Data Reliability II:
A Fundamental Challenge in Social Sensing

With Humans as Sensors

CSE 40437/60437-Spring 2015

Prof. Dong Wang

Last Lecture: Humans as Sensors

Expectation Maximization

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

Estimation parameter | Observed data | Hidden Variable

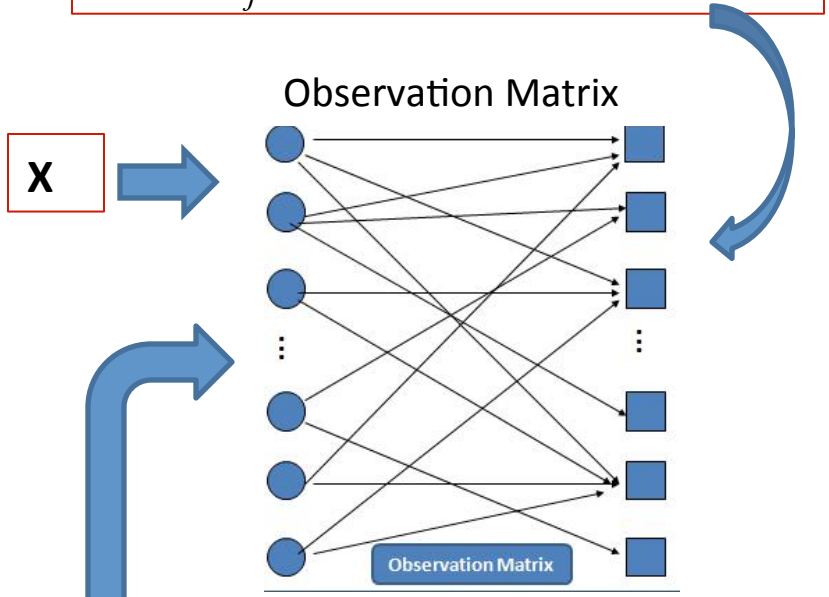
- Expectation Step (E-step)

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

- Maximization Step (M-step)

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

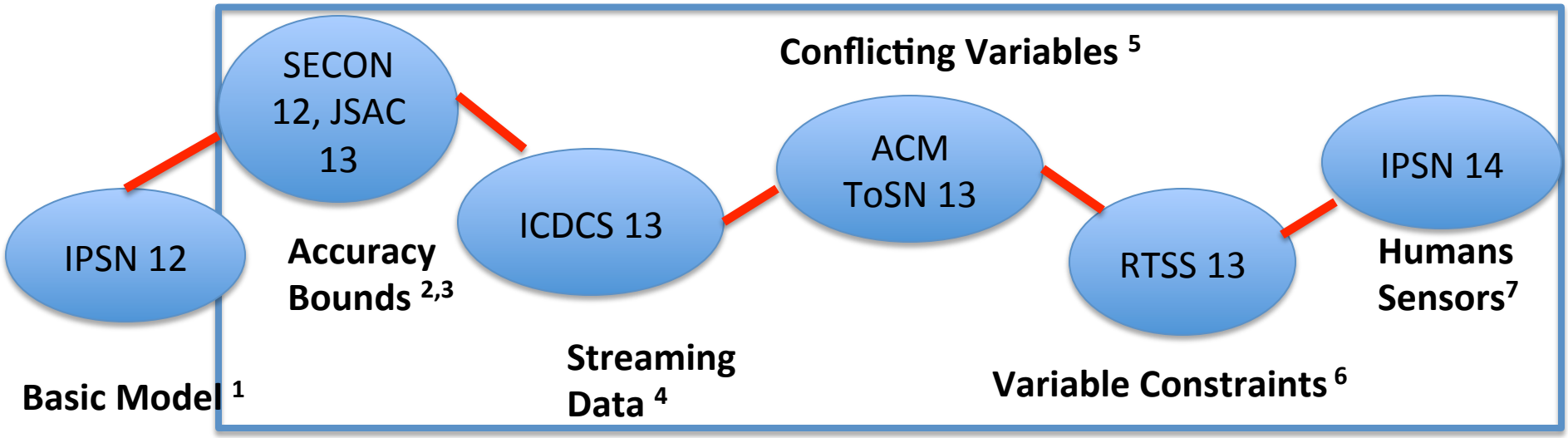
$Z=\{z_1, z_2, \ldots z_N\}$ where z_j =1 when claim C_j is true and 0 otherwise



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

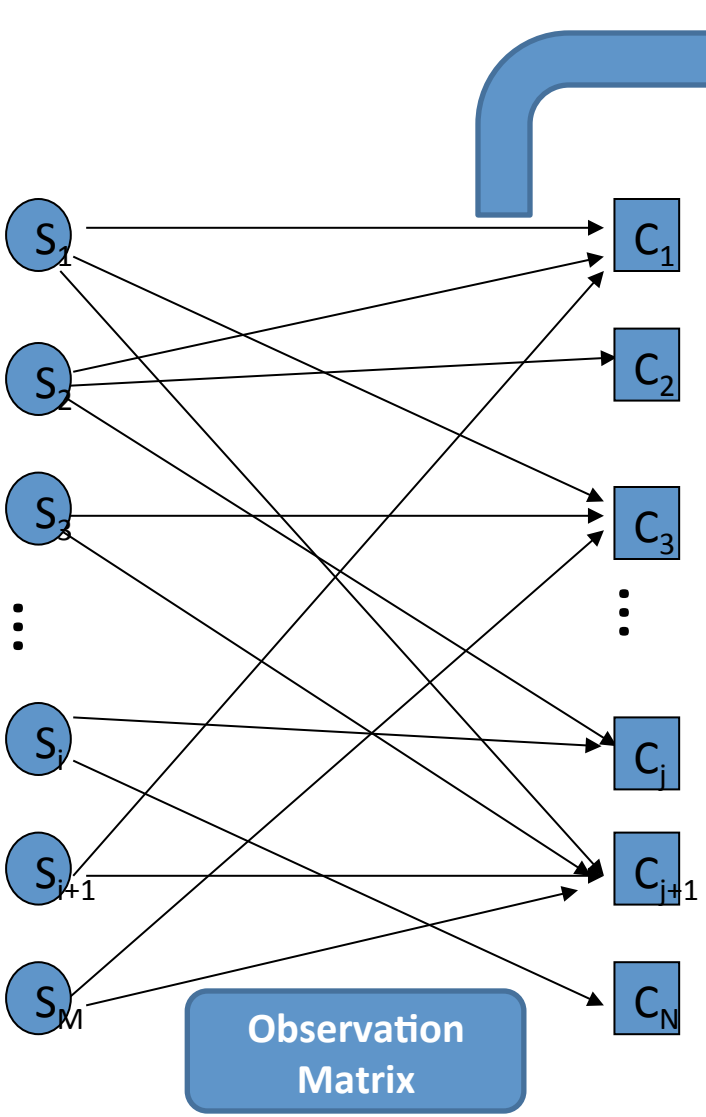
Find MLE of estimation parameter and values of hidden variables

Related Work
on Data Reliability in Social Sensing



1. Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. "On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach." The 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN 12). Beijing, China April 2012.
2. Dong Wang , Lance Kaplan, Tarek Abdelzaher and Charu C. Aggarwal. "On Scalability and Robustness Limitations of Real and Asymptotic Confidence Bounds in Social Sensing." The 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 12), Seoul, Korea, June, 2012.
3. Dong Wang, Lance Kaplan, Tarek Abdelzaher and Charu C. Aggarwal. "On Credibility Tradeoffs in Assured Social Sensing." IEEE JSAC special issue on Network Science, June, Vol. 31, No. 6, 2013.
4. Dong Wang, Tarek Abdelzaher, Lance Kaplan and Charu C. Aggarwal. "Recursive Fact-finding: A Streaming Approach to Truth Estimation in Crowdsourcing Applications.", 33rd International Conference on Distributed Computing Systems (ICDCS 13) Philadelphia, PA, July 2013.
5. Dong Wang, Lance Kaplan and Tarek Abdelzaher. "On Truth Discovery in Social Sensing with Conflicting Observations: A Maximum Likelihood Estimation Approach." ACM Transaction on Sensor Network s (TOSN), in press, 2013
6. Dong Wang, Tarek Abdelzaher, Lance Kaplan and Raghu Ganti. "Exploitation of Physical Constraints for Reliable Social Sensing," IEEE 34th Real-Time Systems Symposium (RTSS'13) Vancouver, Canada, December, 2013.
7. Dong Wang , Tanvir Amin, Shen Li, Tarek Abdelzaher, Lance Kaplan, Siyu Gu, Chenji Pan, Hengchang Liu, Charu Aggrawal, Raghu Ganti, XinLei Wang, Prasant Mohapatra, Boleslaw Szymanski, Hieu Le, "Humans as Sensors: An Estimation Theoretic Perspective," The 13th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN 14), Berlin, Germany, April, 2014.

Estimation Accuracy of MLE from EM?

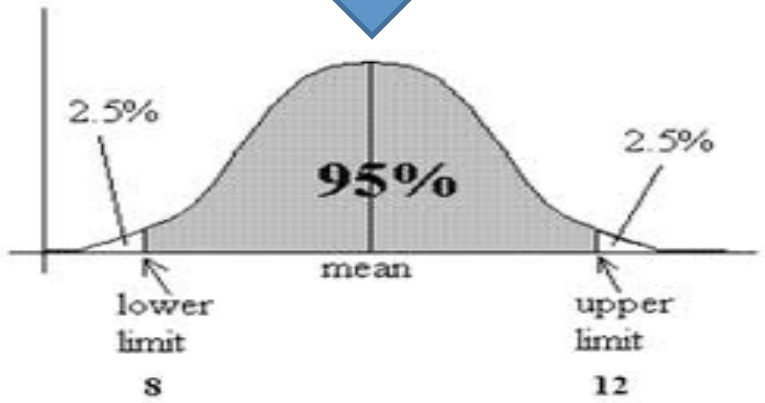


Expectation Maximization

Maximum Likelihood Estimation

Participant Reliability → **Claim Correctness**

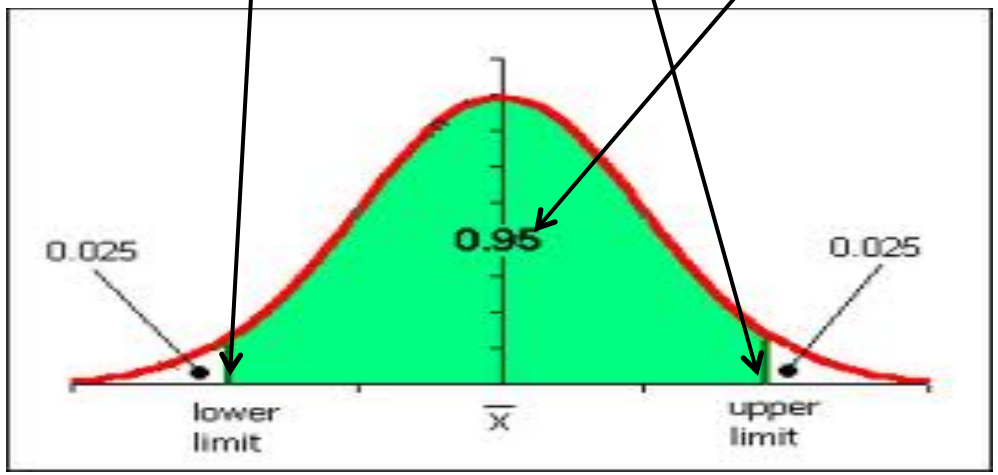
Error Bound?



Establishing Estimation Confidence

Goal: Identify **Confidence Interval** of Source Reliability Estimation from MLE

$$\left(t_i^{MLE} - c_p^{lower}, t_i^{MLE} + c_p^{upper}\right) \text{ with } c\%$$



Establishing Estimation Confidence

Estimation and Statistic Background

Fisher information is defined as

$$I(\theta) = E_X\left[\varphi(x;\theta)\ \varphi(x;\theta)^T\right]$$

→ Fisher Information

Score vector $\varphi(x;\theta)$ for a $k \times 1$ estimation vector $\theta = [\theta_1, \theta_2, ..., \theta_k]^T$

$$\varphi(x;\theta) = \left[\frac{\partial l(x;\theta)}{\partial \theta_1}, \frac{\partial l(x;\theta)}{\partial \theta_2}, ..., \frac{\partial l(x;\theta)}{\partial \theta_k}\right]^T$$

Fisher Information Matrix can rewritten as (under regularity condition of EM) :

$$(I(\theta))_{i,j} = -E_X\left[\frac{\partial^2 l(x;\theta)}{\partial \theta_i \partial \theta_j}\right]$$

Cramer-Rao Lower Bound (CRLB) is defined as the inverse of Fisher information

$$CRLB = I^{-1}(\theta)$$

→ Cramer-Rao Lower Bound

Establishing Estimation Confidence

Deriving the Cramer-Rao Lower Bound

$$CRLB = J^{-1}$$

where

$$J = E[\nabla_\theta \ln p(X|\theta)\ \nabla_\theta^H \ln p(X|\theta)]$$

$$p(X|\theta) = \prod_{j=1}^{N} \left\{ \prod_{i=1}^{M} a_i^{X_{ij}} (1-a_i)^{(1-X_{ij})} \times d + \prod_{i=1}^{M} b_i^{X_{ij}} (1-b_i)^{(1-X_{ij})} \times (1-d) \right\}$$

$$J^{-1} = \frac{1}{N} \begin{bmatrix} \bar{A} & \bar{C} \\ \bar{C}^T & \bar{B} \end{bmatrix}^{-1}$$

Cramer-Rao Lower Bound (CRLB)

Confidence Interval of Participant Reliability

Real Cramer-Rao Lower Bound Derivation

$$J^{-1} = \frac{1}{N}\begin{bmatrix} \bar{A} & \bar{C} \\ \bar{C}^T & \bar{B} \end{bmatrix}^{-1}$$

Can be computed numerically and computation is **exponential** w.r.t to number of participants

$$\bar{A}_{kl} = \sum_{x \in \mathcal{X}_j} \frac{(2X_{kj}-1)(2X_{lq}-1)\prod_{\substack{i=1 \\ i\neq k}}^{M} A_{ij} \prod_{\substack{i=1 \\ i\neq l}}^{M} A_{ij} d^2}{\prod_{i=1}^{M} A_{ij} d + \prod_{i=1}^{M} B_{ij}(1-d)}$$

$$B_{kl} = \sum_{x \in \mathcal{X}_j} \frac{(2X_{kj}-1)(2X_{lq}-1)\prod_{\substack{i=1 \\ i\neq k}}^{M} B_{ij} \prod_{\substack{i=1 \\ i\neq l}}^{M} B_{ij} (1-d)^2}{\prod_{i=1}^{M} A_{ij} d + \prod_{i=1}^{M} B_{ij}(1-d)}$$

$$|\mathcal{X}_j| = 2^M$$

$$\bar{C}_{kl} = \sum_{x \in \mathcal{X}_j} \frac{(2X_{kj}-1)(2X_{lq}-1)\prod_{\substack{i=1 \\ i\neq k}}^{M} A_{ij} \prod_{\substack{i=1 \\ i\neq l}}^{M} B_{ij} d(1-d)}{\prod_{i=1}^{M} A_{ij} d + \prod_{i=1}^{M} B_{ij}(1-d)}$$

Confidence Interval of Participant Reliability

Asymptotic Cramer-Rao Lower Bound Derivation

Log-likelihood Function of EM Scheme:

$$l_{em}(x;\theta) = \sum_{j=1}^{N} \left\{ \begin{array}{l} z_j \times \left[\sum_{i=1}^{M} \left(S_i C_j \log a_i + (1 - S_i C_j) \log(1 - a_i) + \log d \right) \right] \\ +(1 - z_j) \times \left[\sum_{i=1}^{M} \left(S_i C_j \log b_i + (1 - S_i C_j) \log(1 - b_i) + \log(1 - d) \right) \right] \end{array} \right\}$$

where $z_j = 1$ when assertion j is true and 0 otherwise

Assume truthfulness of each variable is estimated correctly from EM.

