zip codes

34

Geo-locating IP addresses on mobile networks is hard

- Balakrishnan et al. (IMC 2009) examined properties of cell-phone IP addresses
- Mobile IPs are ephemeral and their addresses are itinerant
- Example: An individual cell phone can report different IP addresses to various servers within a short time-period

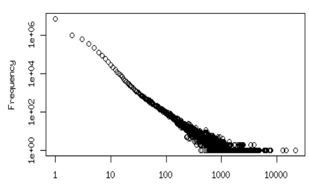
Answers to IP → Location queries provided by 7 geo-location services; the actual cell phone is in Mountain View, CA.

M. Balakrishnan, I. Mohomed, and V. Ramasubramanian. Where's that phone? Geolocating IP addresses on 3G networks. In IMC, pages 294–300, 2009.

35

Many devices often use the same public IP address

- Metwally & Paduano (KDD 2011) estimated the number of users of an IP address by keeping track of the application-specific traffic
- The primary goal of their work was to combat "abusive" traffic (such as DDoS attacks, ad click fraud and email spam) without violating the user privacy



Estimated Size of 10M Random IPs

A. Metwally and M. Paduano. Estimating the number of users behind IP addresses for combating abusive traffic. In KDD, pages 249–257, 2011.

36

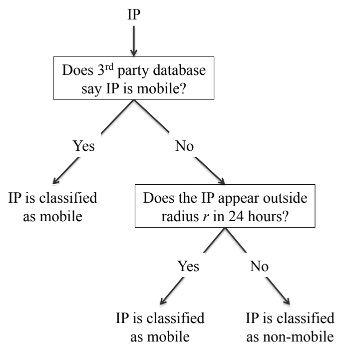
Hyperlocal: A Graph Mining Solution*

1. Classify IPs as mobile vs. non-mobile
2. Construct a movement graph with mobile and non-mobile IP nodes
3. Use a local relational classifier on each unknown node to infer latitude and longitude
4. Assign Census Block Group (CBG) ID to the inferred latitude and longitude using a *k*-nearest neighbor approach

* http://eliassi.org/ESM2013TR.pdf

37

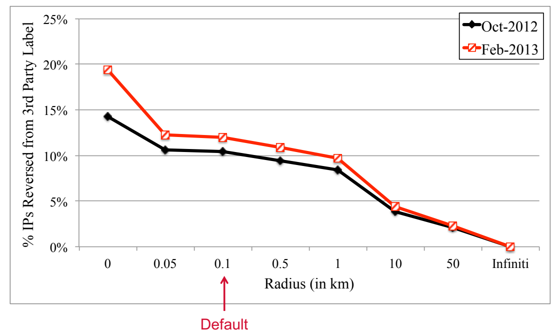
1. Classifying IPs as Mobile vs. Non-mobile



- Mobile IPs tend to change position more quickly than non-mobile IPs
- Default r = 100m (range of current Wi-Fi routers)

38

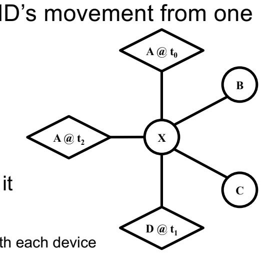
Reversing Third Party Decision



39

2. Constructing IP × IP Movement Graph

- Circle nodes are non-mobile IPs
- Diamond nodes are mobile IPs
  - Mobile IPs are time-stamped because they are transient
- An edge indicates an NUID's movement from one IP to another
  - Each edge stores
    - # of movements between its endpoints, and
    - inter-arrival times (IATs) for all movements across it



NUID: A network exchange ID associated with each device

40

3. Employing Local Relational Classifiers

- Why local?
  - The farther out one moves in the movement graph, the farther away one gets geographically
  - The movement graph is big ➔ non-local approaches can be computationally burdensome
- Local relational classifier used: *wvRN*
  - *wvRN* stands for weighted-vote Relational Neighbor [Macskassy & Provost, JMLR 2007]
  - *wvRN* estimates the class membership probabilities and assumes homophily in the network data

41

What weights should we use in *wvRN*?

