Large-scale Social Sensing (with Humans as Sensors)

Tarek Abdelzaher
Dept. of Computer Science
University of Illinois at Urbana Champaign

Best Presentations

- Total votes received 48.
- The max possible votes per presentation: 24.

- Next: Top 4 presentations from #4 to #1…

In 4th Place

In 4th Place

- Unsupervised Activity Discovery

5 (out of a max of 24)

By:

Tianshi Wang and Siyu Bian

In 3rd Place

In 3rd Place

- Detecting Drinking Episodes

9 (out of a max of 24)

By:

Yuntae Kim, Chenhao Wu, Brian Thompson, Anirudh Madhusudan

In 2nd Place

In 2nd Place

- Smart Eye Mask

13 (out of a max of 24)

By:

Shaima AbdulMajeed and Dipali Ranjan

In 1st Place

In 1st Place

- CovertBand

15 (out of a max of 24)

62.5%!!!

By: Sheng Shen, Chong Lu

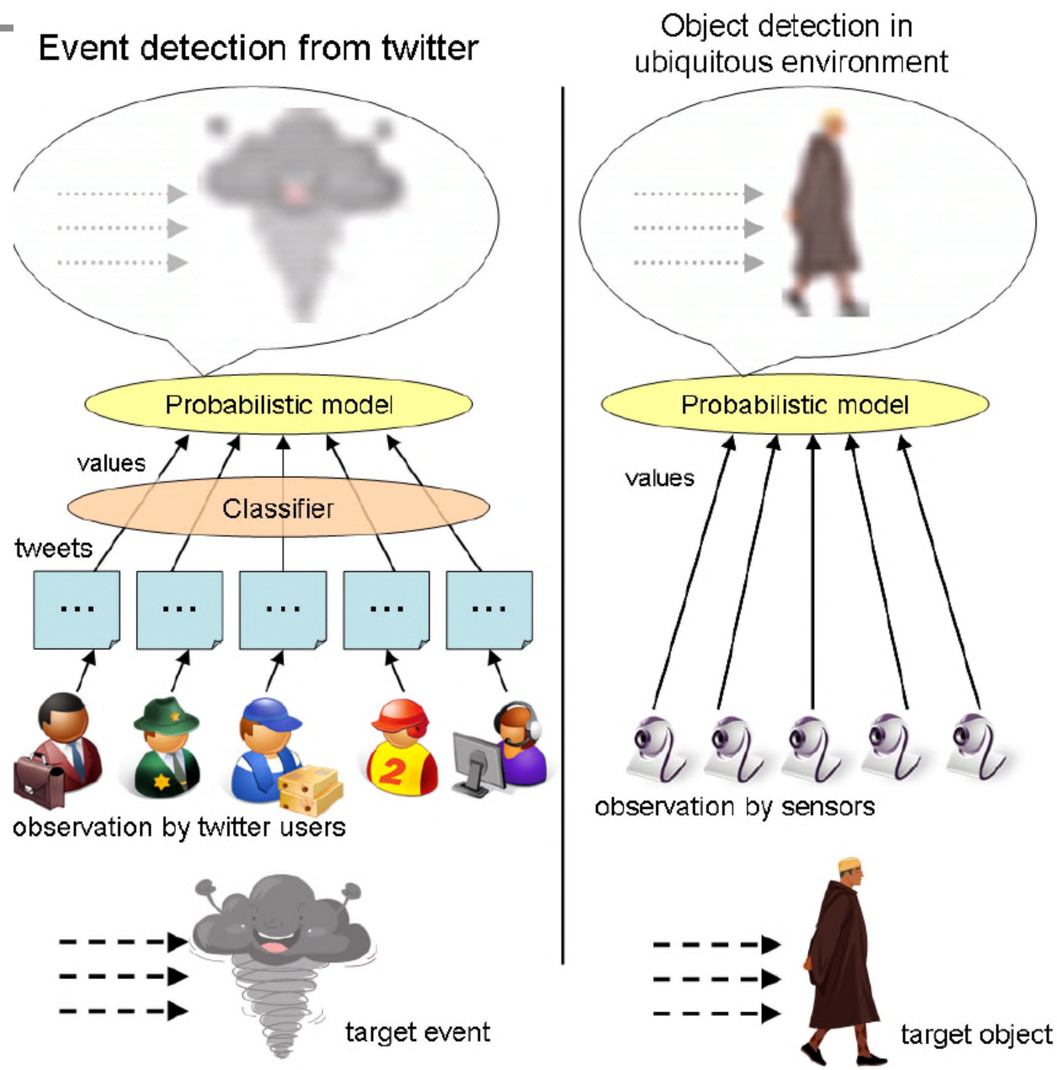
Earthquake Shakes Twitter Users

Using Humans as Sensors:

- Assumption: Each Twitter user is regarded as a sensor. A sensor detects a target event and makes a report probabilistically
- Assumption: Each tweet is associated with a time and location, which is a set of latitude and longitude

Earthquake Shakes Twitter Users

Using Humans as Sensors:

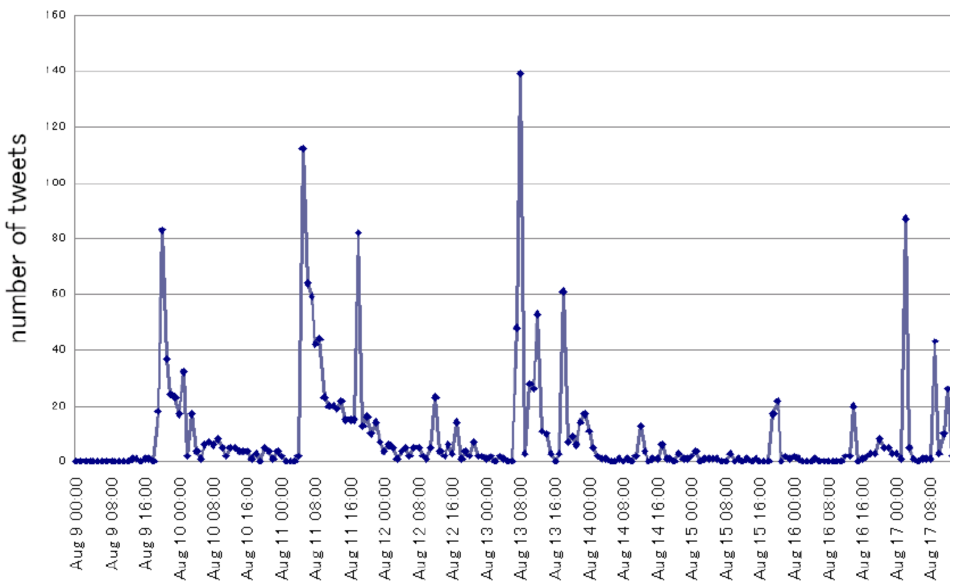


Event Detection: The Classifier

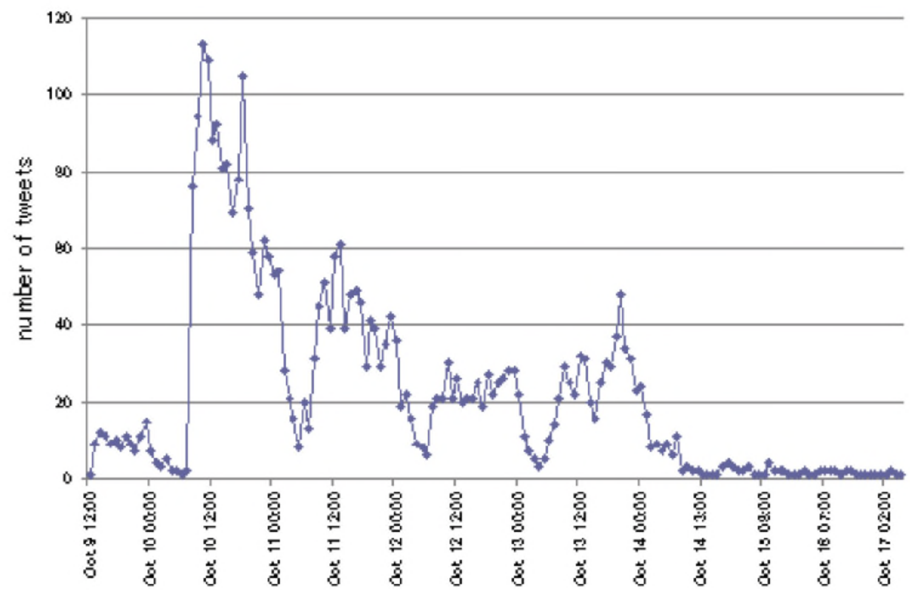
- Not all occurrences of a keyword (e.g., "earthquake" or "shaking" is about an ongoing event:
  - I am afraid of earthquakes
  - Shaking hands with boss
- How to solve this? (How to classify occurrences that constitute *"sensing" of an ongoing event* from others?)

Event Detection: A Probabilistic Model

- Spikes in occurrence of related keywords help detect corresponding events:



Earthquake-related keywords



Typhoon-related keywords

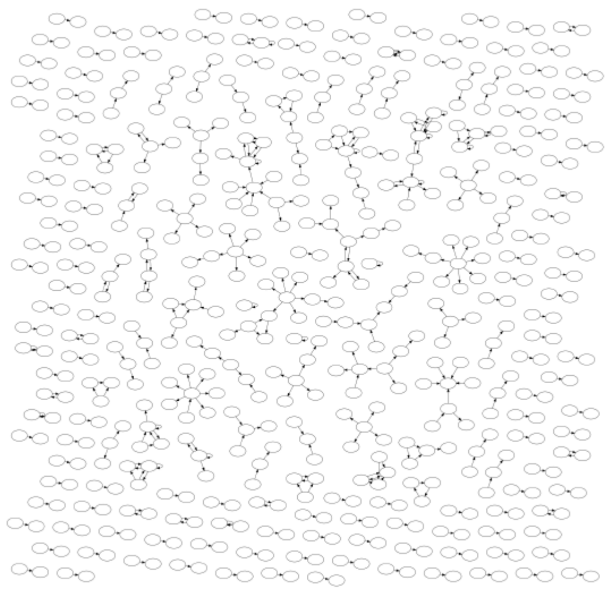
When a user detects an event at time 0, the time to make a tweet follows and exponential distribution.