Confidence bound remains to be ***asymptotic***

Confidence Interval of Participant Reliability

Asymptotic Cramer-Rao Lower Bound Derivation

Plugging $l_{em}(x;\theta)$ into the Fisher Information Matrix:

$$(J(\hat{\theta}_{MLE}))_{i,j} = \begin{cases} 0 & i \neq j \\ -E_X\left[\frac{\partial^2 l_{em}(x;a_i)}{\partial a_i^2}|_{a_i=\hat{a}_i^{MLE}}\right] & i = j \in [1, M] \\ -E_X\left[\frac{\partial^2 l_{em}(x;b_i)}{\partial b_i^2}|_{b_i=\hat{b}_i^{MLE}}\right] & i = j \in (M, 2M] \end{cases}$$

The inverse of above matrix is:

$$\boxed{(J^{-1}(\hat{\theta}_{MLE}))_{i,j}} = \begin{cases} 0 & i \neq j \\ \frac{\hat{a}_i^{MLE} \times (1-\hat{a}_i^{MLE})}{N \times d} & i = j \in [1, M] \\ \frac{\hat{b}_i^{MLE} \times (1-\hat{b}_i^{MLE})}{N \times (1-d)} & i = j \in (M, 2M] \end{cases}$$

Establishing Estimation Confidence

Maximum Likelihood Property

Asymptotic Normality

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{D} N\left(0, J^{-1}(\hat{\theta}_{MLE})\right)$$

CRLB

Confidence Interval of Source S_i

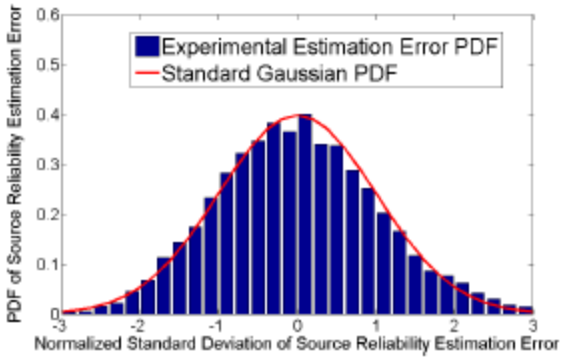
$$\left(\hat{t}_i^{MLE} - c_p\sqrt{Var\left(\hat{t}_i^{MLE}\right)}, \hat{t}_i^{MLE} + c_p\sqrt{Var\left(\hat{t}_i^{MLE}\right)}\right)$$

c_p is the standard score of confidence level p

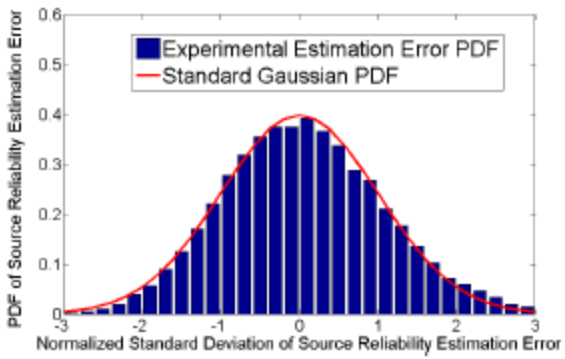
Confidence Interval of Participant Reliability

Asymptotic Normality Evaluation

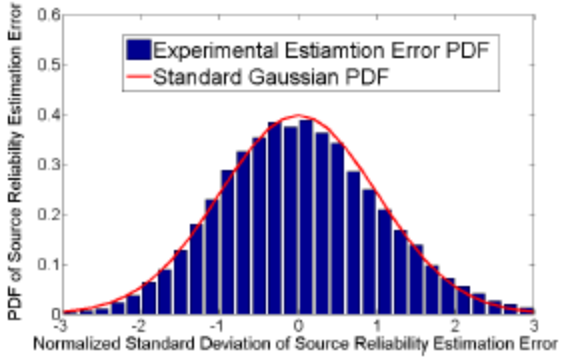
PDF of Participant Reliability Estimation Error



(a) Small Observation Matrix



(b) Medium Observation Matrix



(c) Large Observation Matrix

Observation Matrix Scale	Number of Sources	Number of True Measured Variables	Number of False Measured Variables
Small	100	500	500
Medium	200	1000	1000
Large	300	2000	2000

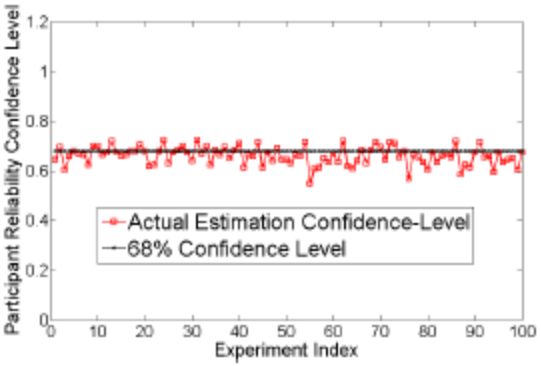
TABLE I
PARAMETERS OF THREE TYPICAL OBSERVATION MATRIX SCALE

PDF of experiment matched the Standard Gaussian Distribution well

Confidence Interval of Participant Reliability

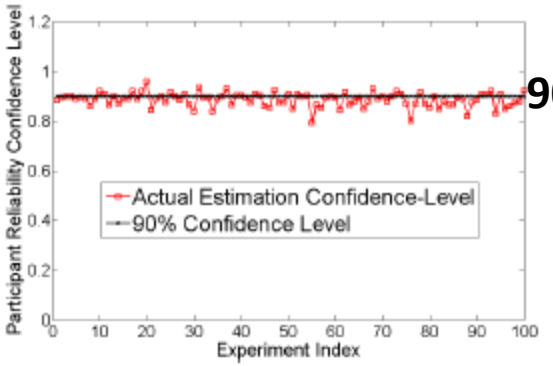
Estimation Confidence Evaluation

Estimation Confidence on Participant Reliability on small observation matrix scale



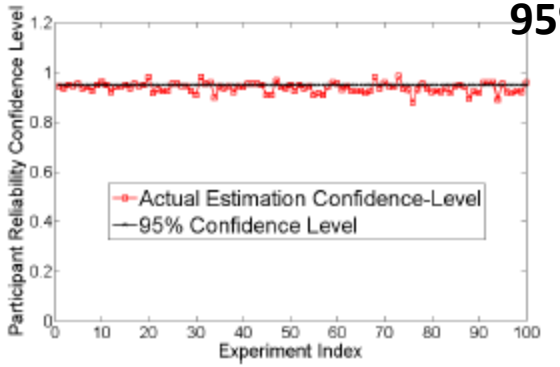
68%

(a) 68% Confidence Level



90%

(b) 90% Confidence Level



95%

(c) 95% Confidence Level

Parameters:

Number of Participants: 100

Number of True Assertions: 500

Number of False Assertions: 500

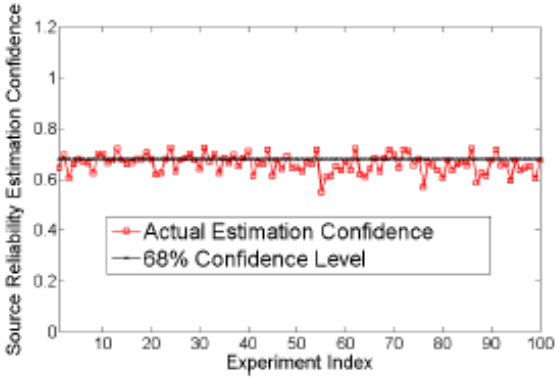
Average Number of Claims per Participant: 100

Estimation confidence stays around the correct confidence level-small scale

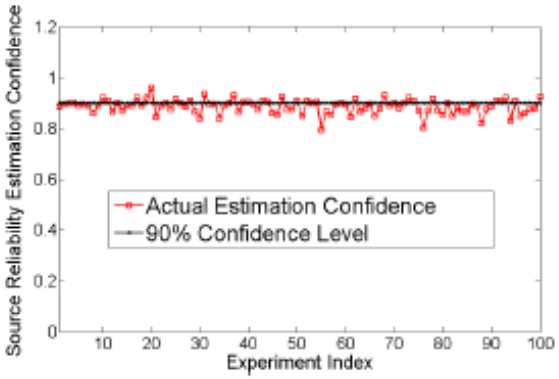
Confidence Interval of Participant Reliability

Estimation Confidence Evaluation

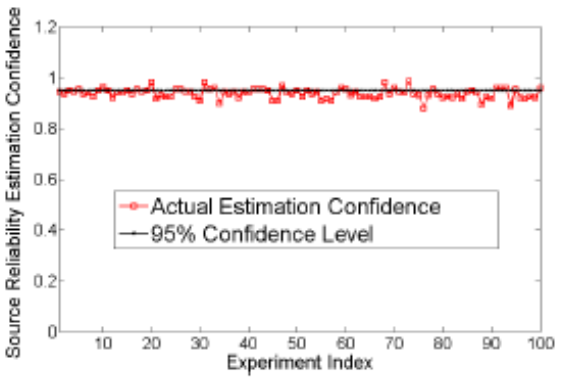
Estimation Confidence on Participant Reliability on medium observation matrix scale



(a) 68% Confidence Level



(b) 90% Confidence Level



(c) 95% Confidence Level

Parameters:

Number of Participants: 200

Number of True Assertions: 1000

Number of False Assertions: 1000

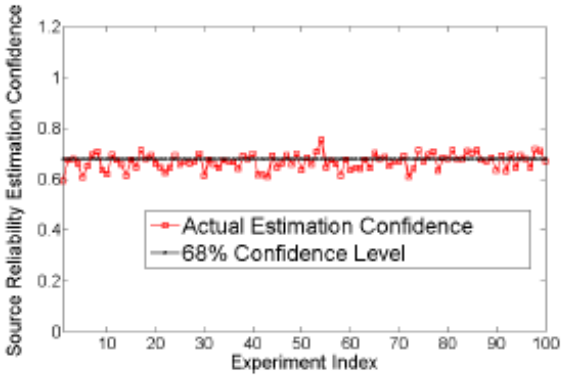
Average Number of Claims per Participant: 100

Estimation confidence stays around the correct confidence level-medium scale

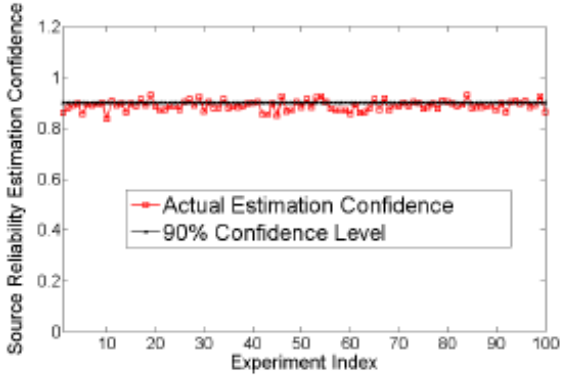
Confidence Interval of Participant Reliability

Estimation Confidence Evaluation

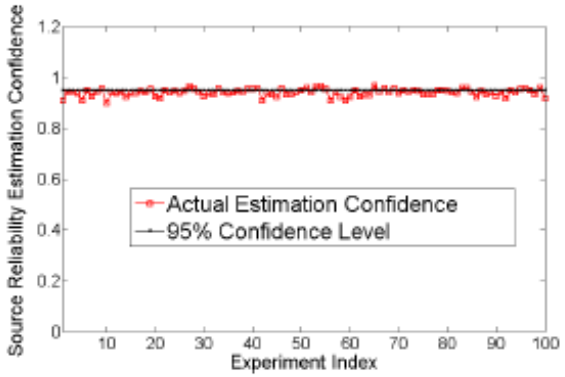
Estimation Confidence on Participant Reliability on large observation matrix scale



(a) 68% Confidence Level



(b) 90% Confidence Level



(c) 95% Confidence Level

Parameters:

Number of Participants: 300

Number of True Assertions: 2000

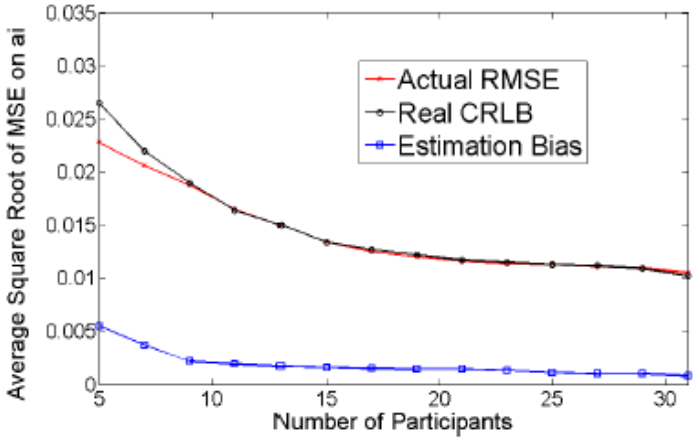
Number of False Assertions: 2000

Average Number of Claims per Participant: 100

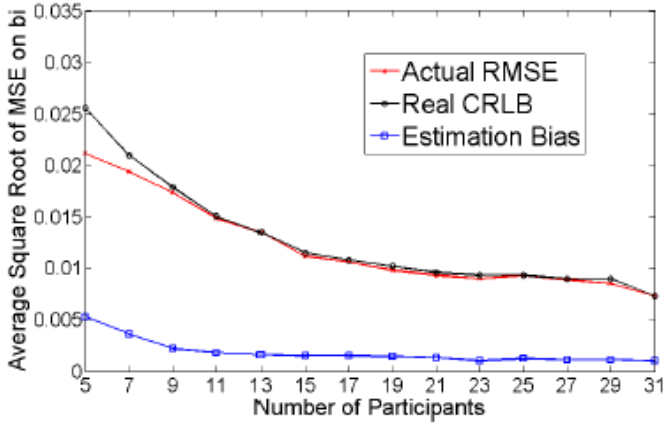
Estimation confidence stays around the correct confidence level-large scale

Scalability of CRLB

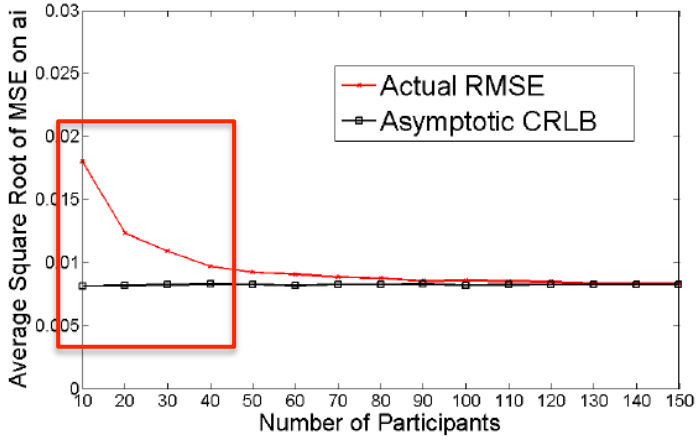
CRLB on Estimation Parameters versus Varying Number of Participants