- Number of movements
  - *Intuition: A node v will be closer in distance to its neighbors with whom it has more movements*
  - The data contains many edges with only one movement
- Minimum IAT
  - *Intuition: The longer the IAT, the longer distance the user has potentially moved (sans traffic)*
  - Uses the normalized minimum IAT between two IPs
  - Weight on the movement between node v and its neighbor i is

$$w_i = \frac{minIAT_v}{min(t), \forall t \in IAT(v,i)}$$

    - $minIAT_v$ is the minimum IAT across all of a given node v's edge

42

wvRN Equations

- Inputs
  - Node v
  - Its neighbors $Nbr(v)$
  - Weights W on the edges between v and its neighbors
- *wvRN* equations

$$latitude(v) = \frac{\sum_{i \in Nbr(v)} w_i \times latitude(i)}{\sum_{i \in Nbr(v)} w_i}$$

$$longitude(v) = \frac{\sum_{i \in Nbr(v)} w_i \times longitude(i)}{\sum_{i \in Nbr(v)} w_i}$$

Restricting IATs on IPs with 1 Known Neighbor

- IP × IP movement graph has a skewed distribution
  - Many nodes have only one neighbor
- Put a constraint on the IAT of IPs with only one neighbor
  - This effectively prunes the noisy links from our graph
  - It also reduces the size of the inference set

4. Assigning CBGs as Proxies for Location

- Infer location of a hashed public IP address at the CBG level <u>and not</u> at the ⟨latitude, longitude⟩ level
- Why use CBG ?
  1. It provides a more consistent labeling (as in location) for IPs
  2. It allows incorporation of external data that uses census data such as demographics
  3. In the majority of mobile applications, this level of location information is sufficient for a successful campaign

A *k*-Nearest Neighbor Approach for Assigning a CBG ID to a ⟨lat, lon⟩

- Inputs
  - Location of interest
    - *loc* = ⟨*lat*, *lon*⟩
  - For each CBG *i* in the US,
    - *i*'s centroid: c_i = ⟨lat_i, lon_i⟩
    - *i*'s area a_i in km
- Output
  - The CBG ID that contains *loc*

- Procedure
  1. *C* = centroids of the *k* nearest CBGs to *loc*
  2. For *j* in *C*
     - Calculate the distance between *loc* & centroid of the j^{th} nearest CBG
       - d_j = distance(loc, c_j)
     - Calculate the radius of CBG corresponding to the j^{th} centroid
       - r_j = sqrt(a_j / π)
     - Calculate the ratio of distance over radius
       - $ratio_j$ = d_j / r_j
  3. Return the CBG ID corresponding to min($ratio_j$), for all *j* in *C*

Recap of *Hyperlocal*

1. Classify IPs as mobile vs. non-mobile
2. Construct a movement graph with mobile and non-mobile IP nodes
3. Use a local relational classifier on each unknown node to infer latitude and longitude
4. Assign Census Block Group (CBG) ID to the inferred latitude and longitude using a *k*-nearest neighbor approach

* http://eliassi.org/ESM2013TR.pdf

Experiments

- Experiments are divided into nine combinations of *infer location for* X *using* Y
- Values for *X* are 'all IPs', 'mobile IPs', and 'non-mobile IPs'
- Values for *Y* are 'all neighbors', 'mobile neighbors', and 'non-mobile neighbors'
- Measure accuracy by checking the predicted CBG ID vs. the actual CBG ID of an IP

Implementation & Runtime

- Hardware & OS: Macbook Pro with
  - CPU 2.66 GHz Intel Core i7
  - RAM 8 GB DDR3
  - hard drive 500 GB SSD
  - OS X 10.8
- Language: Python
- Supporting Software: NetworkX & MongoDB
- Runtime: On average 1.2 milliseconds to process each RTB request

Data

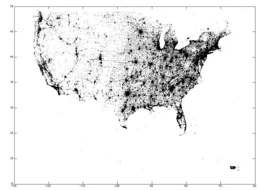
Data Name	Collection Date	# RTB Requests with Valid US NUIDs	% RTB Requests **without** Location	% RTB Requests from **Mobile** IPs
Oct-2012	Mon 10/01/2012	44.1M	36.5%	57.3%
Feb-2013	Wed 02/06/2013	21.6M	56.7%	47.7%

- From Oct-2012 to Feb-2013
  - # of RTB requests decreased by ~50%
    - Due to reductions from SSPs
  - # of requests without location information increased by ~55%
  - # of requests from mobile IPs decreased by ~17%

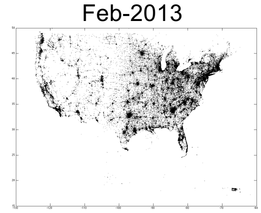
All IPs are hashed public IP addresses.