Event Tracking

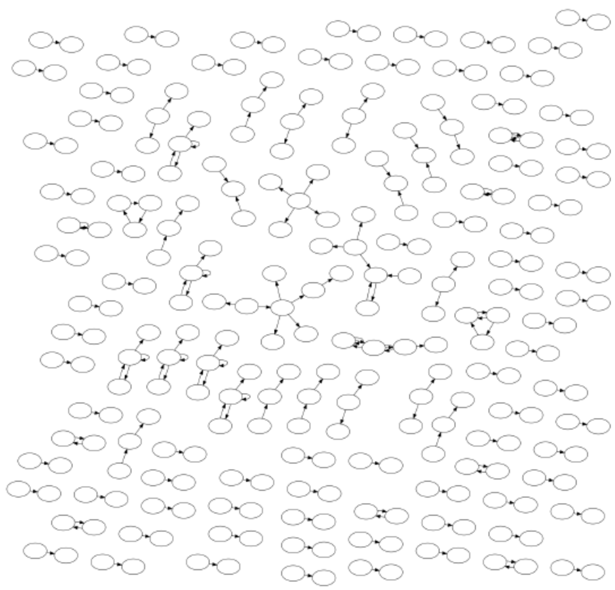
- Given (i) detected noisy location of the event at each point in time and a (ii) mobility model for the event, compute the most likely trajectory
- Multiple tracking techniques available in literature:
  - Kalman filter
  - Particle filter

Information Diffusion

- Assuming little/no diffusion (no retweets)



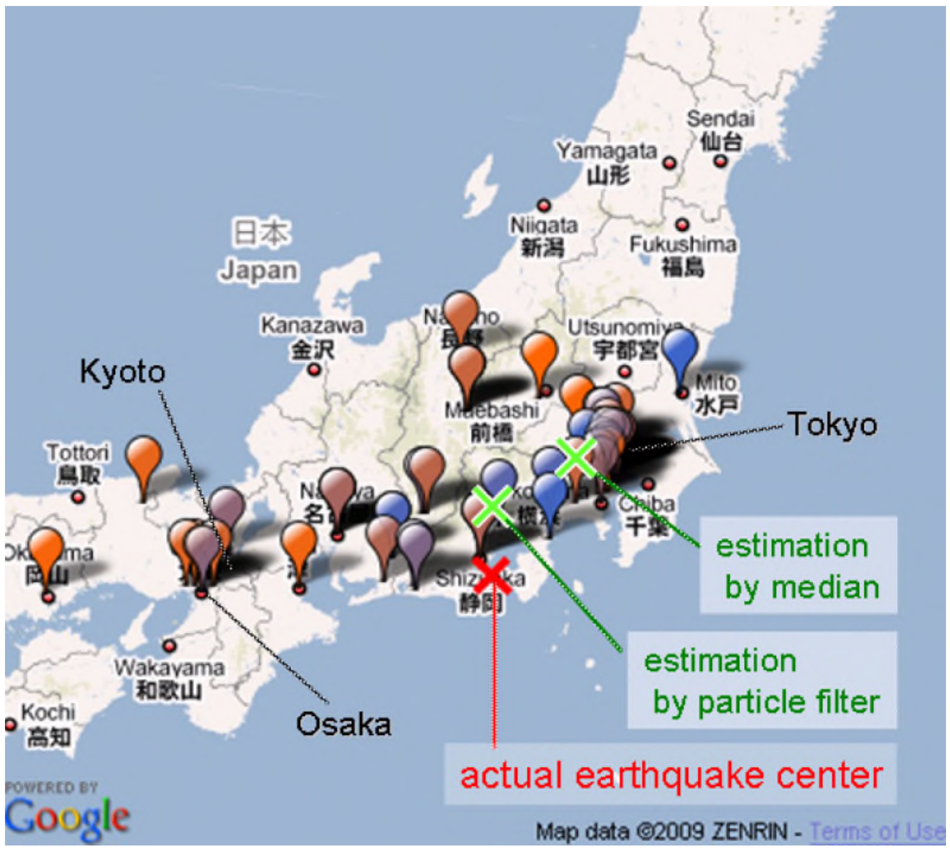
Diffusion of Earthquake tweets



Diffusion of Typhoon tweets

Evaluation

- Detection of an Earthquake



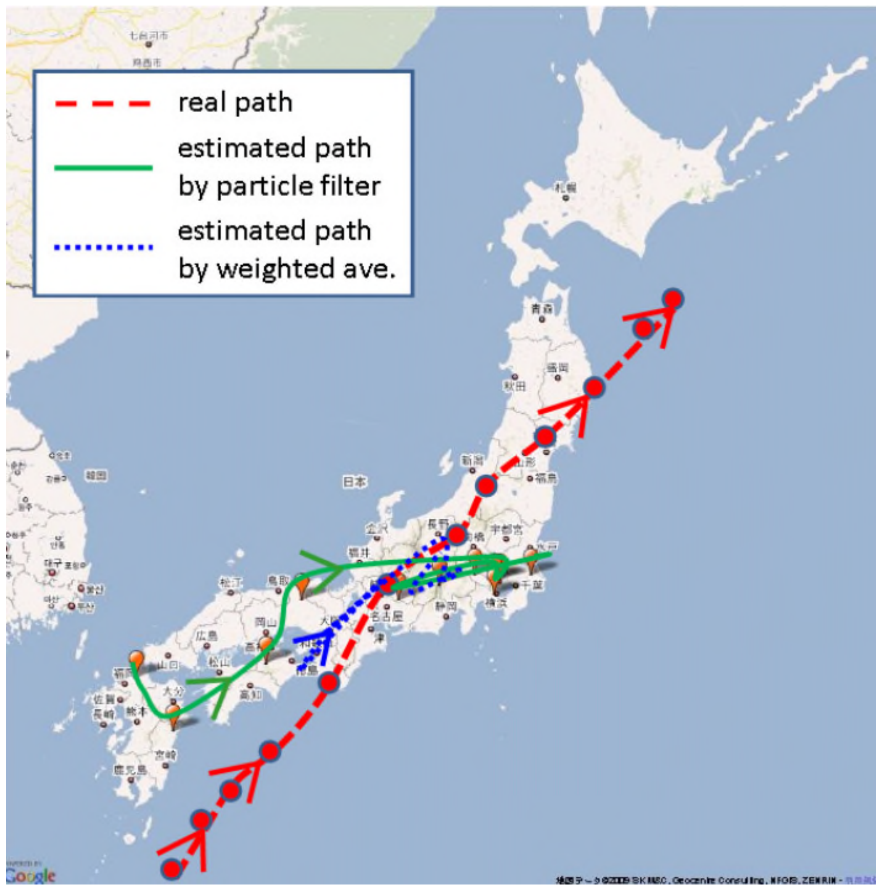
Evaluation

- Detection of an Earthquake

Date	Actual center		Median (baseline)			Weighted ave. (baseline)			Kalman filters			**Particle filters**		
	lat.	long.	lat.	long.	dist.	lat.	long.	dist.	lat.	long.	dist.	lat.	long.	dist.
Aug. 10 01:00	33.10	138.50	3.40	-0.80	3.49	2.70	-0.10	2.70	2.67	-0.50	2.72	2.60	0.50	**2.65**
Aug. 11 05:00	34.80	138.50	0.90	-0.90	1.27	0.70	-0.30	0.76	0.60	-0.20	**0.63**	0.30	-0.90	0.95
Aug. 13 07:50	33.00	140.80	1.30	-9.60	9.69	2.30	-2.30	**3.25**	1.63	-3.75	4.09	2.70	-2.70	3.82
Aug. 17 20:40	33.70	130.20	4.60	6.00	7.56	0.90	3.20	3.32	1.63	4.35	4.65	0.10	-0.80	**0.81**
Aug. 18 22:17	23.30	123.50	7.80	9.90	12.60	8.70	10.90	13.95	8.32	10.13	13.11	5.60	8.10	**9.85**
Aug. 21 08.51	35.70	140.00	0.50	-4.40	4.43	0.10	-1.00	1.00	0.00	-0.60	**0.60**	-0.80	0.48	0.93
Aug. 24 13:30	37.50	138.60	-0.40	0.00	**0.40**	-0.50	0.40	0.64	-0.50	0.30	0.58	2.40	0.70	2.50
Aug. 24 14:40	41.10	140.30	-1.90	1.10	2.20	-1.30	0.50	**1.39**	-1.50	0.50	1.58	3.10	2.00	3.69
Aug. 25 02:22	42.10	142.80	-2.90	-3.90	4.86	-6.10	-3.80	7.19	-5.20	-3.70	6.38	-1.80	-1.90	**2.62**
Aug. 25 20:19	35.40	140.40	1.60	-1.80	2.41	2.20	-0.70	2.31	0.70	-1.60	1.75	1.40	0.10	**1.40**
Aug. 31 00:46	37.20	141.50	-0.40	-3.60	3.62	-1.10	-2.30	2.55	-1.30	-2.20	2.56	-0.30	-0.30	**0.42**
Aug. 31 21:11	33.40	130.90	-4.50	-3.60	5.76	0.50	2.10	2.16	0.70	1.90	2.02	-0.20	-1.70	**1.71**
Sep. 3 22:26	31.10	130.30	6.20	-0.10	6.20	4.00	5.00	6.40	4.90	7.20	8.71	2.40	2.10	**3.19**
Sep. 4 11:30	35.80	140.10	3.10	-1.70	3.54	0.20	-0.90	**0.92**	0.00	-1.00	1.00	0.80	1.40	1.61
Sep. 05 10:59	37.00	140.20	-2.70	-8.30	8.73	-1.40	-3.10	**3.40**	-1.30	-3.30	3.55	-2.10	-5.80	6.17
Sep. 08 01:24	42.20	143.00	-3.60	-8.90	9.60	-2.50	-3.90	4.63	-4.50	-6.00	7.50	1.30	-3.60	**3.83**
Sep. 10 18:29	43.20	146.20	-5.90	-10.20	11.78	-4.90	-7.10	8.63	-4.50	-7.20	8.49	-0.90	-7.00	**7.06**
Sep. 16 21:38	33.40	130.90	1.10	-0.20	**1.12**	0.90	2.10	2.28	0.50	1.40	1.49	-0.20	-2.50	2.51
Sep. 22 20:40	47.60	141.70	-11.10	-7.50	13.40	-10.80	-3.10	11.24	-11.30	-3.80	11.92	-7.80	-3.00	**8.36**
Oct. 1 19:43	36.40	140.70	0.70	-3.80	3.86	-0.60	-1.80	1.90	-0.30	-1.50	1.53	-0.70	0.30	**0.76**
Oct. 5 09:35	42.40	141.60	-3.70	-3.10	4.83	-2.70	-2.00	3.36	-2.60	-1.60	3.05	1.10	-1.70	**2.02**
Oct. 6 07:49	35.90	137.60	0.50	1.20	1.30	-0.20	0.80	0.82	-0.10	0.90	0.91	0.30	0.50	**0.58**
Oct. 10 17:43	41.80	142.20	-3.50	-5.40	6.44	-1.40	-2.10	**2.52**	-2.20	-2.60	3.41	2.40	-1.30	2.73
Oct. 12 16:10	35.90	137.60	2.80	0.50	2.84	0.80	1.20	**1.44**	0.80	1.60	1.79	3.60	1.40	3.86
Oct. 12 18:42	37.40	139.70	-2.00	-4.40	4.83	-1.50	-0.90	1.75	-1.70	-1.40	2.20	-1.00	-0.60	**1.17**
Average distance					5.47			3.62			3.85			**3.01**