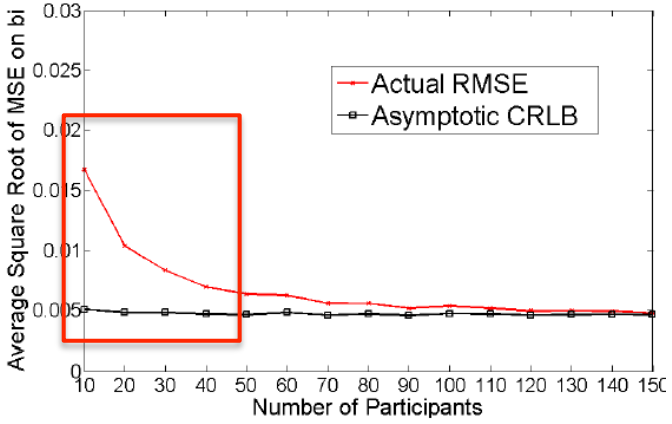
(a) Real CRLB of a_i



(b) Real CRLB of b_i



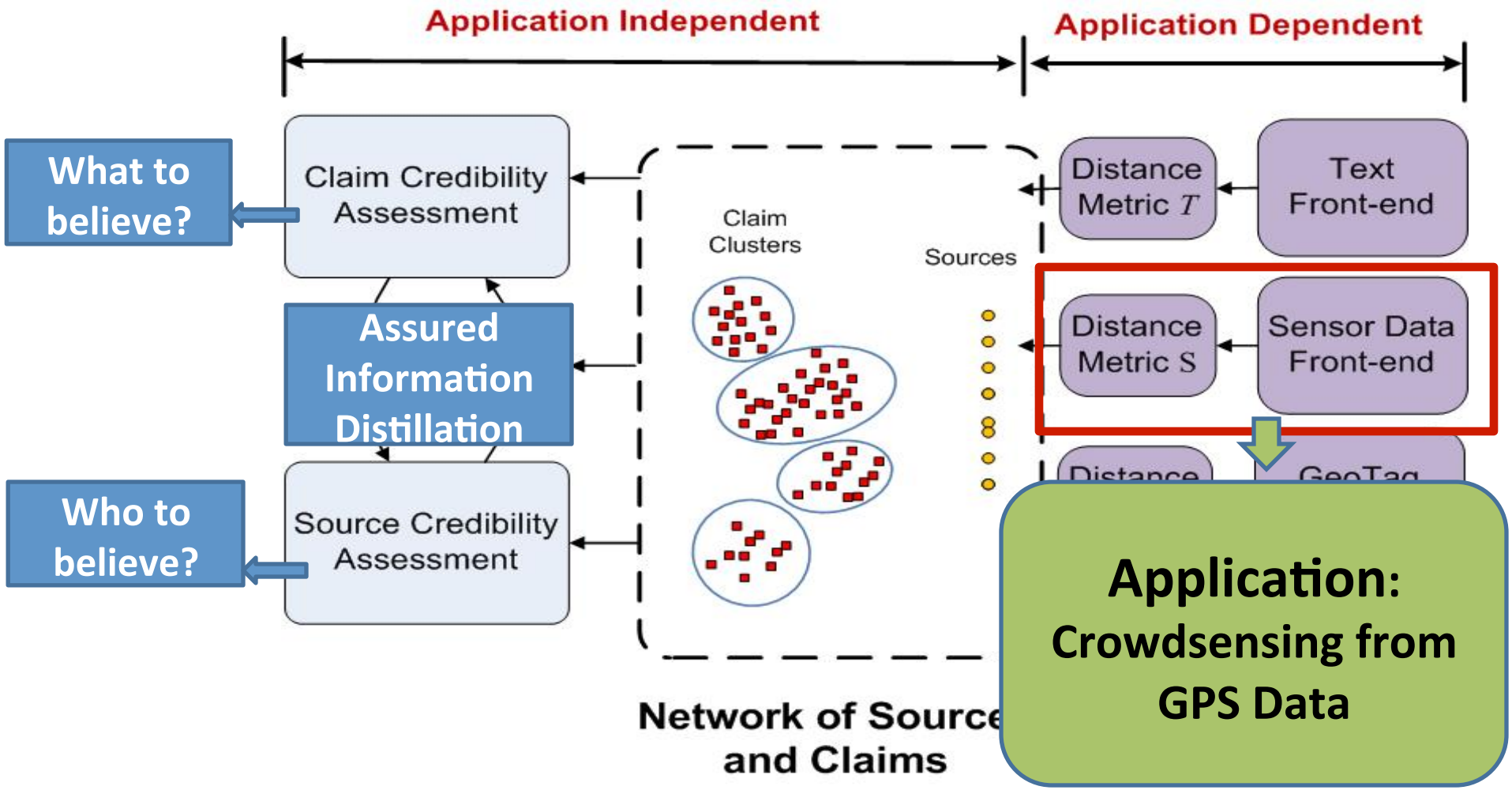
(a) Asymptotic CRLB of a_i



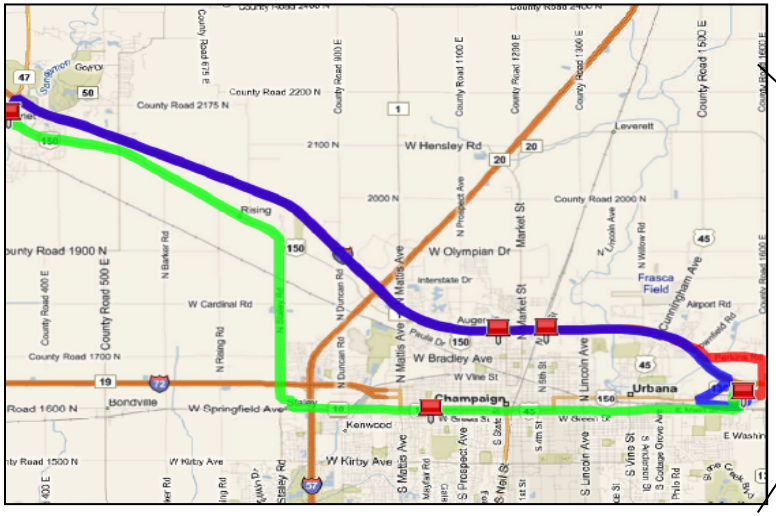
(b) Asymptotic CRLB of b_i

Evaluation using Crowdsensing

The Apollo Fact-finder



Crowdsensing Application Case Study:
Traffic Regulator Mapping from GPS Data



Method: Intentionally **Simple** "Sensors"

15-90 s Stop

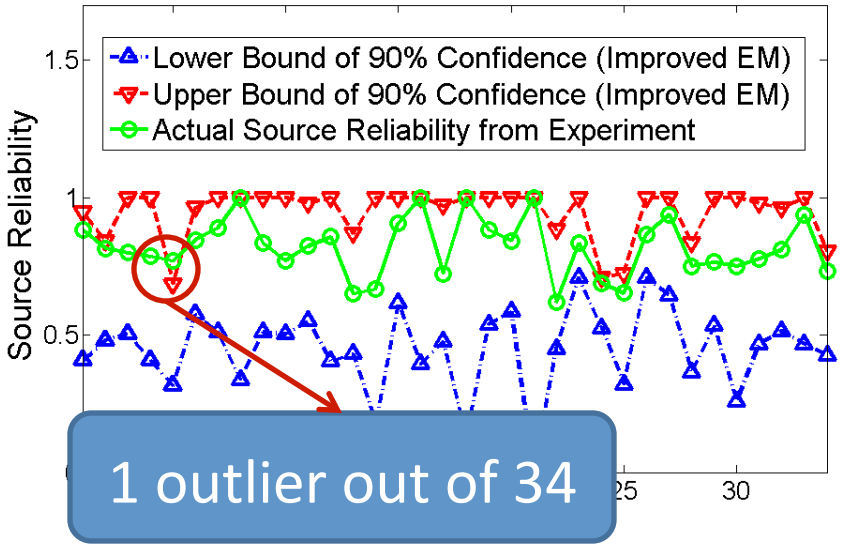
Claim Traffic Light

2-10 s Stop

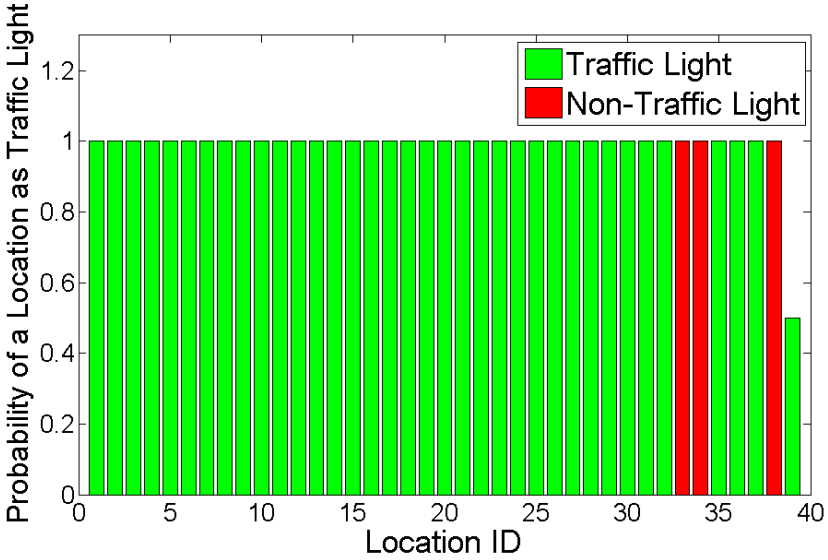
Claim Stop Sign

Unreliable "Sensors"!

Traffic Regulator Mapping From GPS Data



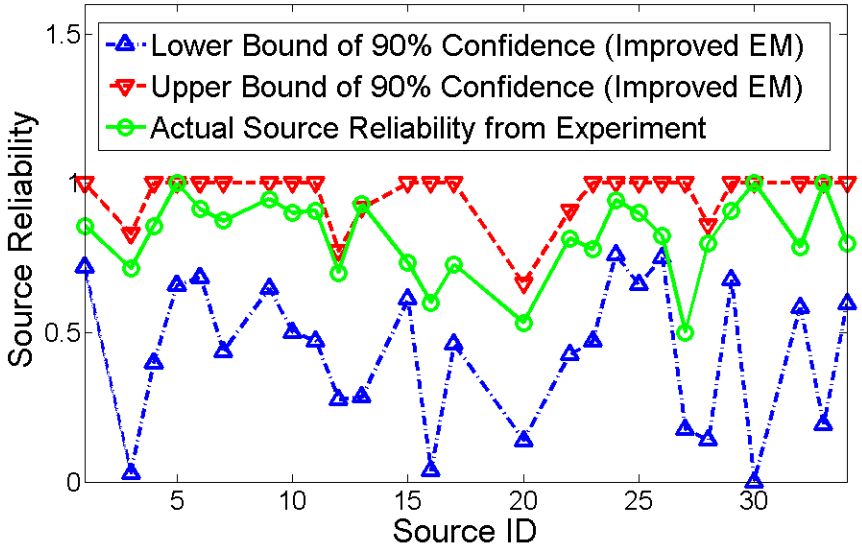
Source Reliability Estimation

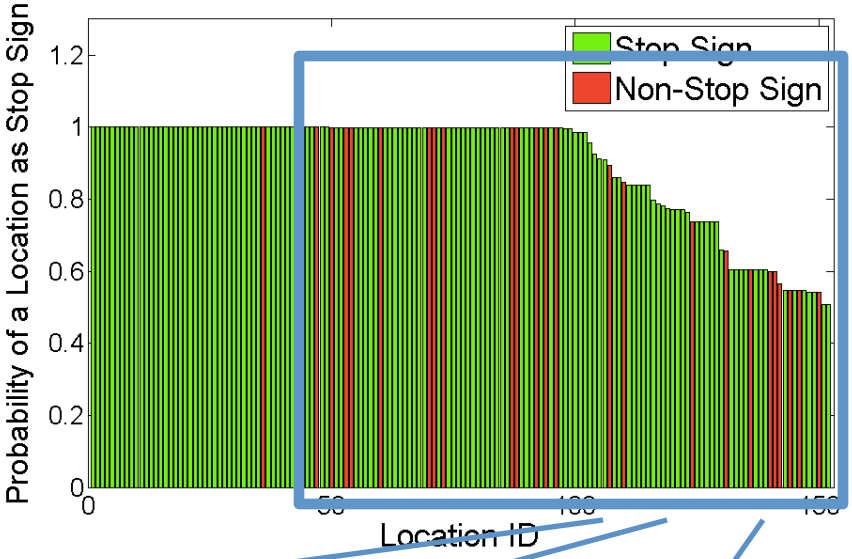


Traffic Light Location Detection

Experiment setup:
34 drivers, **300** hours of driving in Urbana-Champaign
1,048,572 GPS readings, **4865** claims generated by phone (3033 for stop signs, 1562 for traffic lights)

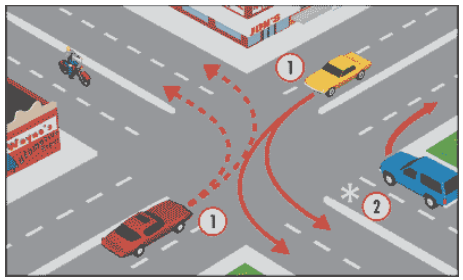
Traffic Regulator Mapping From GPS Data





Source Reliability Estimation

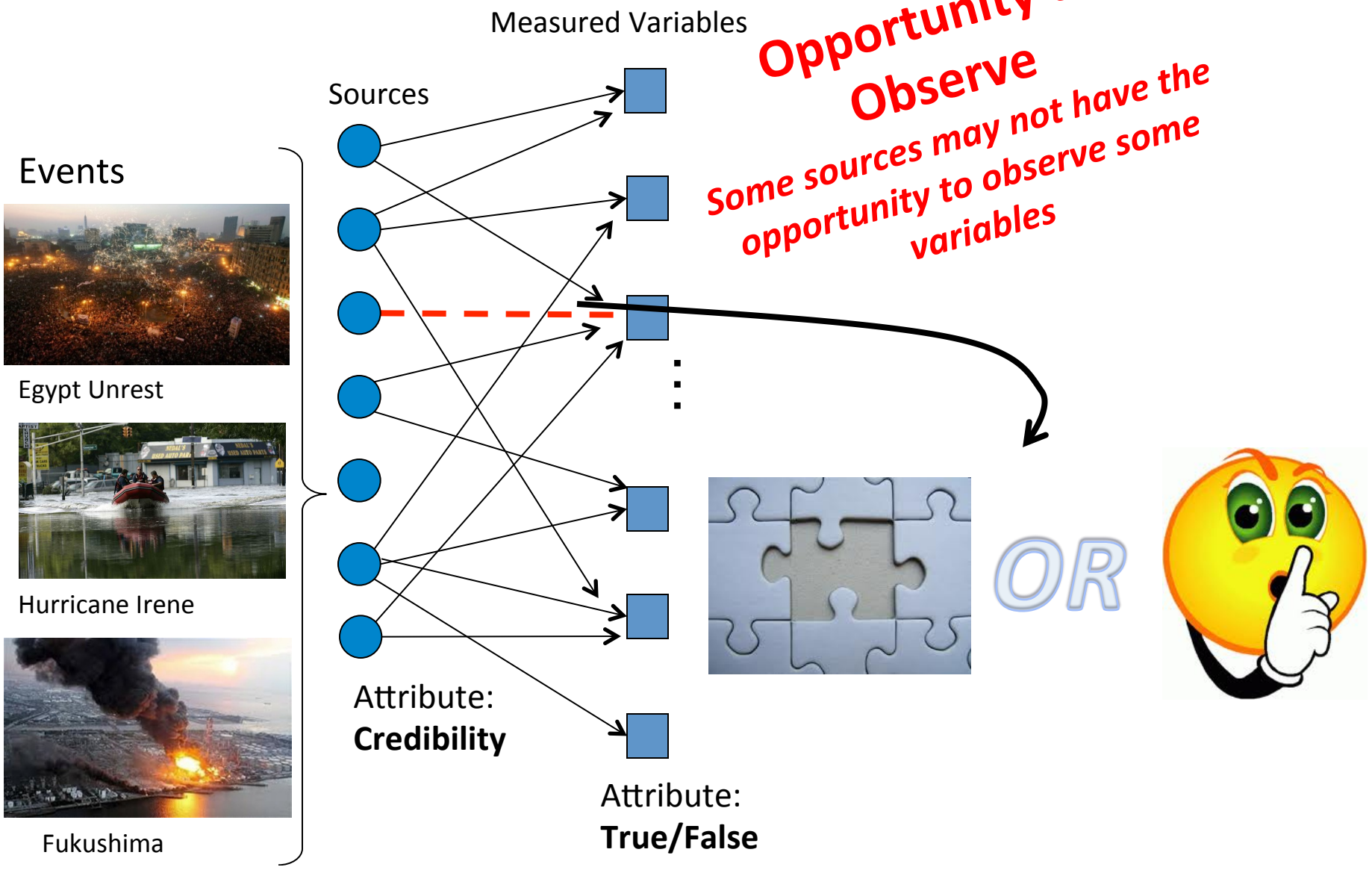
Stop Sign Location Detection



Surrogate Sensing

- Can we use simple sensors (e.g., on our phones) to sense much more complex concepts?
- Examples:
  - Detect free parking spots
  - Find popular places in a city
  - Predict bus arrival time
  - Monitor noise and urban environments