IPs on the US Map

Oct-2012

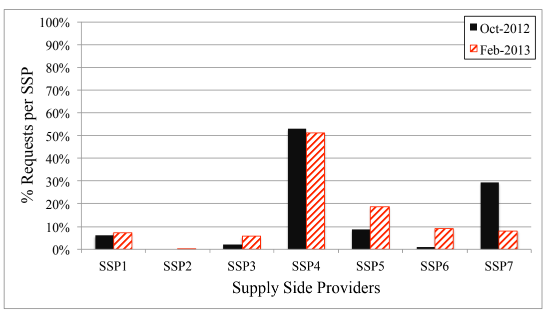


Feb-2013



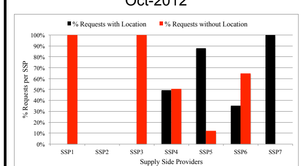
Data Characteristics per SSP

SSP4 provides about 50% of the requests for both datasets.

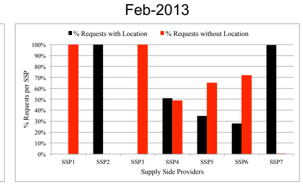


% Requests with and without Location Information per SSP

Oct-2012

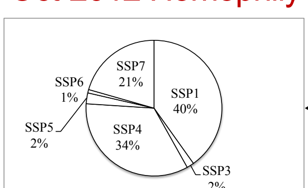


Feb-2013



- None of the requests from SSP1 and SSP3 have location information.
- All the requests from SSP7 have location information.

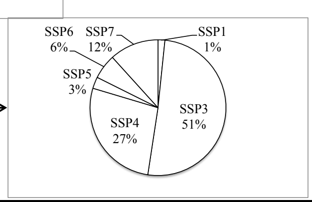
Oct-2012 Homophily per SSP



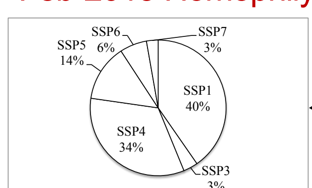
There is a considerable overlap with other SSPs.

% IPs from SSP1 requests that connect to other SSPs

% IPs from SSP3 requests that connect to other SSPs



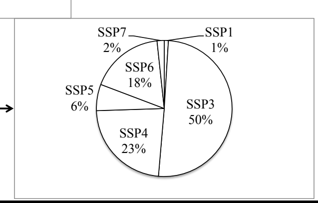
Feb-2013 Homophily per SSP



There is a considerable overlap with other SSPs.

% IPs from SSP1 requests that connect to other SSPs

% IPs from SSP3 requests that connect to other SSPs



Homophily in IP×IP Movement Graphs

$$homophily = \frac{\text{\# of movements whose IP endpoints are from the same SSP}}{\text{total \# of movements}}$$

Homophily	Oct-2012	Feb-2013
All movements	96.5%	90.8%
Mobile to mobile movements	98.6%	99.2%
Non-mobile to non-mobile movements	86.9%	78.1%

There is a considerable homophily in the IP×IP movement graph

Core Results: wvRN(minIAT) vs. wvRN(numMoves)

Infer location for **all** IPs . . .	Accuracy wvRN(minIAT) Oct-2012 | Feb-2013	Accuracy wvRN(numMoves) Oct-2012 | Feb-2013	Number of Predictions Oct-2012 | Feb-2013	% Mobile in Predictions Oct-2012 | Feb-2013
using all neighbors	71.6 | 70.8	69.7 | 69.3	1,189,679 | 328,015	91.0% | 83.8%
using mobile neighbors	**74.4** | **74.8**	**72.6** | **73.5**	1,077,644 | 278,674	93.2% | 87.4%
using non-mobile neighbors	51.9 | 57.7	51.7 | 57.3	98,338 | 40,777	68.7% | 62.1%
Infer location for **mobile** IPs . . .				
using all neighbors	74.2 | 75.0	72.3 | 73.6	1,082,566 | 274,900	100% | 100%
using mobile neighbors	**76.6** | **79.1**	**74.7** | **77.9**	1,004,601 | 243,630	100% | 100%
using non-mobile neighbors	50.4 | 54.7	50.2 | 54.5	67,553 | 25,323	100% | 100%
Infer location for **non-mobile** IPs . . .				
using all neighbors	45.5 | 49.0	44.0 | 46.8	107,113 | 53,115	0% | 0%
using mobile neighbors	44.0 | 45.1	42.5 | 43.0	73,043 | 35,044	0% | 0%
using non-mobile neighbors	**55.4** | **62.7**	**55.0** | **62.0**	30,785 | 15,454	0% | 0%