Evaluation

- Detection and tracking of a Typhoon



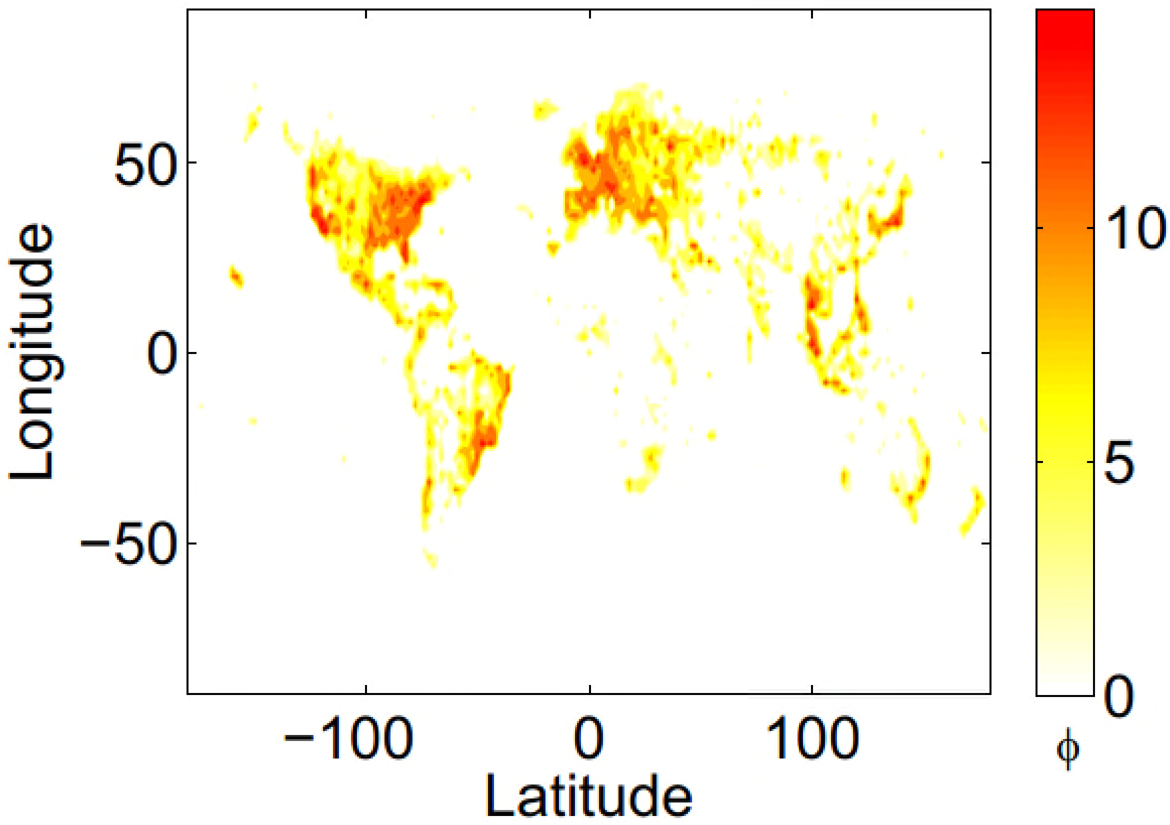
Evaluation

- Detection and tracking of a Typhoon

Date	Location		Median (baseline)			Weighted ave. (baseline)			Kalman filters			**Particle filters**		
	lat.	long.	lat.	long.	dist.	lat.	long.	dist.	lat.	long.	dist.	lat.	long.	dist.
Oct. 7 12:00	29.00	131.80	-1.90	-1.90	**2.69**	-5.20	-3.60	6.32	-3.90	-1.10	4.05	-4.70	1.10	4.83
Oct. 7 15:00	29.90	132.50	-3.70	-2.60	4.52	-3.80	-2.40	4.49	3.20	3.10	4.46	-2.70	0.90	**2.85**
Oct. 7 18:00	30.80	133.20	-4.10	-1.90	4.52	-4.40	-3.50	5.62	-6.40	5.40	8.37	-3.20	-0.70	**3.28**
Oct. 7 21:00	31.60	134.30	-3.90	-3.50	5.24	-3.60	-3.30	4.88	-10.90	-1.60	11.02	-3.70	-0.50	**3.73**
Oct. 8 0:00	32.90	135.60	-2.30	-0.10	**2.30**	-2.30	-0.90	2.47	-12.60	-20.40	23.98	-2.90	-3.50	4.55
Oct. 8 6:00	35.10	137.20	1.60	3.00	3.40	0.80	1.70	**1.88**	4.20	16.00	16.54	-0.60	-2.50	2.57
Oct. 8 9:00	36.10	138.80	-0.60	3.60	3.65	0.00	0.50	**0.50**	0.50	2.60	2.65	0.70	-0.80	1.06
Oct. 8 12:00	37.10	139.70	1.70	3.90	4.25	1.50	1.20	1.92	2.10	1.60	2.64	1.40	0.10	**1.40**
Oct. 8 15:00	38.00	140.90	2.30	3.20	3.94	2.40	2.20	**3.26**	1.70	7.60	7.79	2.40	2.70	3.61
Oct. 8 18:00	39.00	142.30	3.20	7.30	7.97	3.50	5.10	**6.19**	2.10	-18.80	18.92	3.70	5.10	6.30
Oct. 8 21:00	40.00	143.60	4.30	3.90	5.81	4.00	5.30	6.64	1.60	4.50	4.78	4.20	3.10	**5.22**
Average distance					4.39			4.02			9.56			**3.58**