Incorporate Prior Knowledge



Impact of the Opportunity to Observe

Basic Definitions

$$t_i = P(C_j^t | S_i C_j, S_i \text{ knows } C_j)$$

$$a_i = P(S_i C_j | C_j^t, S_i \text{ knows } C_j)$$

$$b_i = P(S_i C_j | C_j^f, S_i \text{ knows } C_j)$$

$$s_i = P(S_i C_j | S_i \text{ knows } C_j)$$

Rebuilt Likelihood Function

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

$$= \prod_{j=1}^{N} \left\{ \prod_{i \in \mathcal{S}_j} a_i^{S_i C_j} (1 - a_i)^{(1 - S_i C_j)} \times d_j \times z_j \right.$$

$$\left. + \prod_{i \in \mathcal{S}_j} b_i^{S_i C_j} (1 - b_i)^{(1 - S_i C_j)} \times (1 - d_j) \times (1 - z_j) \right\}$$

$\mathcal{S}_j$: Set of sources knows C_j

Expectation Step (E-Step)

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

$$= \sum_{j=1}^{N} \left\{ p(z_j = 1|X_j, \theta^{(t)}) \right.$$

$$\times \left[\sum_{i \in \mathcal{S}_j} (S_i C_j \log a_i + (1 - S_i C_j) \log(1 - a_i) + \log d_j) \right]$$

$$+ p(z_j = 0|X_j, \theta^{(t)})$$

$$\left. \times \left[\sum_{i \in \mathcal{S}_j} (S_i C_j \log b_i + (1 - S_i C_j) \log(1 - b_i) + \log(1 - d_j)) \right] \right\}$$

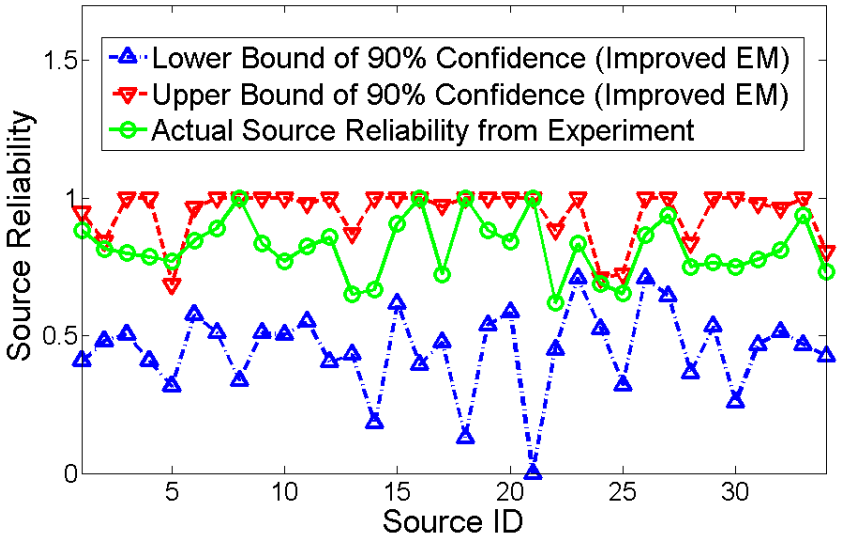
Maximization Step (M-Step)

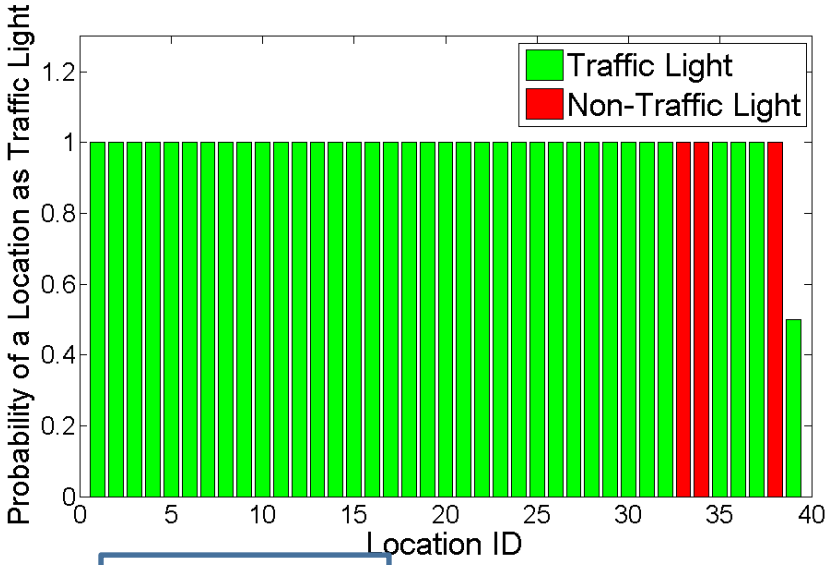
$$a_i^{(t+1)} = a_i^* = \frac{\sum_{j \in SJ_i} Z(t, j)}{\sum_{j \in C_i} Z(t, j)}$$

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

$$d_j^{t+1} = d_j^* = Z(t, j)$$

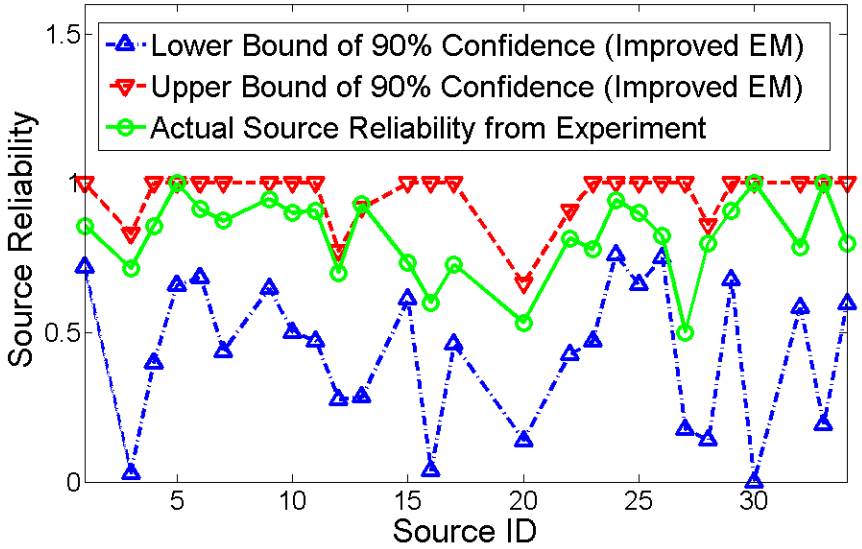
Traffic Regulator Detection (Enhanced)

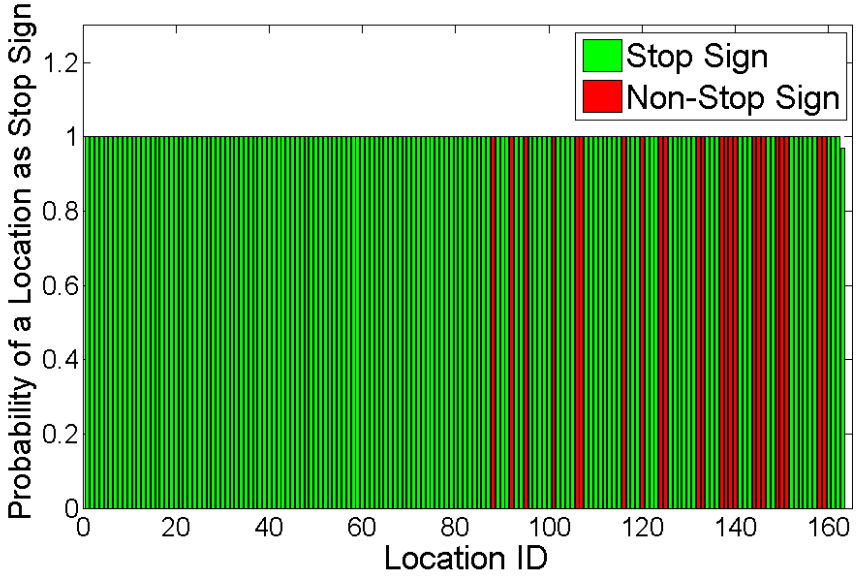




	Original EM	Improved EM
Average Source Reliability Estimation Error	10.19%	**7.74%**
Number of Unbounded Sources	3	**1**
Number of Correctly Identified Traffic Lights	31	**36**
Number of Mis-Identified Traffic Lights	**2**	3

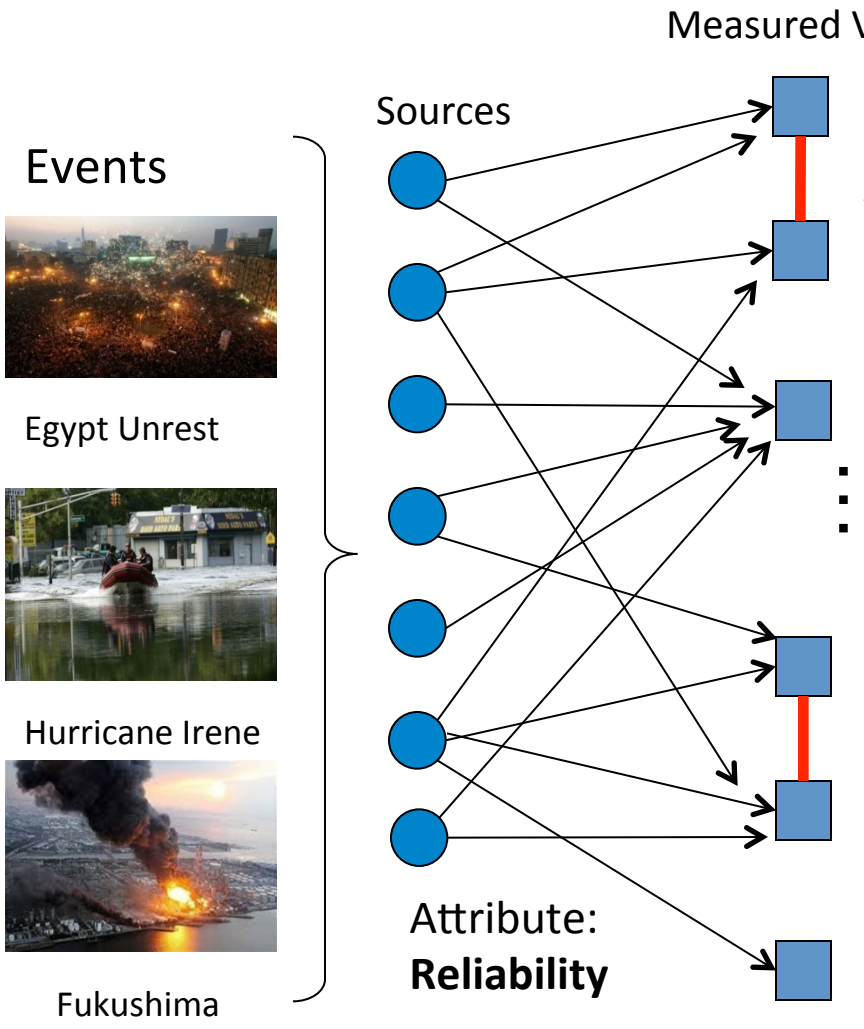
Traffic Regulator Detection (Enhanced)





	Original EM	Improved EM
Average Source Reliability Estimation Error	20.06%	**14.32%**
Number of Unbounded Sources	5	**1**
Number of Correctly Identified Traffic Lights	127	**139**
Number of Mis-Identified Traffic Lights	25	**24**

Relations between Variables

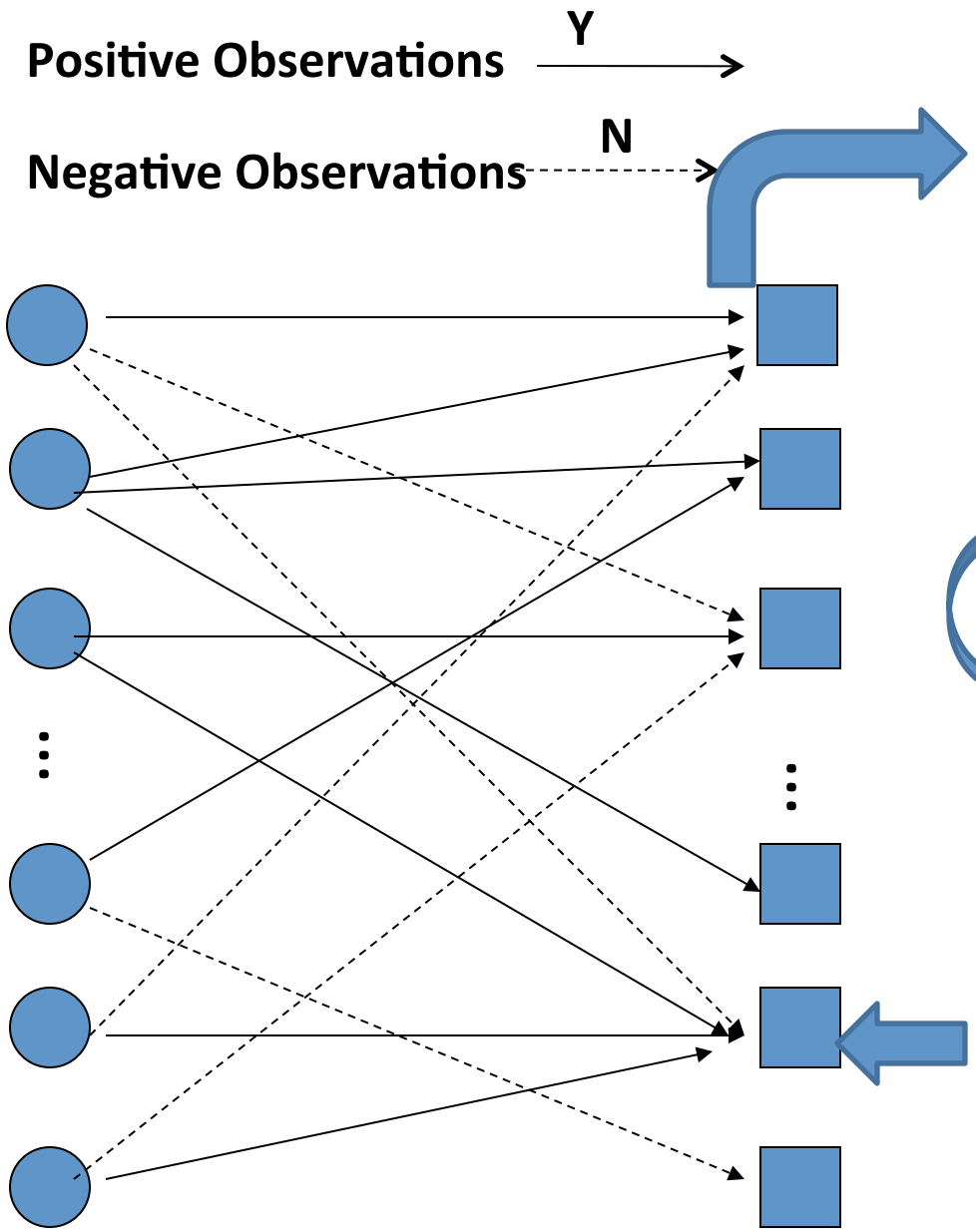


Variables are not independent !