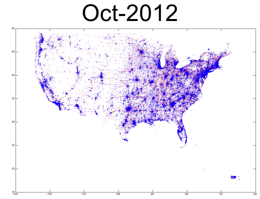
- Differences between the two methods are not statistically significant at the 0.05 level
- Number of predictions varies depending on the particular inference and the neighbor types used in the inference process

IATs on movement edges are correlated with distances

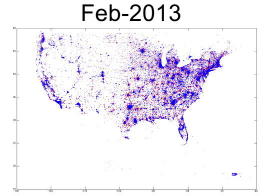
- Shorter IAT, shorter distance
- For IPs with only one known neighbor, restricting IAT to ≤ 60 minutes
  - Improves accuracy by an average of 12% on Oct-2012 and 23% on Feb-2013 data
  - Reduces the number of predictions by an average of 4 times for Oct-2012 and 5 times for Feb-2012
- Restricting IATs to > 60 minutes decrease accuracy

Correct (Blue) & Incorrect (Red) Classification

Oct-2012



Feb-2013

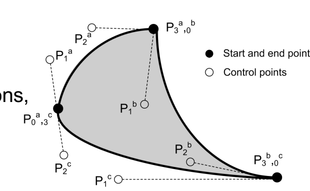


Limitations of a Graph Mining Approach

- Cannot infer location for IPs with no neighbors
  - Use other info – e.g., site visits; subnet info, etc.
- Cannot infer location for IPs with no known neighbors
  - Use collective classification.

Related Work [Wong et al. NSDI 2007]

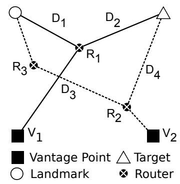
- Locate IPs by
  - representing node positions through regions,
  - expressing constraints as areas, and
  - computing locations by solving a system of geometric constraints
- Relies on pings to estimate the round-trip time between two IPs



B. Wong, I. Stoyanov, and E. G. Sirer. Octant: A comprehensive framework for the geolocalization of internet hosts. In NSDI, pages 23–23, 2007.

Related Work [Wang et al. NSDI 2011]

- A client-independent geo-location system
- Like [Wong et al. NSDI 2007] relies on pings to estimate the round-trip time between two IPs
- Also relies on landmarks, which are collected manually



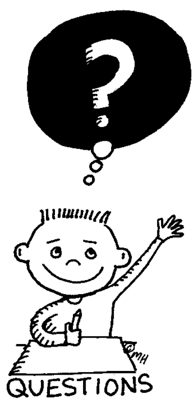
Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang. Towards street-level client-independent IP geolocation. In NSDI, pages 27–27, 2011.

Recap & Open Problems

- Graph mining on just the structure of an IP×IP movement graph to infer locations, in terms of CBGs, for hashed public IP addresses produces an accuracy of ~75%
- Results are impressive since estimating the correct CBG is out of 212K possibilities
- **Open problems**
  - Inference on truncated IP addresses
  - Constrained collective classification

Thank you

- Papers at
  - http://eliassi.org/pubs.html
- Patent: F. Provost, T. Eliassi-Rad, L. Moores, US Patent Number 9014717, issued April 21, 2015.
- Contact me at
  - tina@eliassi.org
- Supported by NSF, DTRA, DARPA, LLNL, and ESM.



Northeastern University
Network Science Institute

Mining Smartphone and Mobility Data: Algorithms & Applications to LBSNs & Mobile Advertising

Tina Eliassi-Rad
tina@eliassi.org
@tinaeliassi

License