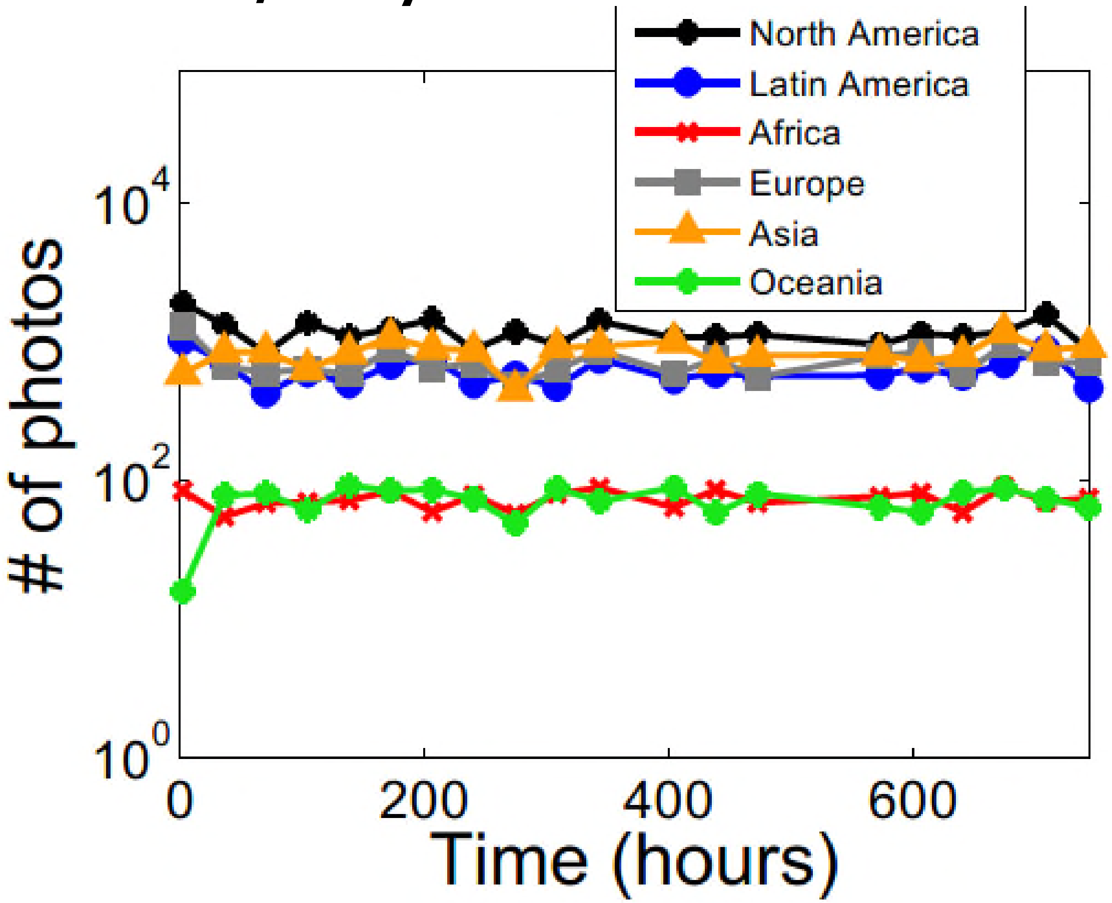
Instagram Coverage

- June/July 2012



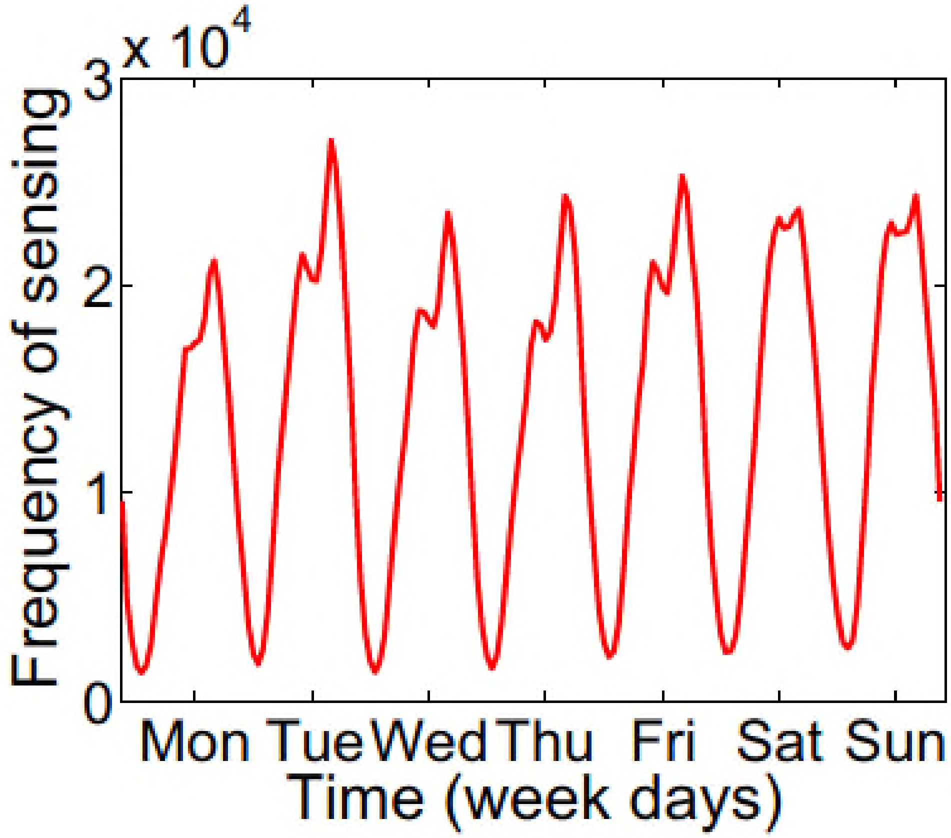
Instagram Coverage

- June/July 2012



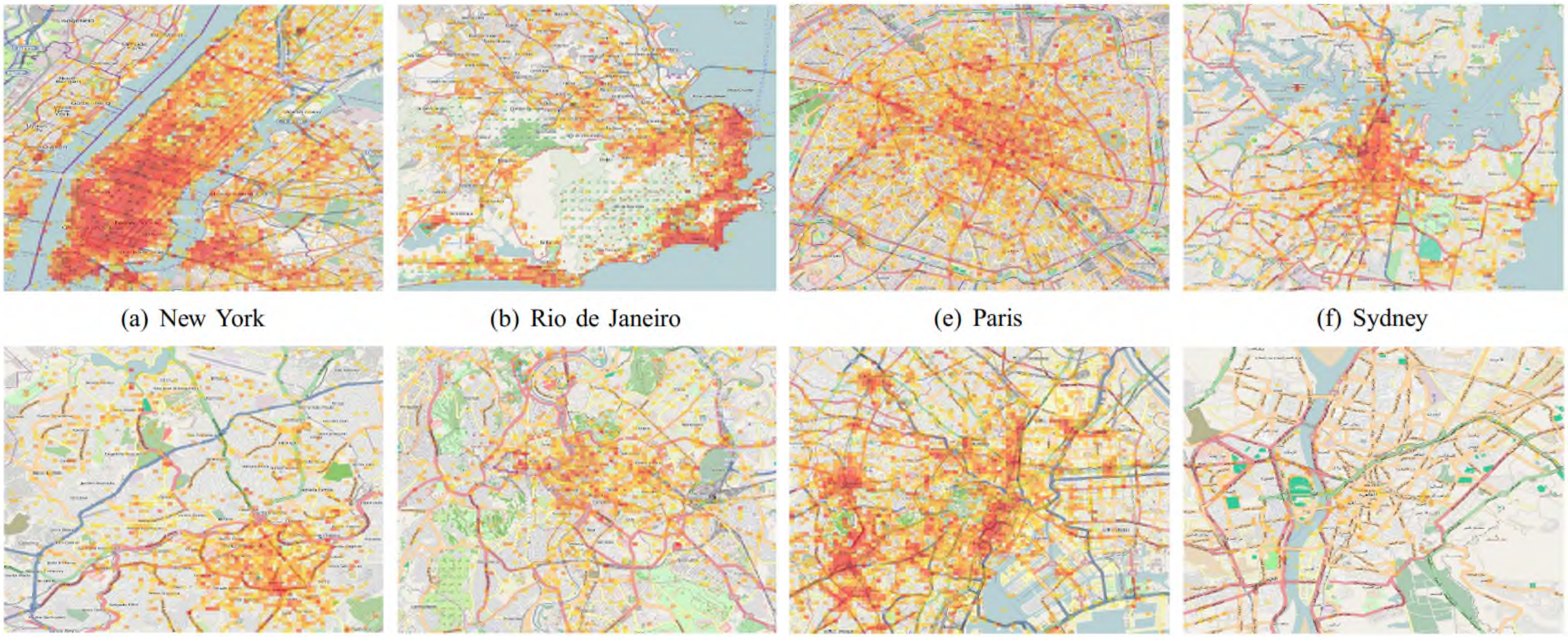
Instagram Coverage

- June/July 2012



Instagram Coverage

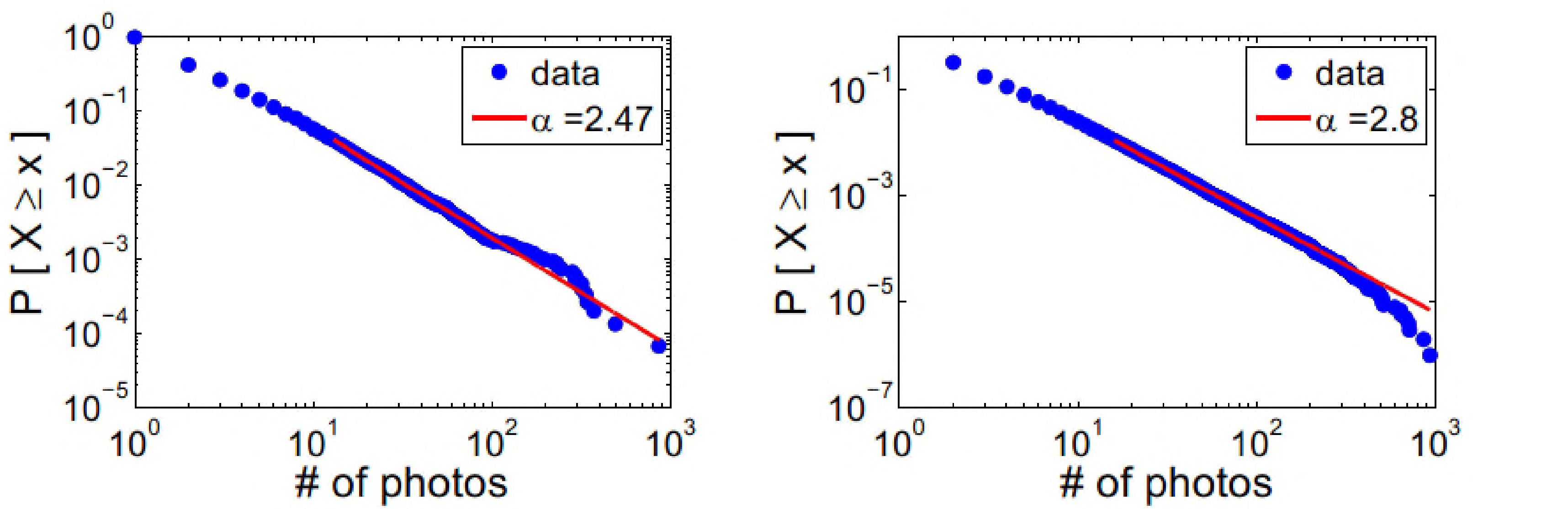
- June/July 2012



(a) New York (b) Rio de Janeiro (e) Paris (f) Sydney
(c) Belo Horizonte (d) Rome (g) Tokyo (h) Cairo

Distribution of the Number of Photos

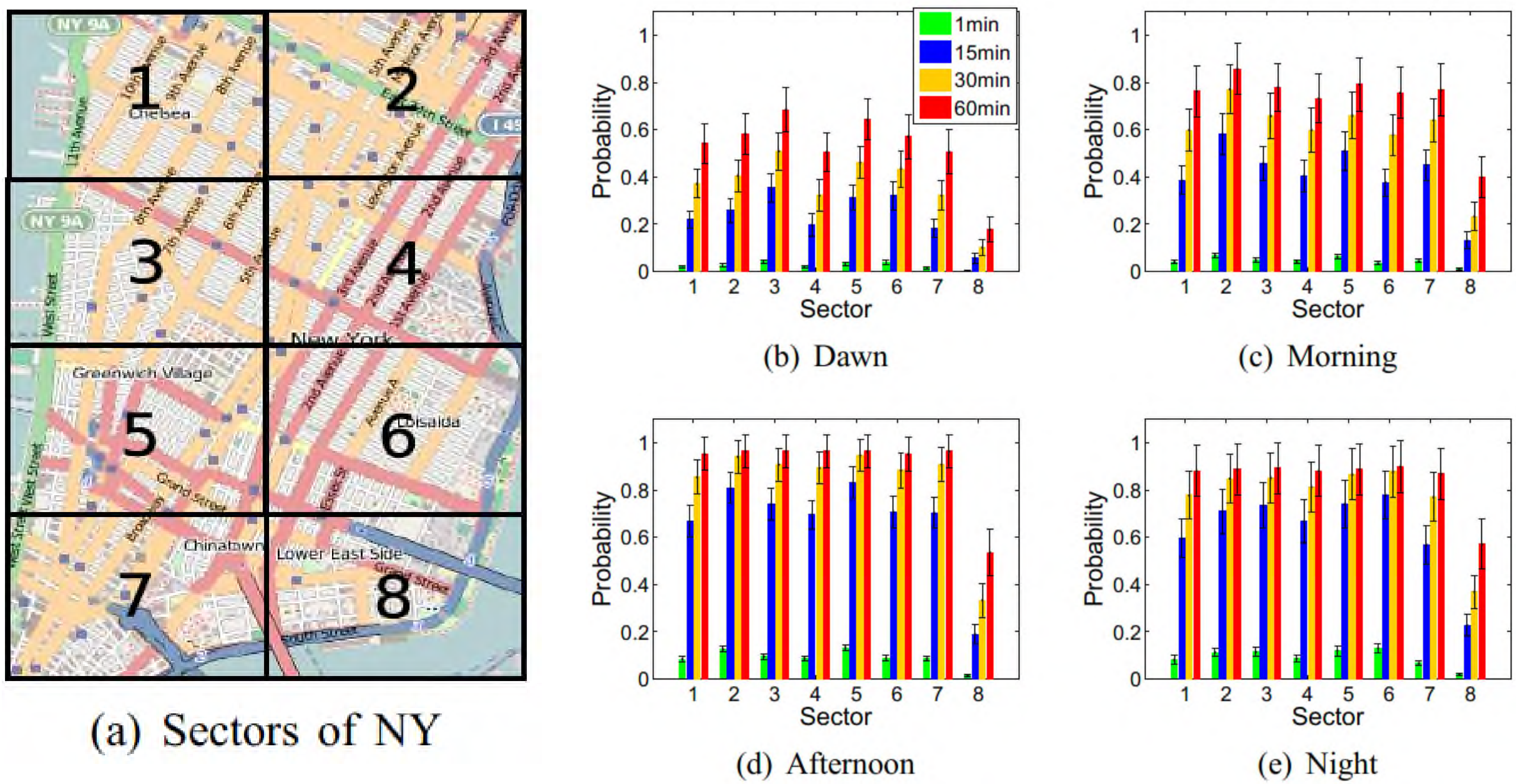
- Most (venue-size quadrants) have a very small photo probability, but some are very popular