Some variables are contradictory and cannot be true at the same time

D. Wang, etc, ACM ToSN 14

MLE with Conflicting Variables



Expectation Maximization

$$\theta = (a_1^T, a_2^T, ..., a_M^T, a_1^F, a_2^F, ..., a_M^F, b_1^T, b_2^T, ..., b_M^T, b_1^F, b_2^F, ..., b_M^F, d)$$

Maximum Likelihood Estimation

Source Reliability

Measured Variable Correctness

Hypothesis that is mostly consistent with observations

MLE with Conflicting Variables

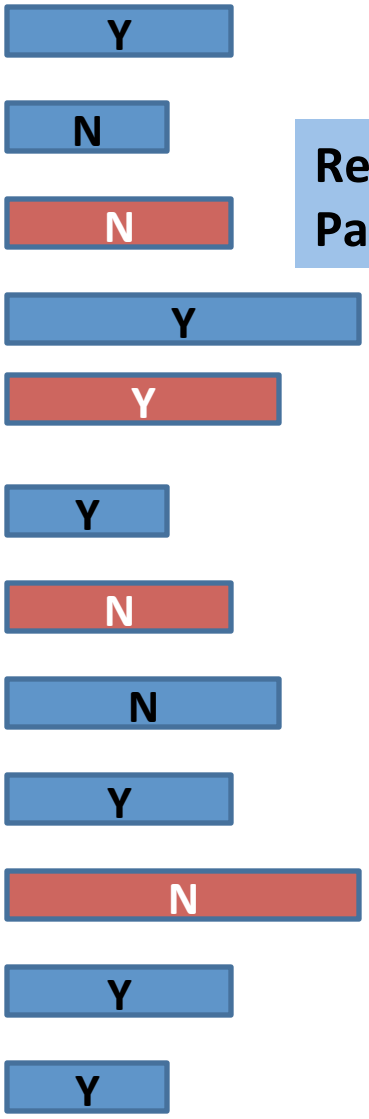
Basic Definition

True Measured Variables

False Measured Variables

Y: Positive Observation

N: Negative Observation



Reliability of Participant *i*

$$= \frac{Y_i}{Y_i + Y_i}$$

$$= \frac{N_i}{N_i + N_i}$$

Participant Reliability

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

$$t_i^F = P(C_j^f \mid S_i C_j^f)$$

$S_i C_j^t$: participant i reports measured variable j to be true

$S_i C_j^f$: participant i reports measured variable j to be false

MLE with Conflicting Variables

Basic Definition

True Measured Variables

False Measured Variables

Y: Positive Observation

N: Negative Observation

Y
N
N
Y
Y
Y
N
N
Y
N
Y
Y

Positive Speak Rate of Participant *i*

$$\propto \frac{Y_i + Y_i}{All + All}$$

Participant i reports positive observations with rate s_i^T

$$s_i^T = P(S_i C_j^t)$$

Negative Speak Rate of Participant *i*

$$\propto \frac{N_i + N_i}{All + All}$$

Participant i reports negative observations with rate s_i^F

$$s_i^F = P(S_i C_j^f)$$

MLE with Conflicting Variables

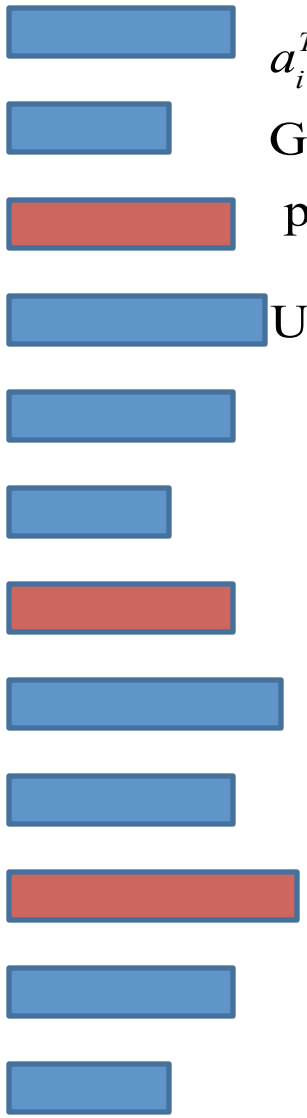
Basic Definition

True Measured Variables

False Measured Variables

Y: Positive Observation

N: Negative Observation



a_i^T

Y

a_i^F N

$a_i^T = P(S_i C_j^t \mid C_j^t)$

Given a measured variable is true, the odds participant S_i will report it positively

Using Bayes Theorem: $a_i^T = \frac{t_i^T \times s_i^T}{d}$

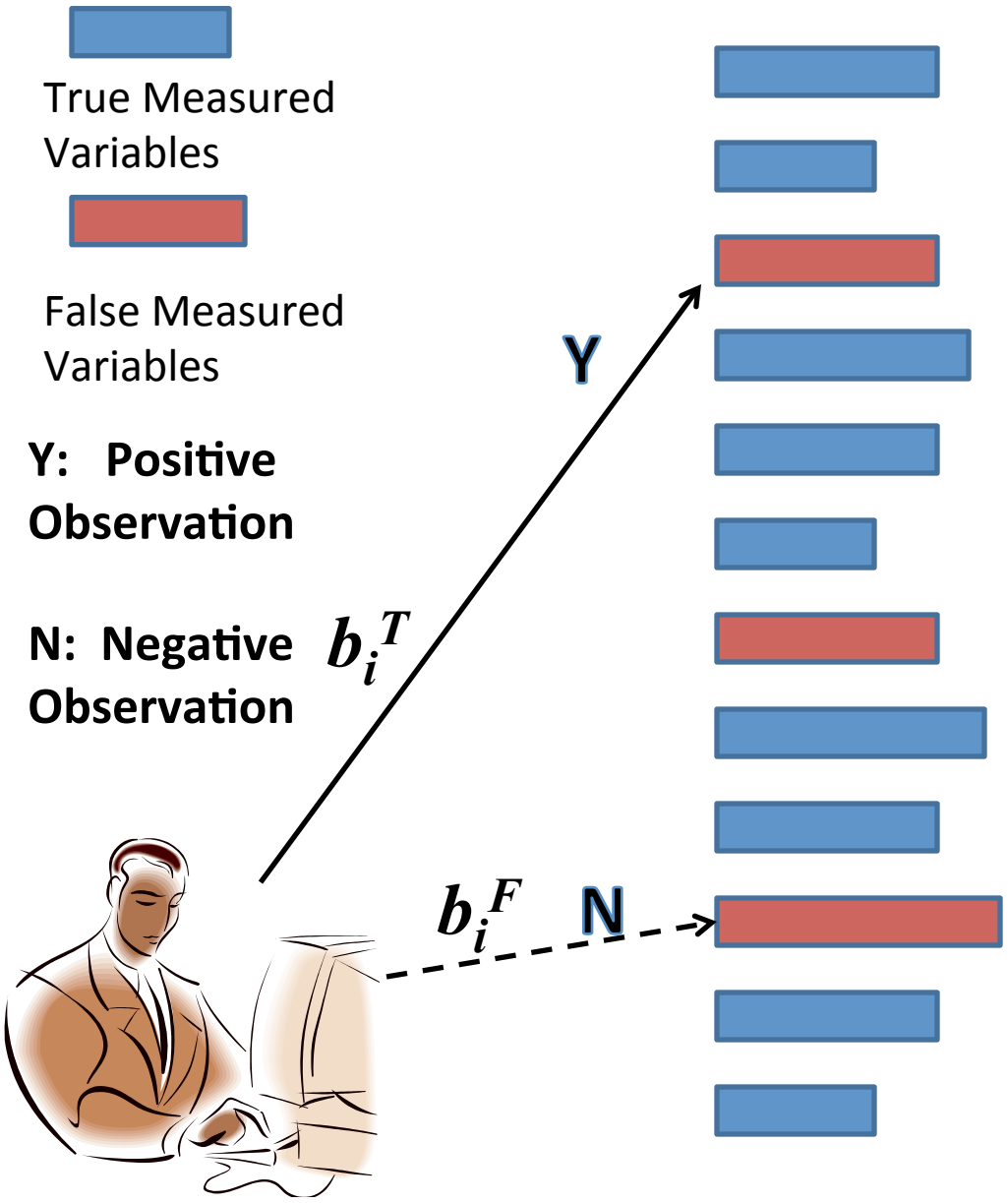
$a_i^F = P(S_i C_j^f \mid C_j^t)$

Given a measured variable is true, the odds that participant S_i will report it negatively

Using Bayes Theorem: $a_i^F = \frac{(1 - t_i^F) \times s_i^F}{d}$

MLE with Conflicting Variables

Basic Definition



$$b_i^T = P(S_i C_j^t \mid C_j^f)$$

Given a measured variable is false, the odds that participant S_i will report it positively

Using Bayes Theorem: $b_i^T = \frac{(1 - t_i^T) \times s_i^T}{1 - d}$

$$b_i^F = P(S_i C_j^f \mid C_j^f)$$

Given a measured variable is false, the odds that participant S_i will report it negatively

Using Bayes Theorem: $b_i^F = \frac{t_i^F \times s_i^F}{1 - d}$

Approach: EM with Conflicting Variables

Likelihood function of Extended EM

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

Expectation Step (E-Step)

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

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

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

Iterate

Maximization Step (M-Step)

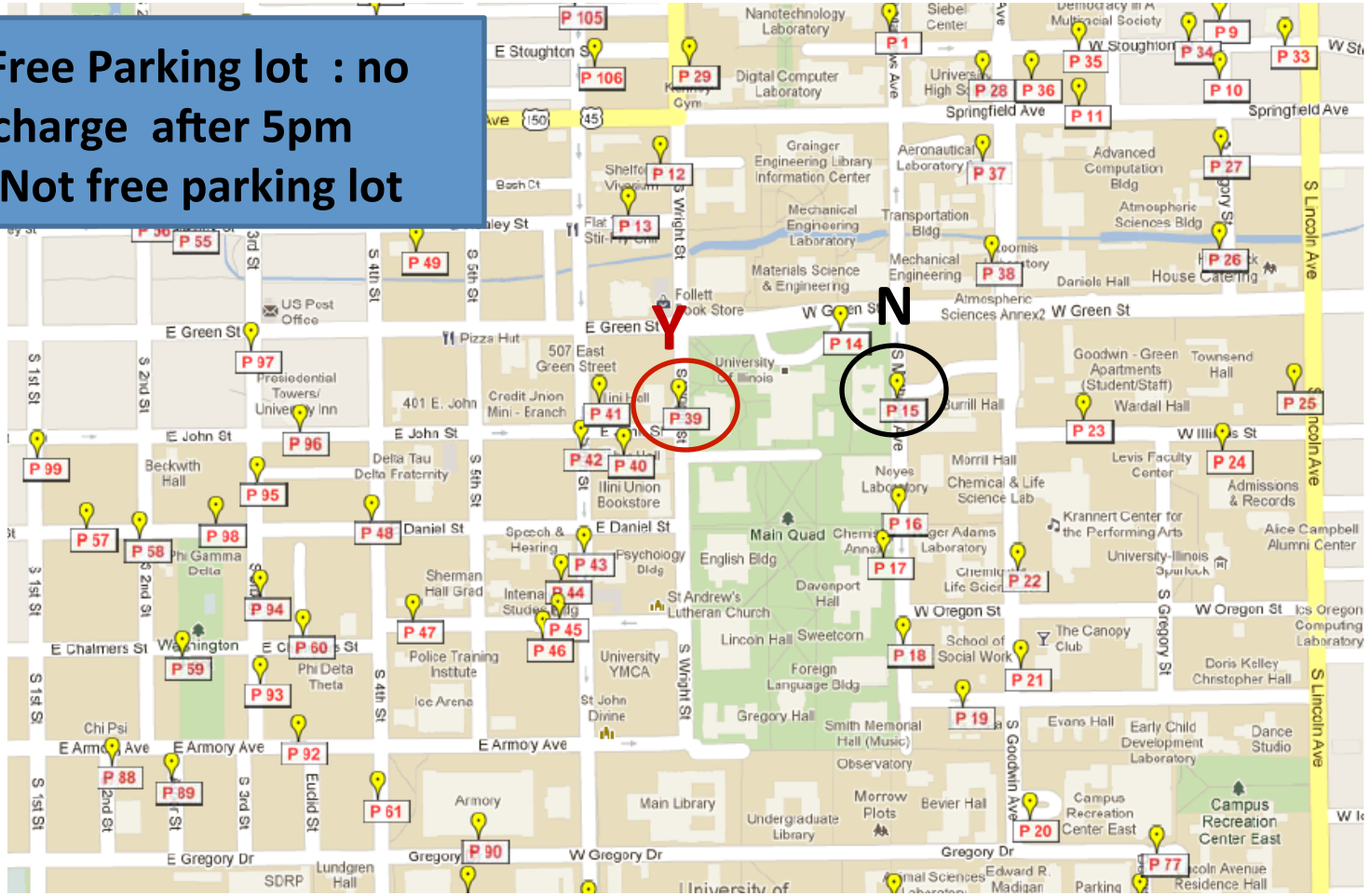
$$a_i^{T(t+1)} = a_i^{T*} = \frac{\sum_{j \in SJ_i^T} Z(t, j)}{\sum_{j=1}^{N} Z(t, j)} \qquad a_i^{F(t+1)} = a_i^{F*} = \frac{\sum_{j \in SJ_i^F} Z(t, j)}{\sum_{j=1}^{N} Z(t, j)}$$

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

Free Parking Lot Identification from GeoTag Data

Goal: Find Free Parking Lots on UIUC Campus

Y: Free Parking lot : no charge after 5pm
N: Not free parking lot



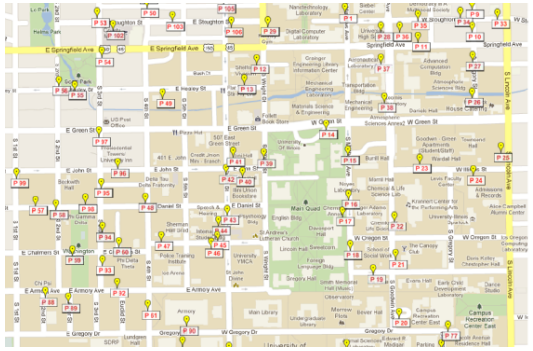
Free Parking Lot Identification from GeoTag Data

Results of Extended EM vs Baselines

Table I. Accuracy of Finding Free Parking Lots on Campus

Schemes	False Positives	False Negatives
EM-Conflict	**6.67%**	**10.87%**
EM-Regular	11.67%	17.39%
Average-Log	16.67%	19.57%
Truth-Finder	18.33%	15.22%
Voting	21.67%	23.91%

Experiment setup:
106 parking lots of interests, **46** indeed free
30 participants, **901** marks collected



Address Variable Constraints

Events

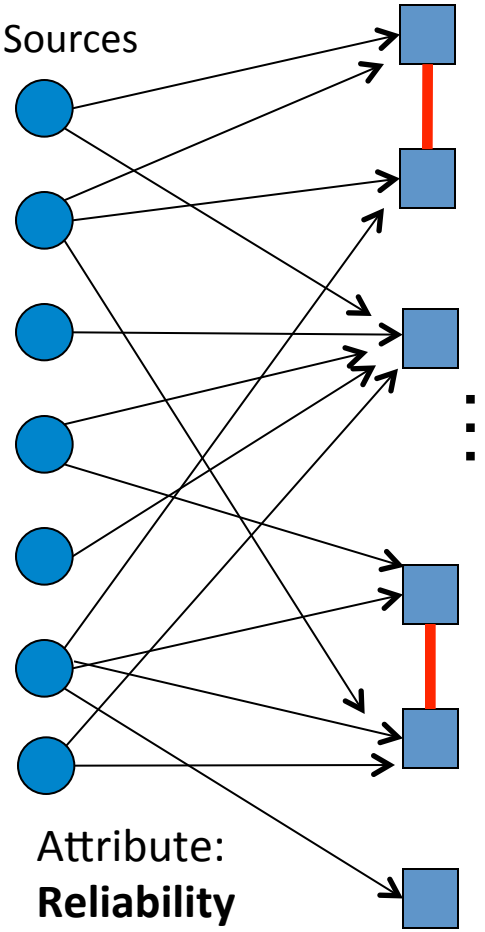
Hurricane Sandy

Boston Marathon Explosion

Egypt President Arrest

Measured Variables

Sources

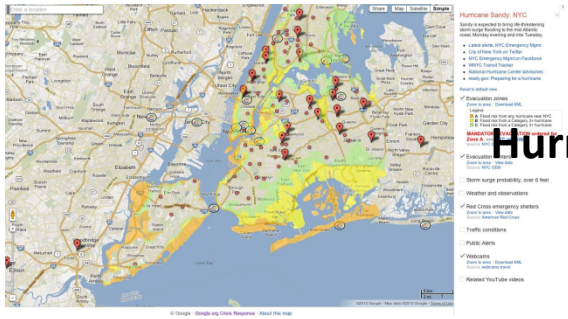


Attribute:
Reliability

Attribute:
True/False

Claims constraints can be general!