(a) NY (b) All locations

Time Between Photos

- Mean probability of another photo within…



(a) Sectors of NY

(b) Dawn

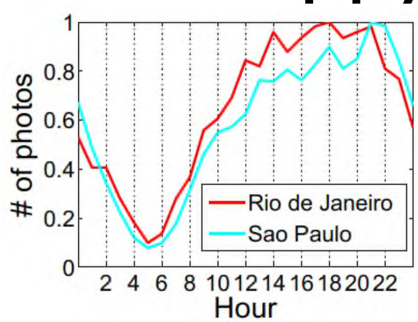
(c) Morning

(d) Afternoon

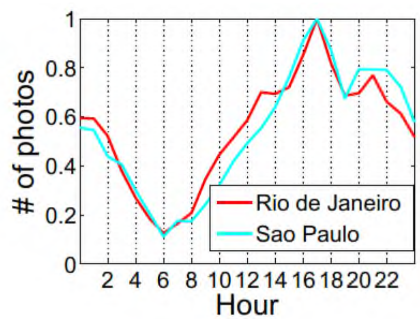
(e) Night

Daily Pattern

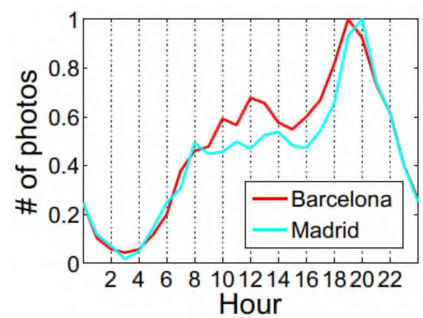
- Picture taking aligned with meal time and “happy hour”



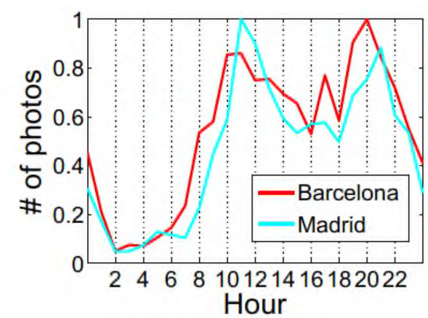
(a) Brazil – Mon to Fri



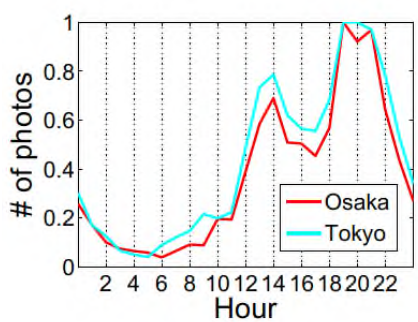
(b) Brazil – Sat to Sun



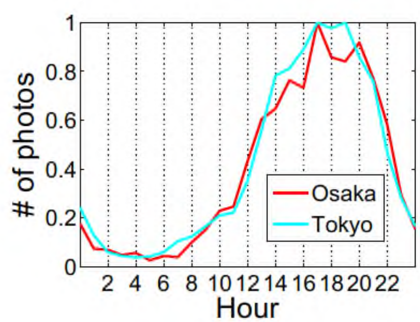
(e) Spain – Mon to Fri



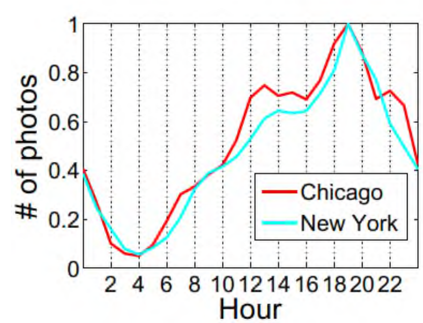
(f) Spain – Sat to Sun



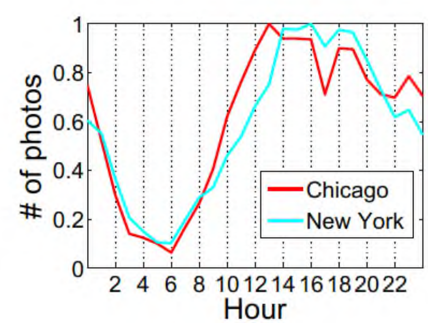
(c) Japan – Mon to Fri



(d) Japan – Sat to Sun



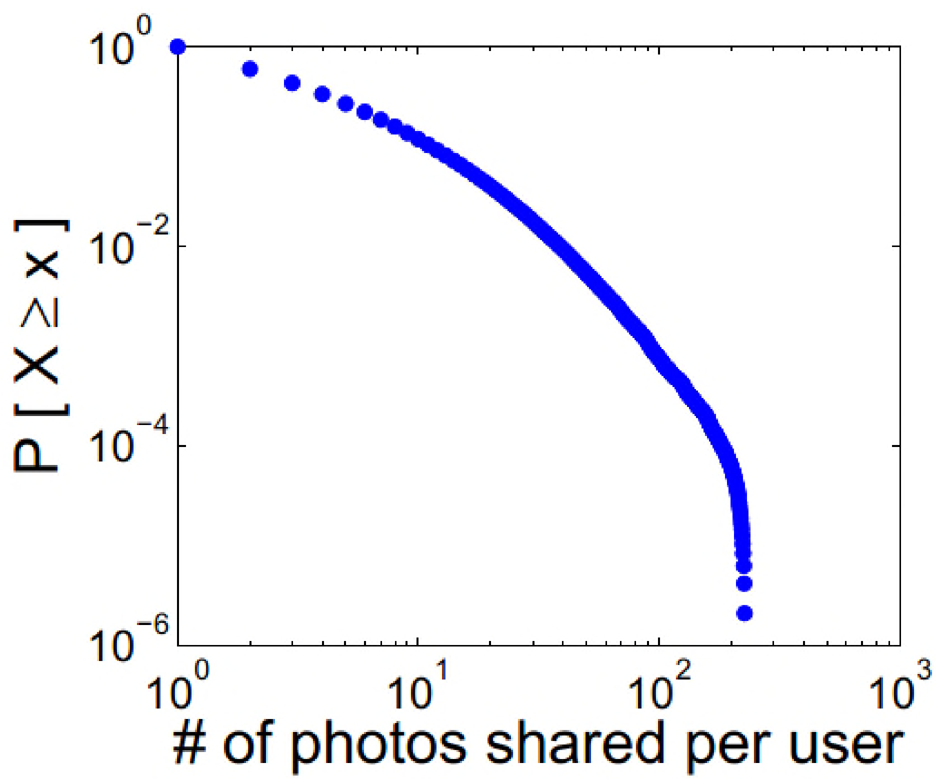
(g) USA - Mon to Fri



(h) USA – Sat to Sun

Distribution of Photos Shared

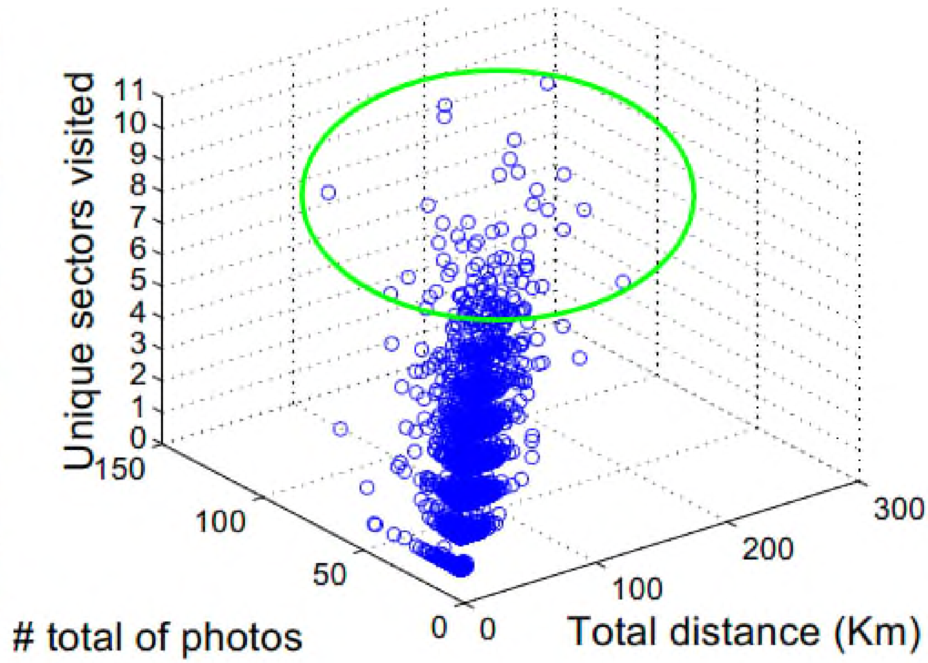
- Some users share hundreds… but most share few



Distribution of Photos Shared

- Some users share hundreds… but most share few

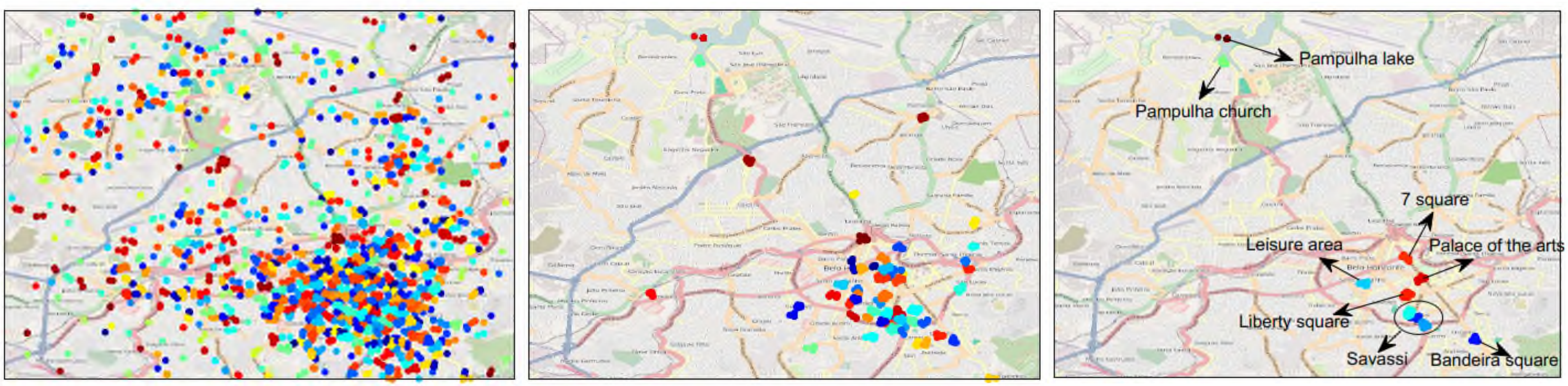
(a) NY in 27 sectors



(b) Node performance

Application: Points of Interest

- Finding clusters related to points of interest
  - How to eliminate places where lots of pictures are taken but only by very few people?
  - How to distinguish tourist attractions?