Road Speed Map



Hurricane Risk Map

D. Wang, et al., RTSS, 2013

Approach: EM with General Correlated Claims and Opportunity to Observe

Likelihood function of Extended EM

$$L(\theta; X, Z) = \prod_{g \in G} p(X_g, Z_g|\theta) = \prod_{g \in G} p(Z_g) \times p(X_g|Z_g, \theta)$$

$$= \prod_{g \in G} \left\{ \sum_{g_1, \ldots, g_k \in \mathcal{Y}_g} p(z_{g_1}, \ldots, z_{g_k}) \prod_{i \in S_j} \prod_{j \in c_g} \alpha_{i,j} \right\}$$

General Claim Correlations

Expectation Step (E-Step)

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

$$= \sum_{g \in G} p(z_{g_1}, \ldots, z_{g_k}|X_g, \theta^{(t)})$$

$$\times \left\{ \sum_{i \in SJ_j} \sum_{j \in c_g} \log \alpha_{i,j} + \log p(z_{g_1}, \ldots, z_{g_k}) \right\}$$

$$Z(t,j) = f(a_i^{(t)}, b_i^{(t)}, d^{(t)}, j)$$

Maximization Step (M-Step)

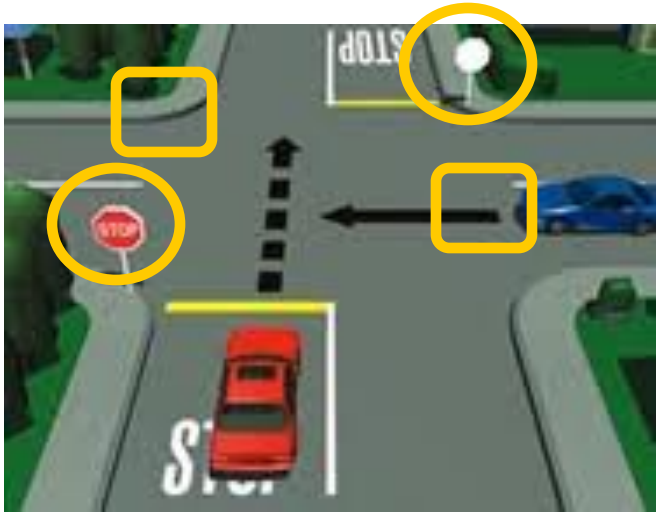
$$a_i^{(t+1)} = a_i^* = \frac{\sum_{j \in SJ_i} Z(t,j)}{\sum_{j \in \mathcal{C}_i} Z(t,j)}$$

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

$$d_j^{t+1} = d_j^* = Z(t,j)$$

Iterate

Traffic Regulator Mapping From GPS Data



A = stop sign 1 exists; B = stop sign 2 exists	Percentage
p(A,B)	36%
p(not A, not B)	49%
p(A,not B) = p(not A, B)	7.5%

	Regular EM	OtO EM	DV EM	DV+OtO EM
Average Source Reliability Estimation Error	25.34%	16.75%	15.99%	11.98%
Number of Correctly Identified Stop Signs	127	139	141	146
Number of Mis-Identified Stop Signs	25	24	29	25

Address Source Dependency

Events

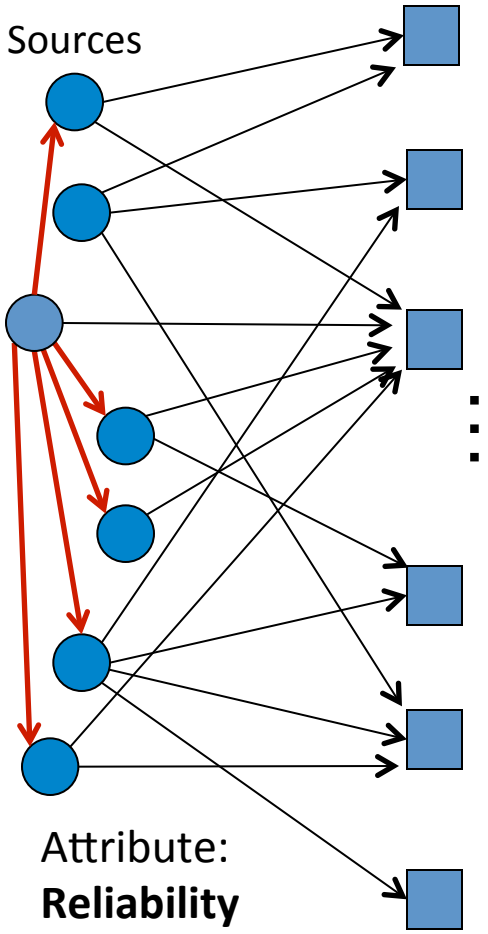
Egypt Unrest

Hurricane Irene

Fukushima

Measured Variables

Sources



Attribute:
Reliability

Attribute:
True/False

Sources are not independent ! Consider the social network and source forwarding behaviors

Examples

Examples of Twitter Observations
Crash blocking lanes on I-5S @ McBean Pkwy in Santa Clarita
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!

Egypt Unrest, 2011

Hurricane Sandy, 2012

Boston Marathon Explosion,2013

Chile Earthquake, 2014

Humans as Sensors Challenges

- How to Model Networked Humans as Participatory Sensors?
- How to Filter Out “Bad Data” from Human Sensors?
- How Good is the Necessarily Simplified Model?

A Binary Human Sensor Model

1: Model Humans as Sensors of

- Unknown Reliability
- Binary Observations
- Uncertain Data Provenance

2: Build An Estimation-theoretic Framework to

- Solve the Reliable Sensing Problem
- Validate through Real-world Twitter Traces

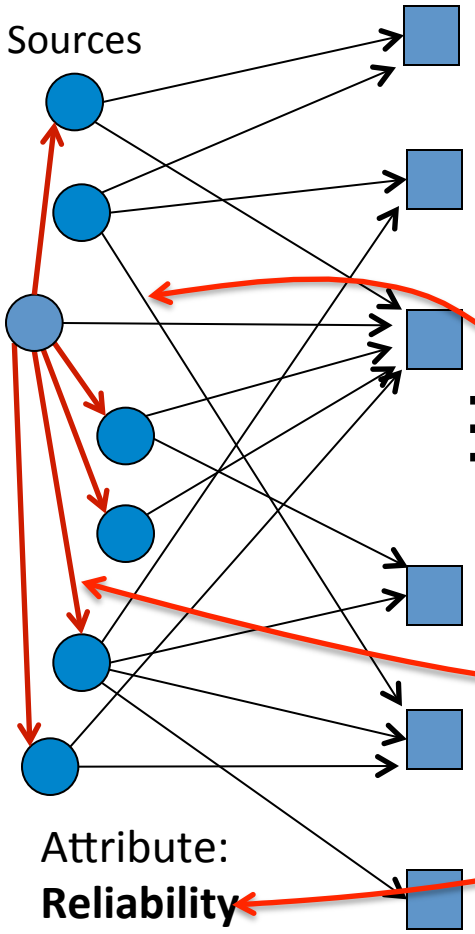
Formulate the Likelihood Function

Events

Hurricane Sandy

Boston Marathon Explosion

Egypt President Arrest



SC: Source Variable Graph

SD: Social Dissemination Graph

$$P(SC|SD,\theta) = \sum_{z} P(SC,z|SD,\theta)$$

D. Wang, et al., IPSN 2014

Social Expectation Maximization (1/2)

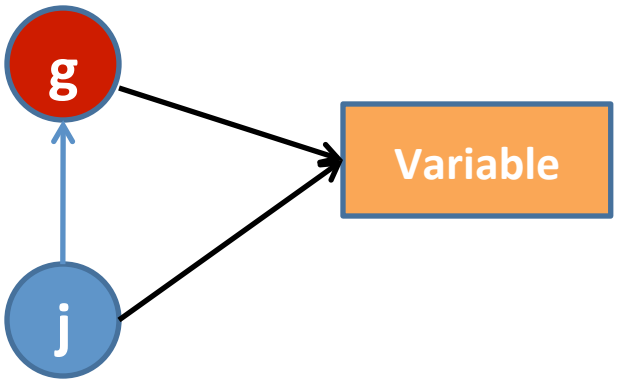
Likelihood Function Incorporating Source Dependency

$$P(SC, z|SD, \theta) = \prod_{j=1}^{N} P(z_j) \times \{ \prod_{g \in M_j} P(S_g C_j|\theta, z_j) \prod_{i \in c_g} P(S_i C_j|S_g C_j) \}$$

$$P(z_j) = \begin{cases} d & z_j = 1 \\ (1-d) & z_j = 0 \end{cases}$$

$$P(S_g C_j|\theta, z_j) = \begin{cases} a_g & z_j = 1, S_g C_j = 1 \\ (1-a_g) & z_j = 1, S_g C_j = 0 \\ b_g & z_j = 0, S_g C_j = 1 \\ (1-b_g) & z_j = 0, S_g C_j = 0 \end{cases}$$

$$P(S_i C_j|S_g C_j) = \begin{cases} p_{ig} & S_g C_j = 1, S_i C_j = 1 \\ 1-p_{ig} & S_g C_j = 1, S_i C_j = 0 \end{cases}$$



Dependent Sources

Social Expectation Maximization (2/2)

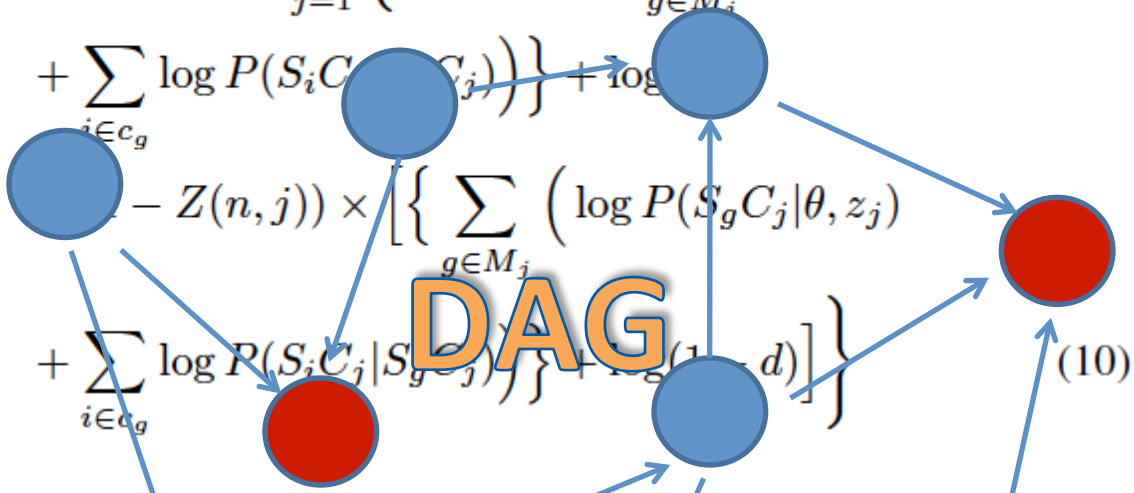
E-Step

$$Q\left(\theta|\theta^{(n)}\right) = \sum_{j=1}^{N}\left\{Z(n,j)\times\left[\left\{\sum_{g\in M_j}\left(\log P(S_gC_j|\theta,z_j)\right.\right.\right.$$

$$\left.\left.\left.+\sum_{i\in c_g}\log P(S_iC_j|\theta,z_j)\right)\right\}+\log\right.$$

$$\left.-Z(n,j))\times\left[\left\{\sum_{g\in M_j}\left(\log P(S_gC_j|\theta,z_j)\right.\right.\right.\right.$$

$$\left.\left.\left.\left.+\sum_{i\in c_g}\log P(S_iC_j|S_gC_j)\right)\right\}+\log(1-d)\right]\right\} \quad (10)$$



M-Step

$$a_g^{(n+1)} = a_g^* = \frac{\sum_{j\in SJ_g} Z(n}{\sum_{j=1}^{N} Z(n,j)}$$

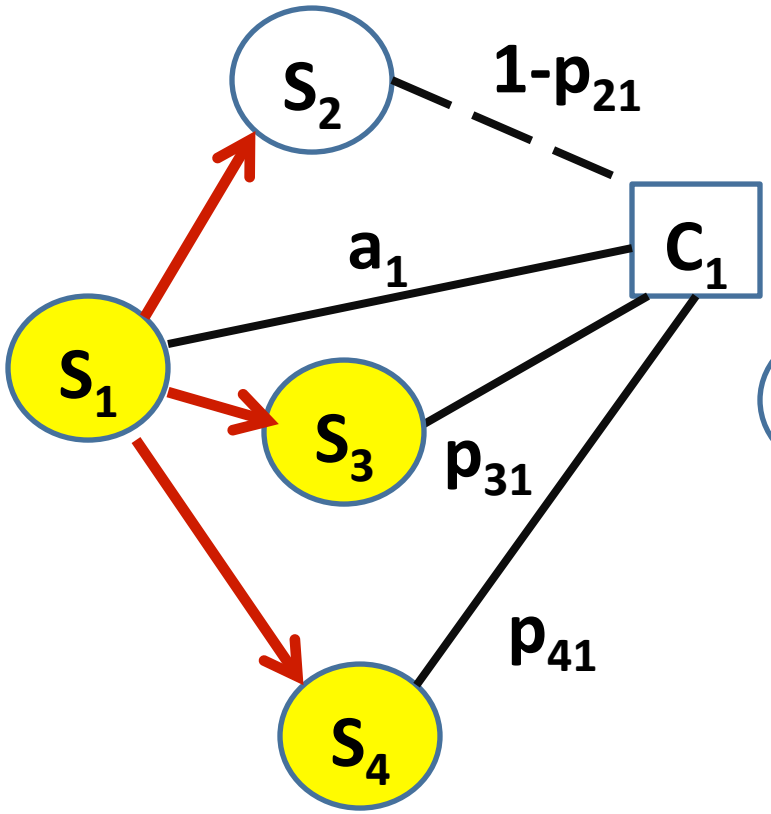
$$a_i^{(n+1)} = a_i^* = \frac{\sum_{j\in \overline{SJ}_g\cap SJ_i} Z(n,j)}{\sum_{j\in \overline{SJ}_g} Z(n,j)} \quad \text{for } i\in c_g$$

$$b_g^{(n+1)} = b_g^* = \frac{\sum_{j\in S} (1-Z(n,j))}{\sum_{j=} Z(n,j))}$$

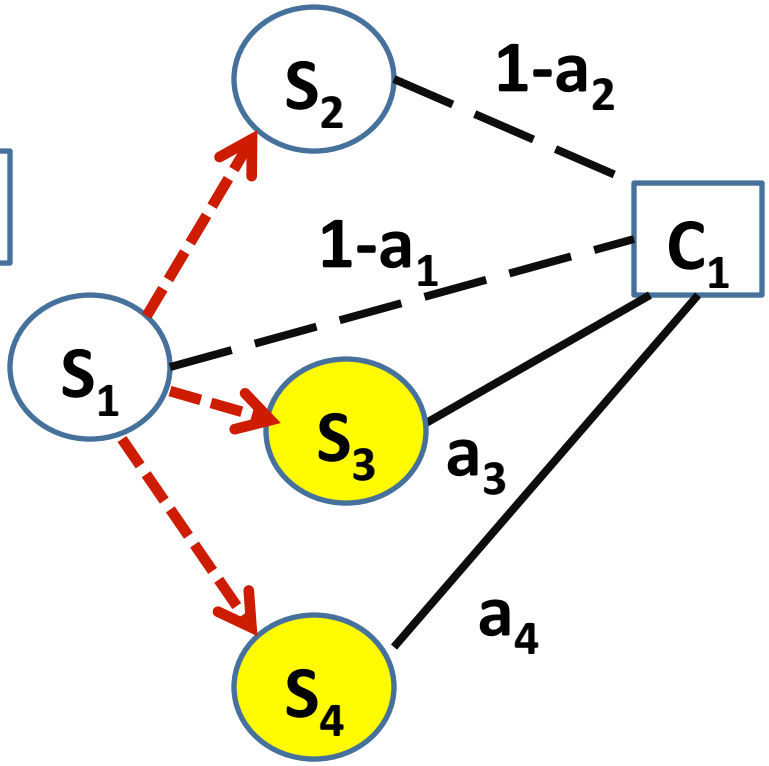
$$b_i^{(n+1)} = b_i^* = \frac{\sum_{j\in \overline{SJ}_g\cap SJ_i}(1-Z(n,j))}{\sum_{j\in \overline{SJ}_g}(1-Z(n,j))} \quad \text{for } i\in c_g$$