The Human Sensor Model

- Humans are better at binary observations. For measurements on a scale, use sensors
- Examples of actual Twitter feeds that can be thought of as “binary observations”:
  - “Crash blocking lanes on I-5S @ McBean Pkwy in Santa Clarita”
  - “105E past LakewoodB: traffic stopped to clear tire debris out of lanes”
  - “@BostonGlobe: BREAKING NEWS: Shots fired in Watertown; source says Boston Marathon terror bomb suspect has been pinned down.”
  - “The police chief of Afghanistan's southern Kandahar province has died in a suicide attack on his headquarters.”
  - “Yonkers mayor has lifted his gas rationing order. Fill it up! #SandyABC7”

Dow Jones Hickup

- Dow Jones lost 150 points on a rumor of two explosions in the White House on April 23rd, 2013

BREAKING: Two Explosions in the White House and Barack Obama is Injured

Reconstructing Event Timelines

The Apollo Fact-finder

Boston Bombing

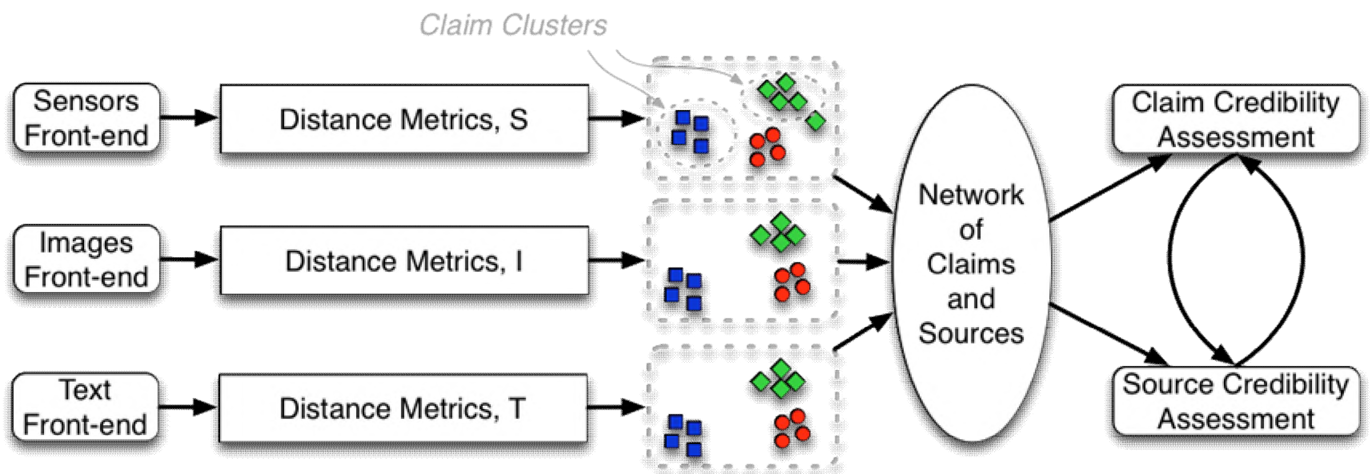
Hurricane Sandy

Egypt unrest

twitter

People

Clean Event Summary?

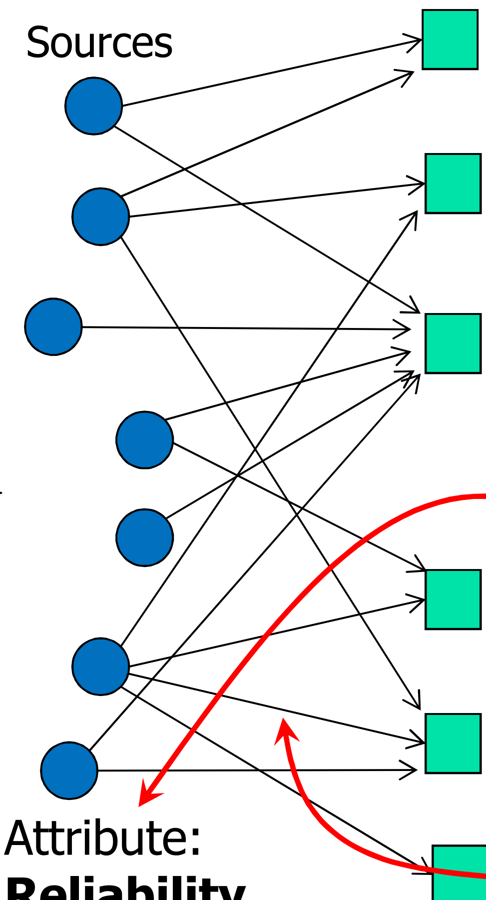


The Apollo Fact-finder

Boston Bombing

Hurricane Sandy

Egypt unrest



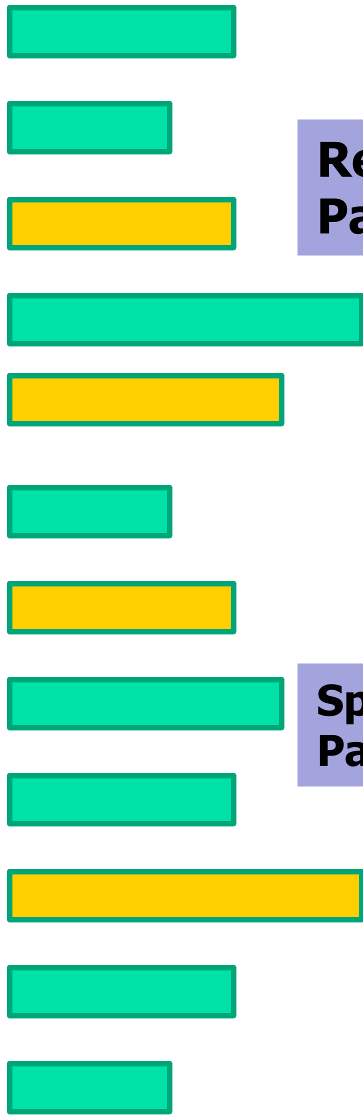
- Define a_i as:
  - P (source$_i$ makes an original observation | it is true)
- Define b_i as:
  - P (source$_i$ makes an original observation | it is false)
- What are the source reliability parameters that maximize the probability of received observations?