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

Simple Illustrative Examples



Example 1



Example 2

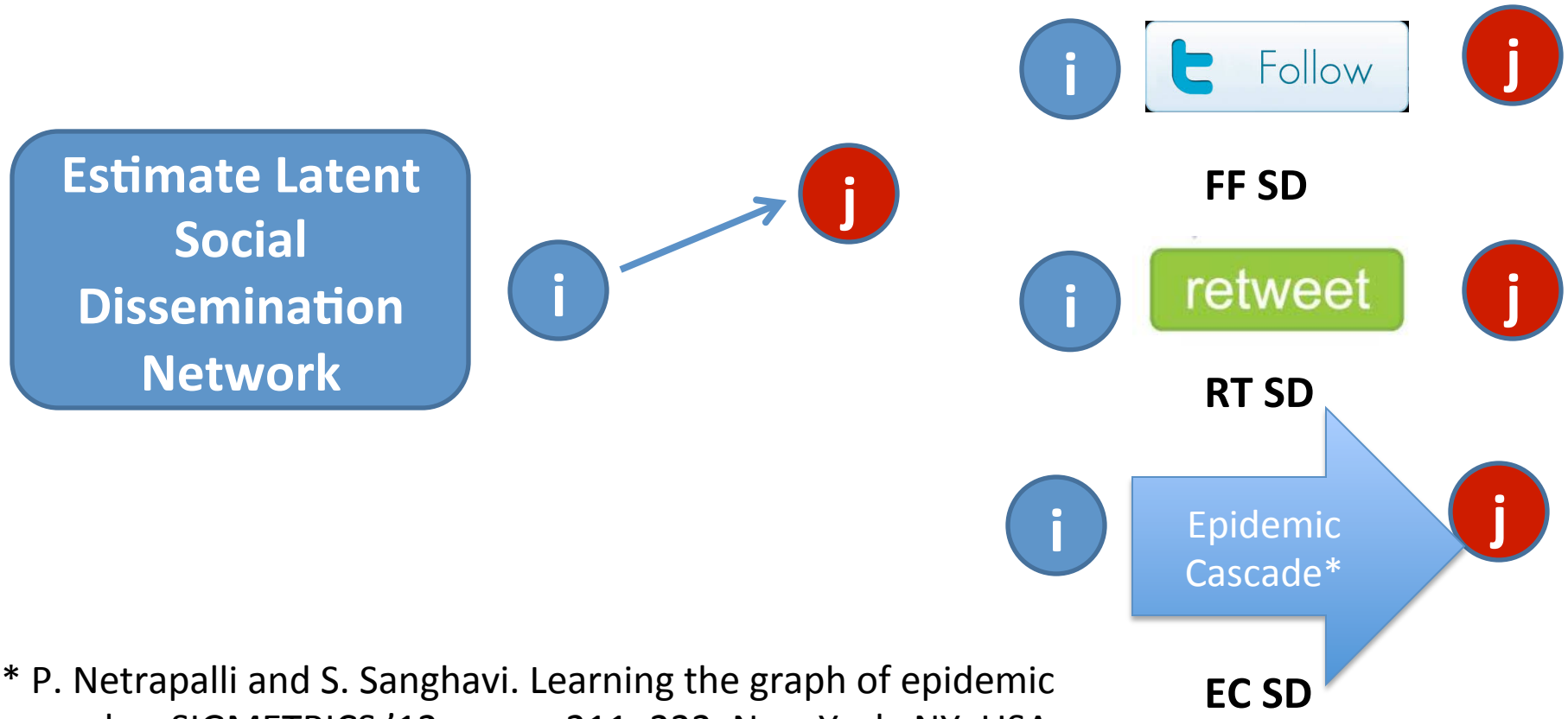
Evaluation

- Integrated Social EM with Apollo
  - Apollo is an Information Distillation Tool for Reliable Social Sensing
- Evaluated Through Real-world Twitter Based Case Studies
  - Hurricane Sandy, Hurricane Irene, Egypt Unrest
- Compared with State-of-the-art Baselines
  - Regular EM in IPSN 12, EM with AD, Voting, etc.

Evaluation using Real Twitter Traces

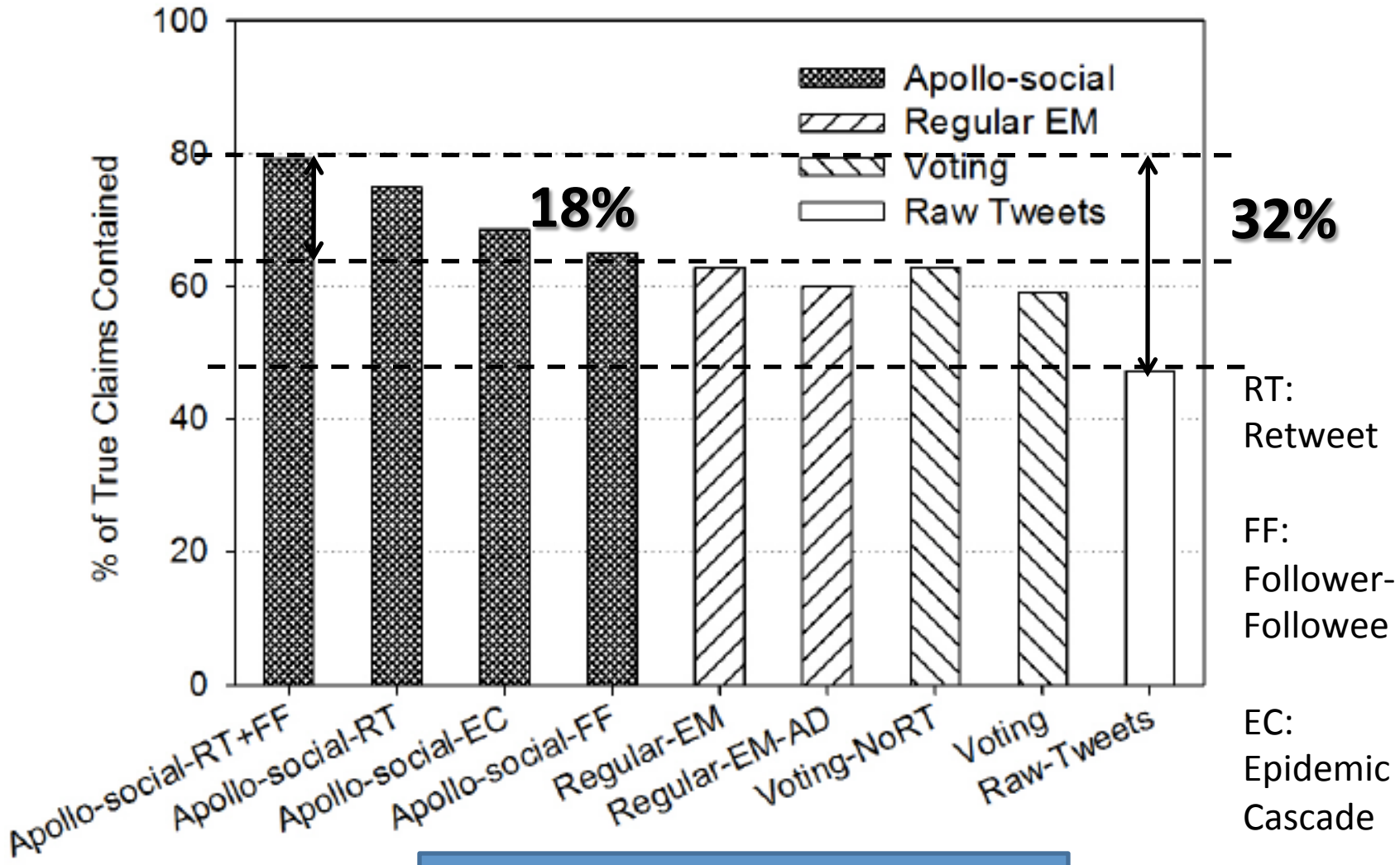
Trace	Hurricane Sandy	Hurricane Irene	Egypt Unrest
Time duration	14 days (Nov. 2-15, 2012)	8 days (Aug.26-Sept. 2, 2011)	18 days (Feb.2-Feb. 19,2011)
Locations	16 cities in East Coasts	New York	Cairo, Egypt
# of users tweeted	7,583	207,562	13,836
# of tweets	12,931	387,827	93,208
# of users crawled in social network	704,941	2,510,316	5,285,160
# of follower-followee links	37,597	3,902,713	10,490,098

Estimate Latent Social Dissemination (SD) Network



* P. Netrapalli and S. Sanghavi. Learning the graph of epidemic cascades. SIGMETRICS '12, pages 211–222, New York, NY, USA, 2012. ACM.

Evaluation on Sandy Trace



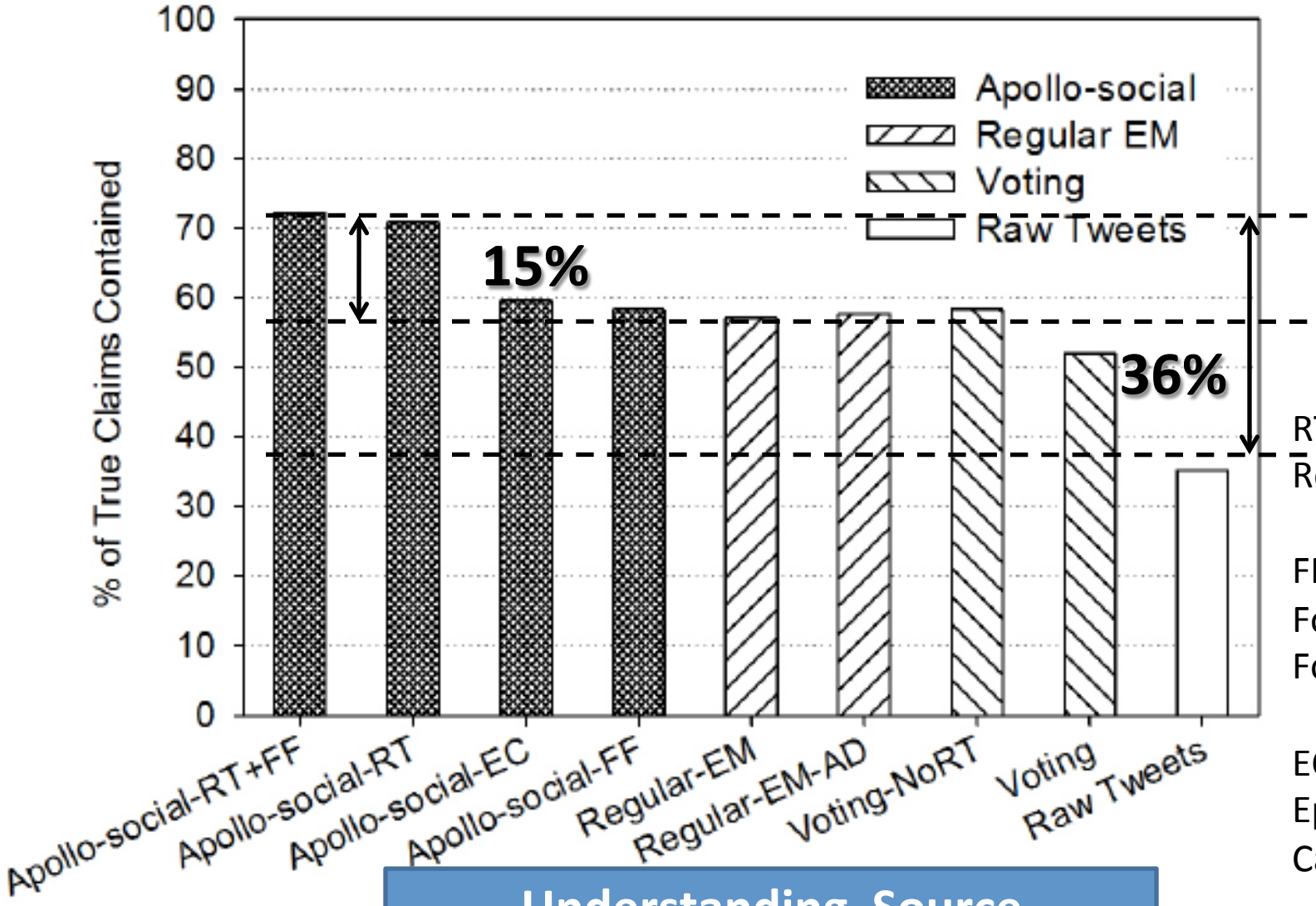
RT: Retweet

FF: Follower-Followee

EC: Epidemic Cascade

Understanding Source Dependency Helps !

Evaluation on Irene Trace



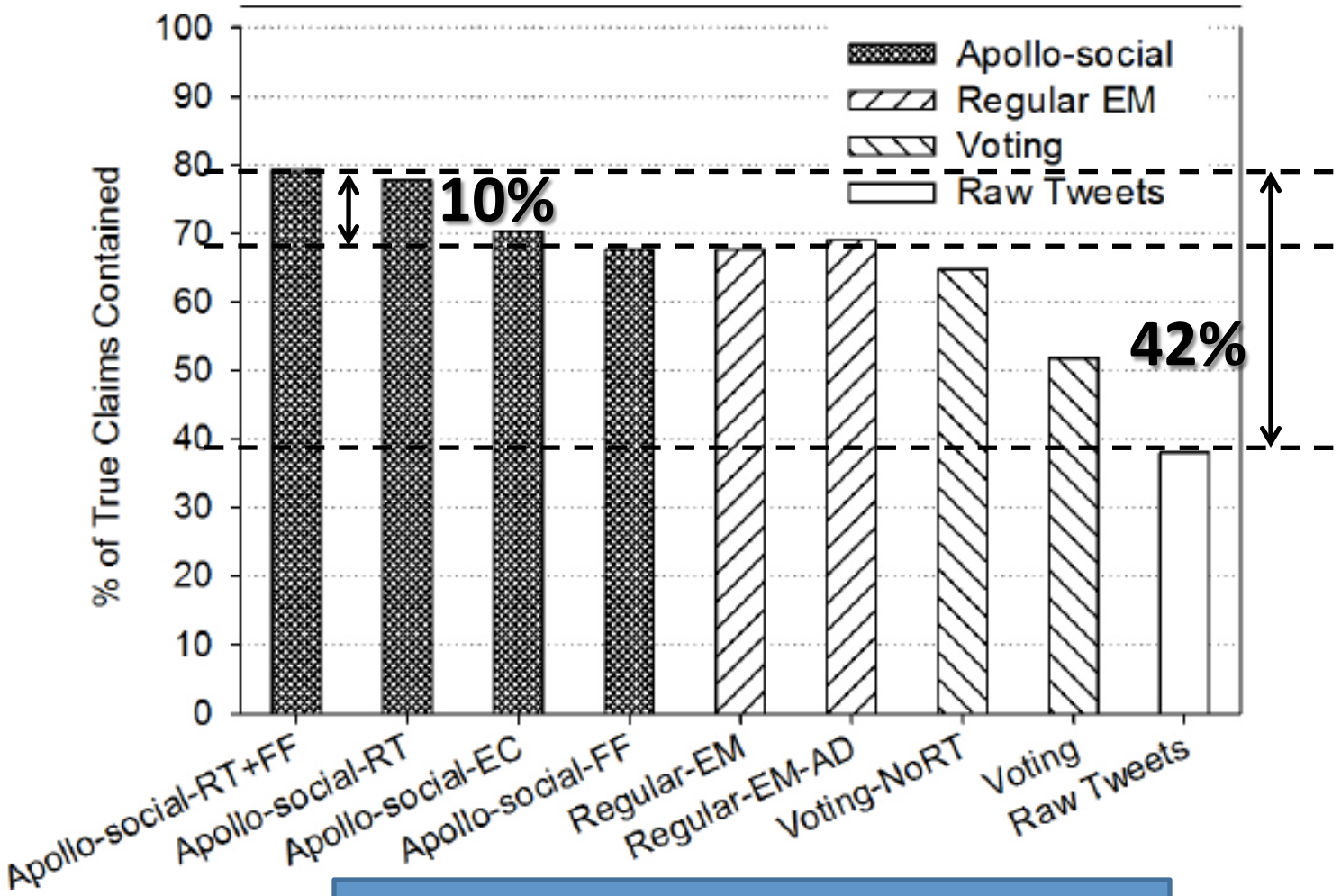
RT: Retweet

FF: Follower-Followee

EC: Epidemic Cascade

Understanding Source Dependency Helps !

Evaluation on Egypt Trace



RT: Retweet

FF: Follower-Followee

EC: Epidemic Cascade

Understanding Source Dependency Helps !

Ground Truth Events Found by Social EM vs Regular EM

#	Media	Tweet found by Apollo-social	Tweet found by Regular EM
1	Rockland County Executive C. Scott Vanderhoef is announcing a Local Emergency Order restricting the amount of fuel that an individual can purchase at a gas station.	Rockland County Orders Restrictions on Gas Sales - Nyack-Piermont, NY Patch http://t.co/cDSrqpa2	**MISSING**
2	New York City Mayor Michael Bloomberg has announced that the city will impose an indefinite program of gas rationing after fuel shortages led to long lines and frustration at the pump in the wake of superstorm Sandy.	Gas rationing plan set for New York City: The move follows a similar announcement last week in New Jersey to eas... http://t.co/nkmF7U9I	RT @nytimes: Breaking News: Mayor Bloomberg Imposes Odd-Even Gas Rationing Starting Friday, as Does Long Island http://t.co/eax7KMVi
3	New Jersey authorities filed civil suits Friday accusing seven gas stations and one hotel of price gouging in the wake of Hurricane Sandy.	RT @MarketJane: NJ plans price gouging suits against 8 businesses. They include gas stations and a lodging provider.	**MISSING**
4	The rationing system: restricting gas sales to cars with even-numbered license plates on even days, and odd-numbered on odd days will be discontinued at 6 a.m. Tuesday, Gov. Chris Christie announced on Monday.	# masdirin City Room: Gas Rationing in New Jersey to End Tuesday # news	RT @nytimes: City Room: Gas Rationing in New Jersey to End Tuesday http://t.co/pYIVOmPo
5	New Yorkers can expect gas rationing for at least five more days: Bloomberg.	Mayor Bloomberg: Gas rationing in NYC will continue for at least 5 more days. @eyewitnessnyc #SandyABC7	Bloomberg: Gas Rationing To Stay In Place At Least Through The Weekend http://t.co/mmqqjYRx

TABLE III. GROUND TRUTH EVENTS AND RELATED CLAIMS FOUND BY APOLLO-SOCIAL VS REGULAR EM IN SANDY

One Interesting Example

Shark in the street!

Suppressed by Social EM

The Washington Post

Posted at 08:53 AM ET, 08/26/2011

Hurricane Irene: 'Photo' of shark swimming in street is fake

By Sarah Anne Hughes

Holy moly! A (fake) picture of a shark swimming on a Puerto Rico street! (Reddit)

http://www.washingtonpost.com/blogs/blogpost/post/hurricane-irene-photo-of-shark-swimming-in-street-is-fake/2011/08/26/gIQABHAvfJ_blog.html

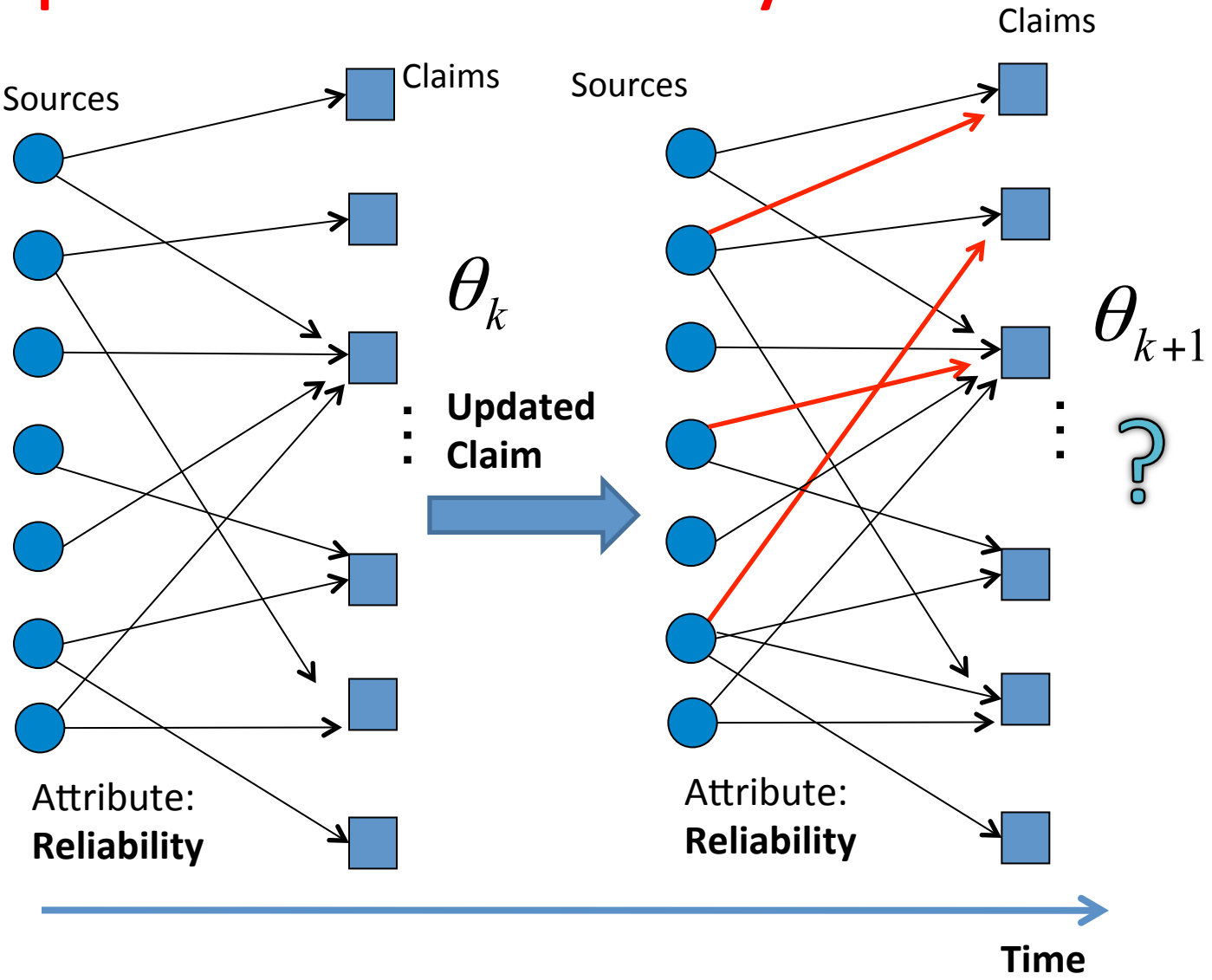
Design for Streaming Data

Update Estimation On the Fly !

Hurricane Sandy

Boston Marathon Explosion

Egypt President Arrest



D. Wang, et al., ICDCS, 2013

Design for Streaming Data

Recursive Formula of EM

Previous Estimate **Updated Observations**

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \{(k+1) I_c(\hat{\theta}_k)\}^{-1} \psi(X_{k+1}, \hat{\theta}_k)$$

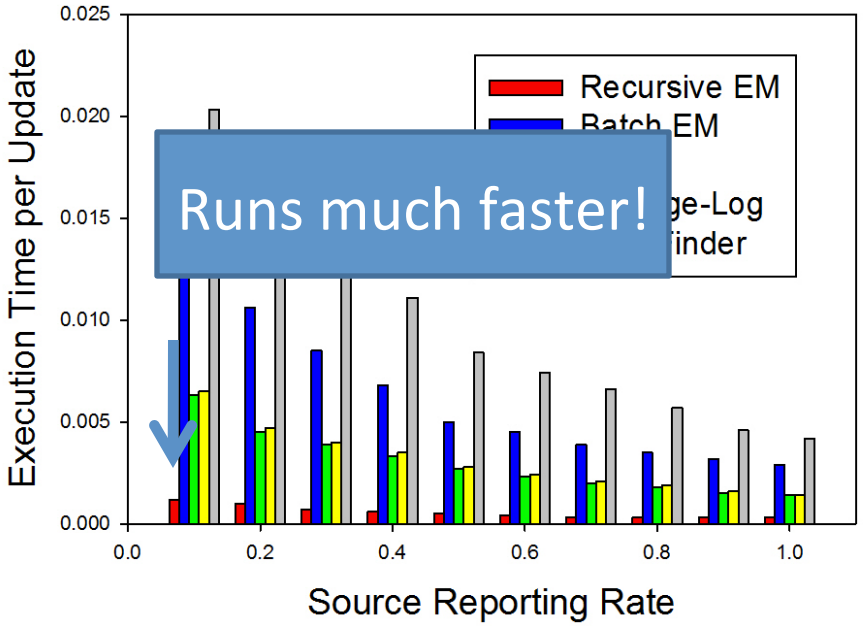
CRLB

Plugging the CRLB derived before

$$\hat{a}_i^{k+1} = \hat{a}_i^k + \frac{1}{Nd(k+1)} \times \left[\sum_{j \in SJ_i^{k+1}} g(\hat{a}_i^k, \hat{b}_i^k, X_k, X_{k+1})(1 - \hat{a}_i^k) - \sum_{j \in \bar{SJ}_i^{k+1}} g(\hat{a}_i^k, \hat{b}_i^k, X_k, X_{k+1}) \hat{a}_i^k \right]$$

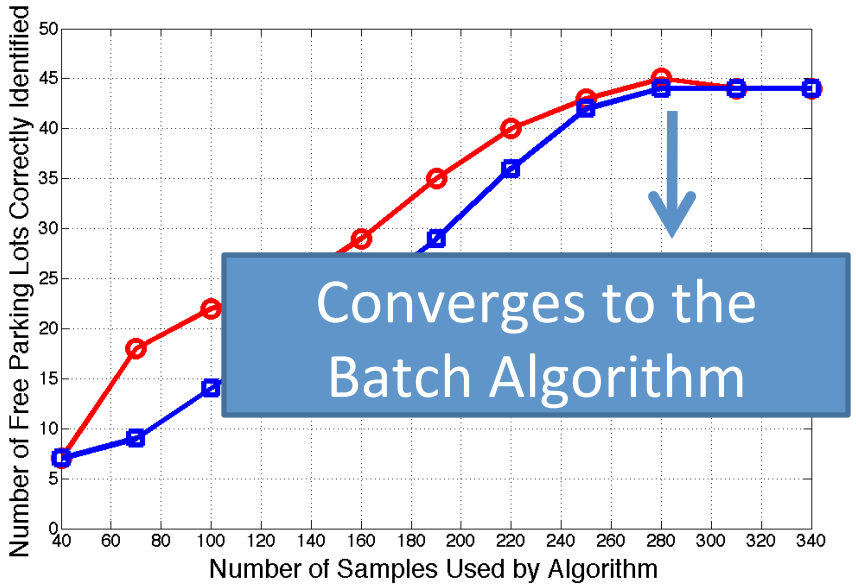
$$\hat{b}_i^{k+1} = \hat{b}_i^k + \frac{1}{Nd(k+1)} \times \left[\sum_{j \in SJ_i^{k+1}} (1 - g(\hat{a}_i^k, \hat{b}_i^k, X_k, X_{k+1}))(1 - \hat{b}_i^k) - \sum_{j \in \bar{SJ}_i^{k+1}} (1 - g(\hat{a}_i^k, \hat{b}_i^k, X_k, X_{k+1})) \hat{b}_i^k \right]$$

Design for Streaming Data



Improved Algorithm Efficiency

Synthetic Dataset



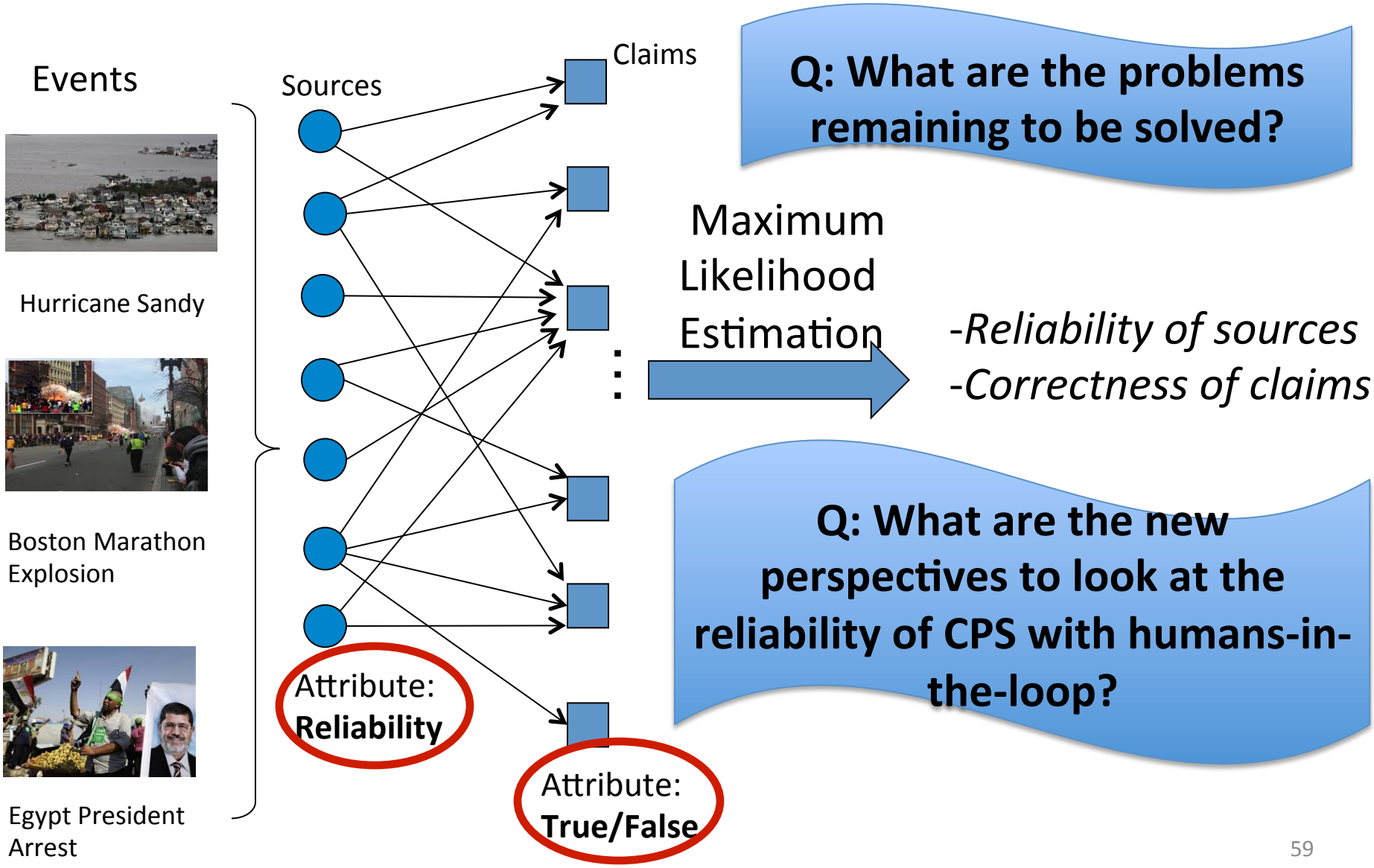
Converge to the Optimal Estimation

Parking Lot Dataset

Human Sensor vs Physical Sensor

Human Sensor	Physical Sensor
Broad spectrum of observations	Narrow spectrum but accurate measurements
Good at reporting binary observations	Good at measuring continuous variables
Unreliable and unknown failure model	Reliable and know failure model
Uncertain data provenance	Certain data provenance
Human specific energy limitation	Battery limitation and other resource constraints
Mostly unstructured data (e.g., text, images)	Mostly structured data
Unvetted sources and often unknown to applications	Verified sensors and often known to applications

What is next ?


Events
Hurricane Sandy
Boston Marathon Explosion
Egypt President Arrest
Sources
Claims
Attribute: Reliability
Attribute: True/False
Maximum Likelihood Estimation
-Reliability of sources
-Correctness of claims
Q: What are the problems remaining to be solved?
Q: What are the new perspectives to look at the reliability of CPS with humans-in-the-loop?

Future Work

- Time dimension is an interesting direction to follow up
  - *Accommodate dynamic states and dependencies of observed variables*
- Measured Variables are not always equal
  - *Generalize the model to handle "hardness" of variables*
- Uncertainty and Bias of Sources
  - *Model the bias and uncertainty of data sources*
- Expertise of Sources
  - *Model the reliability of sources in different expertise dimensions*
- Apply the Model beyond Twitter-based Applications
  - *Apply to a much broader set of crowdsourcing and mobile sensing applications*