$$P(SC|\theta) = \sum_z P(SC, z|\theta)$$

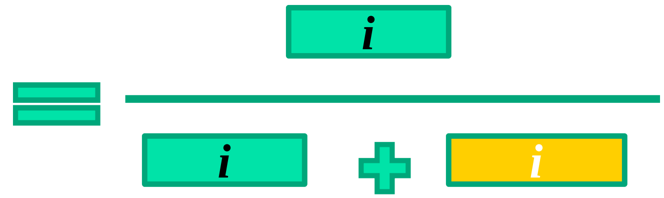
Humans as Sensors

True Assertion

False Assertion



Reliability of Participant *i*

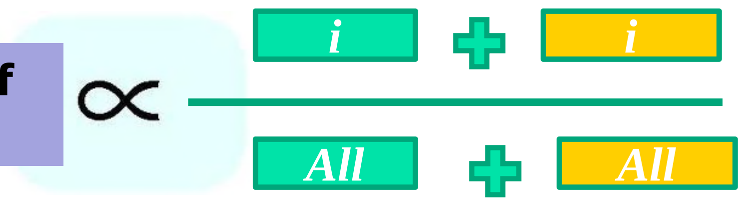


Participant Reliability

$t_i = P(C_j^t \mid S_i C_j)$

$S_i C_j$: participant i claims assertion j

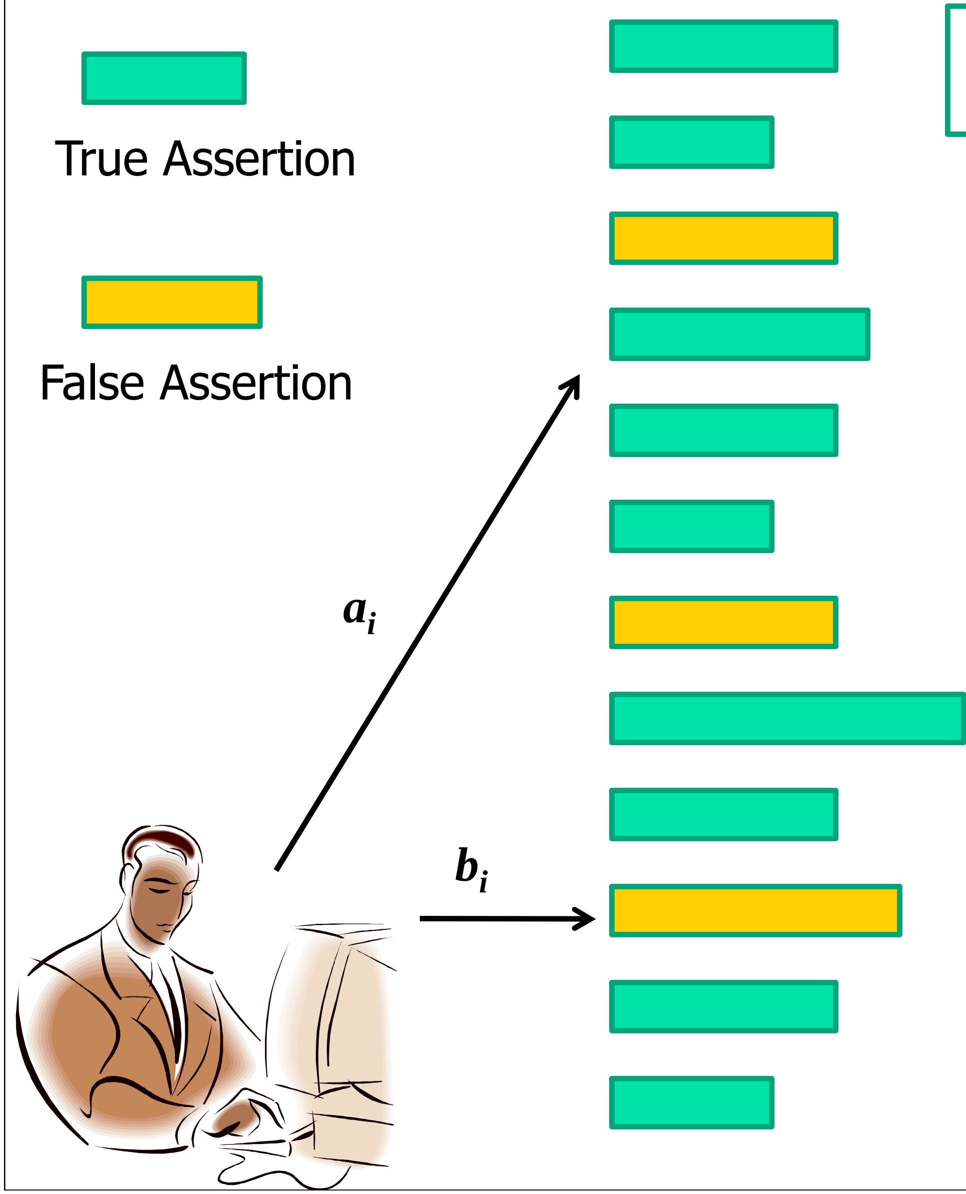
Speak Rate of Participant *i*



Participant i speak with rate s_i

$s_i = P(S_i C_j)$

Expectation Maximization



$$\boxed{a_i = P(S_i C_j \mid C_j^t)}$$

Using Bayesian Theorem: $a_i = \dfrac{t_i \times s_i}{d}$

where d is the overal prior that a randomnly chozen assertion is true

$$\boxed{b_i = P(S_i C_j \mid C_j^f)}$$

Using Bayesian Theorem: $b_i = \dfrac{(1-t_i) \times s_i}{1-d}$

where d is the overal prior that a randomnly chozen assertion is true

Expectation Maximization

Expectation Maximization

$$L(\theta; X) = p(X|\theta) = \sum_Z p(X, Z|\theta)$$

Estimation parameter

Observed data

Hidden Variable

Expectation Step (E-step)

Apply EM

$$Q\left(\theta|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)]$$

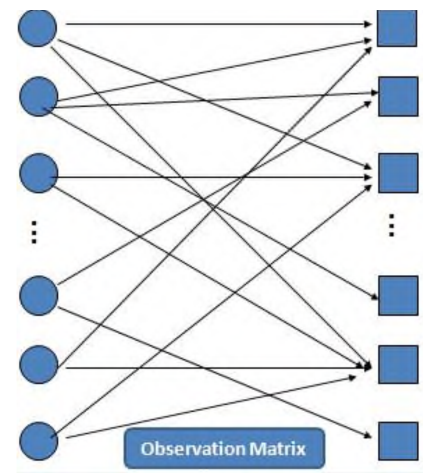
Maximization Step (M-step)

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta|\theta^{(t)}\right)$$

$Z=\{z_1, z_2, \ldots z_N\}$ where $z_j = 1$ when assertion C_j is true and 0 otherwise

Observation Matrix

X



$$\theta = (a_1, a_2, \ldots a_M; b_1, b_2, \ldots b_M; d)$$

Find MLE of estimation parameter and values of hidden variables

Expectation Maximization

Likelihood function of EM

$$L(\theta; X, Z) = p(X, Z|\theta)$$
$$= \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{S_i C_j} (1 - a_i)^{(1 - S_i C_j)} \times d \times z_j + \prod_{i=1}^{M} b_i^{S_i C_j} (1 - b_i)^{(1 - S_i C_j)} \times (1 - d) \times (1 - z_j) \right\}$$

Expectation Step (E-Step)

$$Q\left(\theta|\theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)]$$

$$Z(t, j) = f(a^{(t)}, b^{(t)}, d^{(t)} j)$$

$$= \sum_{j=1}^{N} \left\{ p(z_j = 1|X_j, \theta^{(t)}) \times \left[\sum_{i=1}^{M} (S_i C_j \log a_i + (1 - S_i C_j) \log(1 - a_i) + \log d) \right] \right.$$
$$\left. + p(z_j = 0|X_j, \theta^{(t)}) \times \left[\sum_{i=1}^{M} (S_i C_j \log b_i + (1 - S_i C_j) \log(1 - b_i) + \log(1 - d)) \right] \right\}$$

Maximization Step (M-Step)

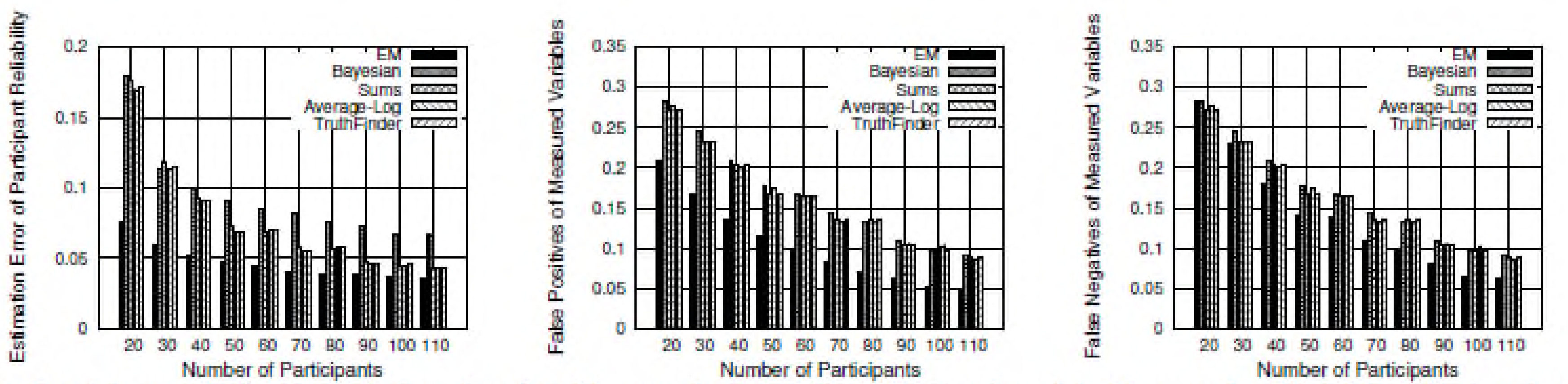
$$a_i^{(t+1)} = a_i^* = \frac{\sum_{j \in SJ_i} Z(t, j)}{\sum_{j=1}^{N} Z(t, j)}$$

$$b_i^{(t+1)} = b_i^* = \frac{K_i - \sum_{j \in SJ_i} Z(t, j)}{N - \sum_{j=1}^{N} Z(t, j)}$$

$$d_i^{(t+1)} = d_i^* = \frac{\sum_{j=1}^{N} Z(t, j)}{N}$$

Iterate

Simulation



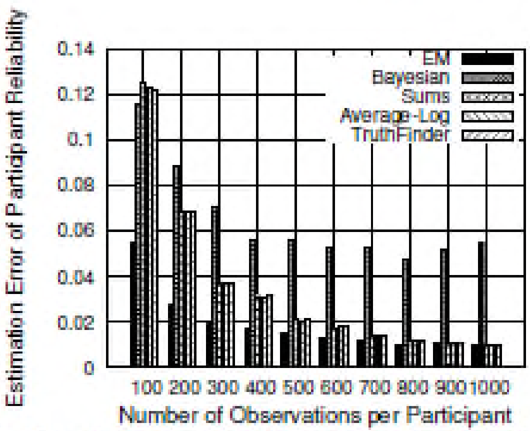
(a) Participant Reliability Estimation Accuracy (b) Measured Variable Estimation: False Positives (c) Measured Variable Estimation: False Negatives

EM outperforms state-of-art heuristics

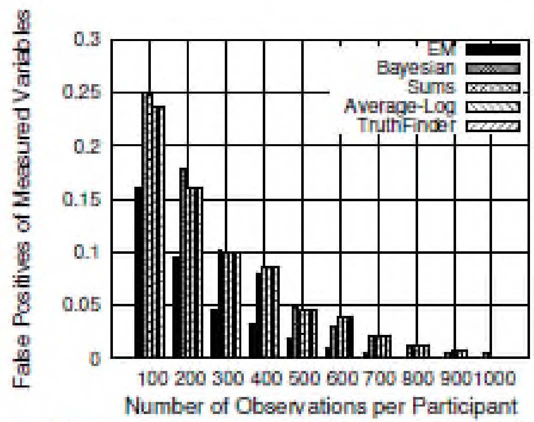
Parameters:

Number of Participants: 20-110, Number of True Assertions: 1000, Number of False Assertions: 1000, Average Number of Claims per Participant: 100

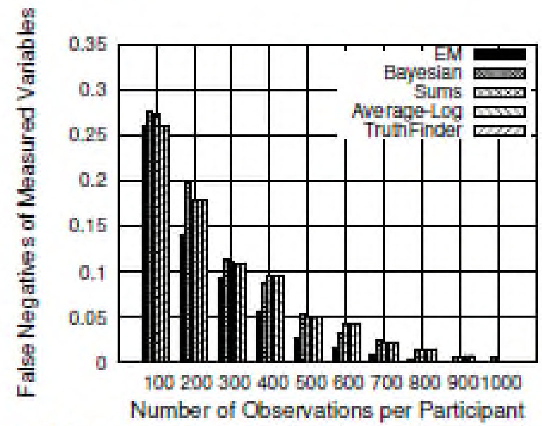
Simulation



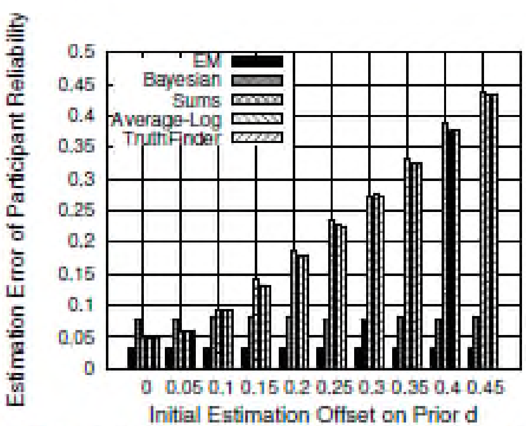
(a) Participant Reliability Estimation Accuracy



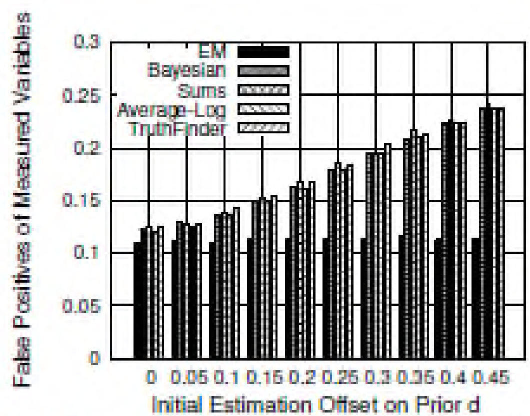
(b) Measured Variable Estimation: False Positives



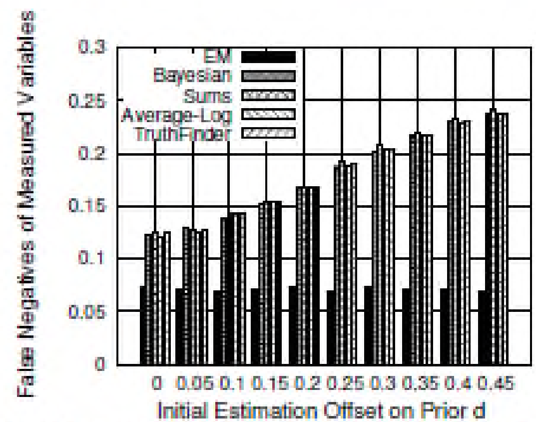
(c) Measured Variable Estimation: False Negatives



(a) Participant Reliability Estimation Accuracy

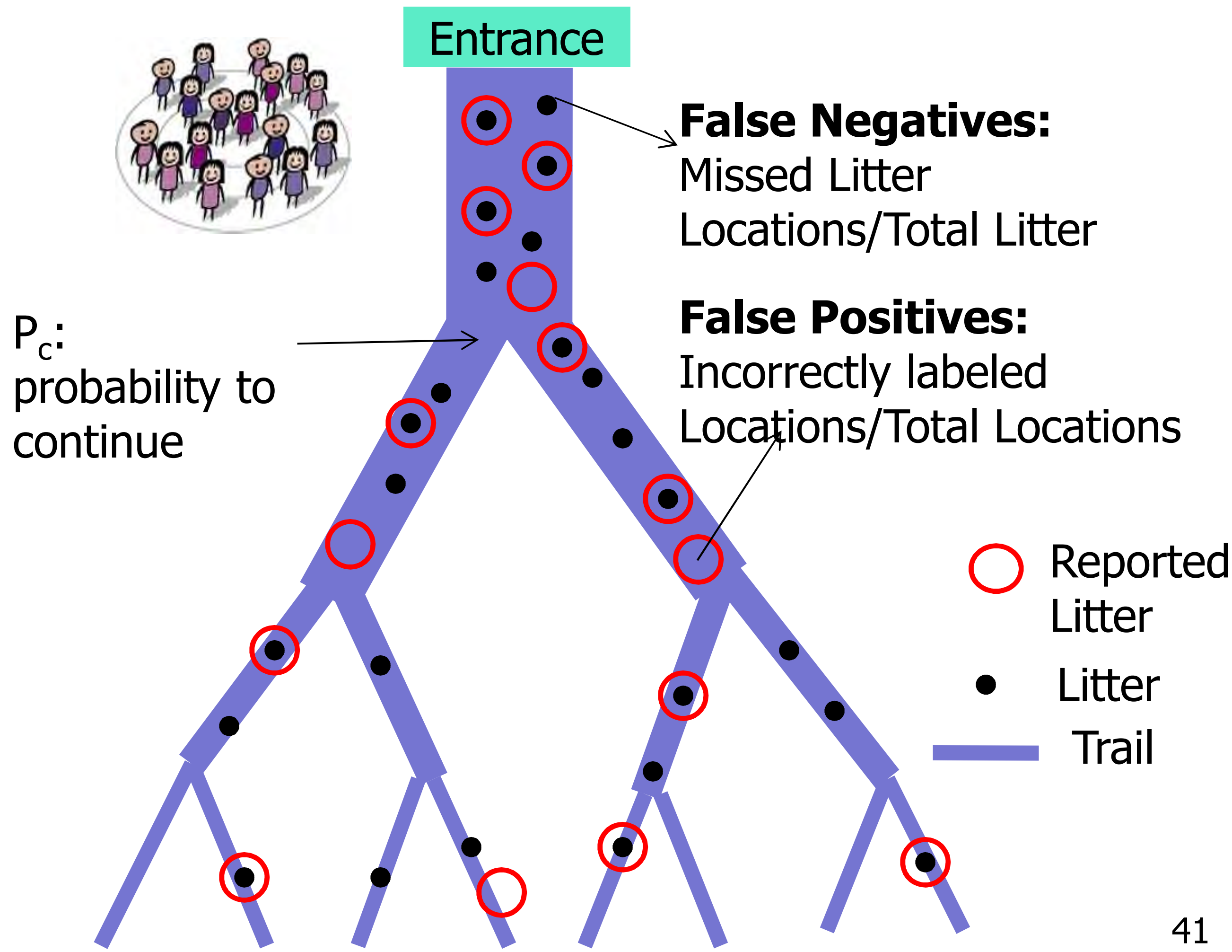


(b) Measured Variable Estimation: False Positives

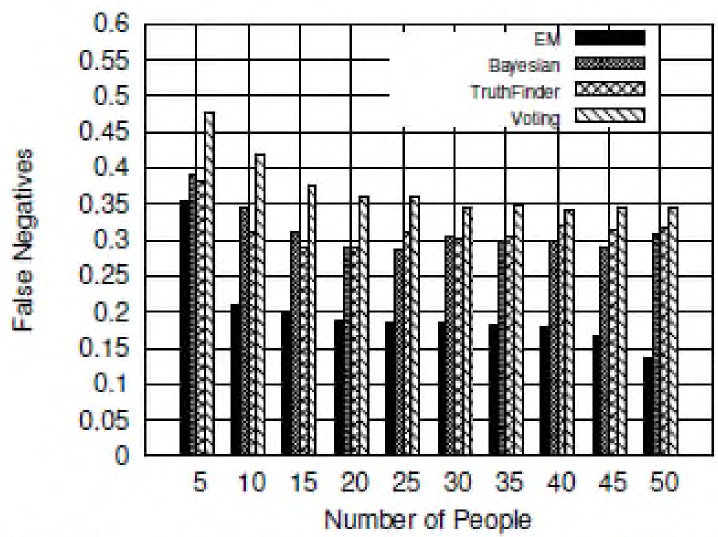


(c) Measured Variable Estimation: False Negatives

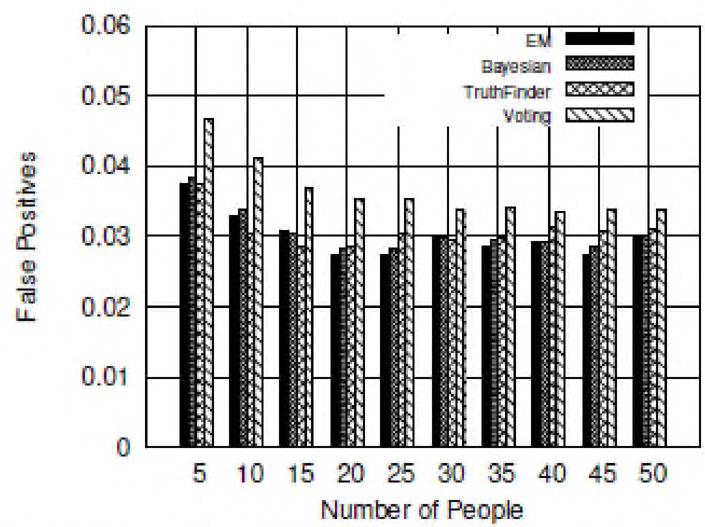
Simulated Geotagging



Simulated Geotagging



(a) False Negatives (missed/total litter)



(b) False Positives (false/total locations)

Litter Geotagging Accuracy <u>versus Number of People</u>

Twitter-based Evaluation (Hurricane Irene)

#	Media	Tweet found by EM
1	East Coast Braces For Hurricane Irene; Hurricane Irene is expected to follow a path up the East Coast	@JoshOchs A #hurricane here on the east coast
2	Hurricane Irene's effects begin being felt in NC, The storm, now a Category 2, still has the East Coast on edge.	Winds, rain pound North Carolina as Hurricane Irene closes in http://t.co/0gVOSZk
3	Hurricane Irene charged up the U.S. East Coast on Saturday toward New York, shutting down the city, and millions of Americans sought shelter from the huge storm.	Hurricane Irene rages up U.S. east coast http://t.co/u0XiXow
4	The Wall Street Journal has created a way for New Yorkers to interact with the location-based social media app Foursquare to find the nearest NYC hurricane evacuation center.	Mashable - Hurricane Irene: Find an NYC Evacuation Center on Foursquare ... http://t.co/XMtpH99
5	Following slamming into the East Coast and knocking out electricity to more than a million people, Hurricane Irene is now taking purpose on largest metropolitan areas in the Northeast.	2M lose power as Hurricane Irene moves north - Two million homes and businesses were without power ... http://t.co/fZWkEU3
6	Irene remains a Category 1, the lowest level of hurricane classification, as it churns toward New York over the next several hours, the U.S. National Hurricane Center said on Sunday.	Now its a level 1 hurricane. Let's hope it hits NY at Level 1
7	Blackouts reported, storm warnings issued as Irene nears Quebec, Atlantic Canada.	DTN Canada: Irene forecast to hit Atlantic Canada http://t.co/MjhmeJn
8	President Barack Obama declared New York a disaster area Wednesday, The New York Times reports, allowing the release of federal aid to the state's government and individuals.	Hurricane Irene: New York State Declared A Disaster Area By President Obama
9	Hurricane Irene's rampage up the East Coast has become the tenth billion-dollar weather event this year, breaking a record stretching back to 1980, climate experts said Wednesday.	Irene is 10th billion-dollar weather event of 2011.
10	WASHINGTON- On Sunday, September 4, the President will travel to Paterson, New Jersey, to view damage from Hurricane Irene.	White House: Obama to visit Paterson, NJ Sunday to view damage from Hurricane Irene

Top correct tweets found by EM matches well with Media Reports