Exploitation of Physical Constraints for Reliable Social Sensing

Dong Wang*, Lance Kaplan+, Tarek Abdelzaher*, Raghu Ganti§, Shaohan Hu*, Hengchang Liu*

*Department of Computer Science University of Illinois at Urbana Champaign
+Networked Sensing and Fusion Branch, US Army Research Labs
§IBM Research, Yorktown Heights, NY

RTSS 13, Vancuvour
Social sensing

A set of applications where data are collected from *human sources* or devices on their behalf.
Reliable Social Sensing

Social Sensing with Quantified Data Reliability Guarantees
(Building Reliable Systems on Unreliable Data)

Who to believe?

What to believe?

Numeric data

Images

Smart Devices
Cyber-Physical System Challenges

• Temporal Correctness Guarantees
• Functional Correctness Guarantees
• Data Correctness Guarantees
Cyber-Physical System Challenges

• Temporal Correctness Guarantees
• Functional Correctness Guarantees
• Data Correctness Guarantees
Why Study Data Correctness Guarantees?
Crowdsensing: A Confluence of Three Trends

Mass Dissemination Media
- twitter
- flickr
- facebook
- YouTube
- twitpic

Connectivity
- Cars on Internet
- 4G
- WiMax

Sensors
- Smart Phone
- Smart Meter
- GPS
Problem: How to Explore the Physical Constraints in **Reliable** Social Sensing?

How to explore Physical Constraints for Reliable Social Sensing?
Contributions

1: Explore the Constraints on Sources
   – Consider the opportunity of sources to observe co-located variables

2: Explore the Constraints on Observed Variables (Claims)
   – Consider the Observed Variables may be generally correlated
Related Work on Reliable Social Sensing


Background: Maximum Likelihood Estimation

Events
- Hurricane Sandy
- Boston Marathon Explosion
- Egypt President Arrest

Sources → Claims

Maximum Likelihood Estimation

- Reliability of sources
- Correctness of claims

Probability a claim is true

# of True claims / Total # of claims a source makes

Attribute: Reliability

Attribute: True/False

Unknown a priori!

D. Wang, et al., IPSN, 2012
Contribution 1: Address Source Constraints

Events

- Hurricane Sandy
- Boston Marathon Explosion
- Egypt President Arrest

Sources

Claims

Attribute: Reliability

Attribute: True/False

Opportunity to Observe

OR
Contribution 1: Address Source Constraints

Events

Hurricane Sandy

Boston Marathon Explosion

Egypt President Arrest

Sources

Claims

$L_1$

$L_2$

$L_j$

$L_N$

Attribute: Reliability

Attribute: True/False

$S_i$ Observes $C_j$:
$S_i$ visited location $L_j$

$S_iC_j$ ($S_i$ made a claim $C_j$):
$S_i$ reported $C_j$ was true

$$P(SC|\theta) = \sum_{Z} P(SC, Z|\theta)$$
Contribution 1: Address Source Constraints

Basic Definition

Reliability of Source $i$

Source Reliability

$$ t_i = P(C_j = true | S_iC_j, S_i \text{ Observes } C_j) $$

$S_iC_j$: source $i$ makes claim $j$

Speak Rate of Source $i$

Source $i$ speak with rate $s_i$

$$ s_i = P(S_iC_j | S_i \text{ Observes } C_i) $$
Contribution 1: Address Source Constraints

**Basic Definition**

\[ a_i = P(S_i C_j | C_j = \text{true}, S_i \text{ Observes } C_j) \]

Using Bayesian Theorem: \[ a_i = \frac{t_i \times s_i}{d} \]

where \( d \) is the overall prior that a randomly chosen claim is true

\[ b_i = P(S_i C_j | C_j = \text{false}, S_i \text{ Observes } C_j) \]

Using Bayesian Theorem: \[ b_i = \frac{(1-t_i) \times s_i}{1-d} \]

where \( d \) is the overall prior that a randomly chosen claim is true
Approach: EM with Opportunity to Observe (OtO EM)

Likelihood function of OtO EM

\[
L(\theta; X, Z) = p(X, Z|\theta) = \prod_{j=1}^{N} p(z_j) \times p(X_j|z_j, \theta) = \prod_{j=1}^{N} \prod_{i \in S_j} p(z_j) \times \alpha_{i,j}
\]

where \( S_j \): Set of sources observed \( C_j \)

Expectation Step (E-Step)

\[
Q \left( \theta|\theta^{(t)} \right) = E_{Z|X,\theta^{(t)}} \left[ \log L(\theta; X, Z) \right] = \sum_{j=1}^{N} \left\{ p(z_j = 1|X_j, \theta^{(t)}) \times \sum_{i \in S_j} (\log \alpha_{i,j} + \log d_j) \right. \\
\left. + p(z_j = 0|X_j, \theta^{(t)}) \times \sum_{i \in S_j} (\log \alpha_{i,j} + \log(1 - d_j)) \right\}
\]

\[
Z(t, j) = f \left( a_i^{(t)}, b_i^{(t)}, d^{(t)}, j \right)
\]

Maximization Step (M-Step)

\[
a_i^{(t+1)} = a_i^* = \frac{\sum_{j \in S_i} Z(t, j)}{\sum_{j \in C_i} Z(t, j)}
\]

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

\[
d_j^{(t+1)} = d_j^* = Z(t, j)
\]

\[
d_i^* = \frac{\sum_{j \in C_i} Z(t, j)}{|C_i|}
\]

Iterate
Contribution 2: Address Claim Constraints

Events

- Hurricane Sandy
- Boston Marathon Explosion
- Egypt President Arrest

Sources

Claims

Attribute: Reliability

Attribute: True/False

Road Speed Map

Hurricane Risk Map

Claims constraints can be general!
Approach: EM with Dependent Variables (DV EM)

Likelihood function of OtO EM

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

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

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

Expectation Step (E-Step)

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

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

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

Maximization Step (M-Step)

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

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

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

Iterate
Approach: EM with General Correlated Claims and Opportunity to Observe (OtO+DV EM)

Likelihood function of Extended EM

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

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

Expectation Step (E-Step)

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

\[ = \sum_{g \in G} p(z_{g_1}, \ldots, z_{g_k} | X_g, \theta^{(t)}) \times \left\{ \sum_{i \in S_j} \sum_{j \in c_g} \log \alpha_{i,j} + \log p(z_{g_1}, \ldots, z_{g_k}) \right\} \]

Maximization Step (M-Step)

\[ a_i^{(t+1)} = a_i^* = \frac{\sum_{j \in S_j} Z(t, j)}{\sum_{j \in c_i} Z(t, j)} \]

\[ b_i^{(t+1)} = b_i^* = \frac{\sum_{j \in S_j} (1 - Z(t, j))}{\sum_{j \in c_i} (1 - Z(t, j))} \]

\[ d_j^{(t+1)} = d_j^* = Z(t, j) \]

General Claim Correlations

Opportunity to Observe

\[ Z(t, j) = f(a_i^{(t)}, b_i^{(t)}, d^{(t)}, j) \]

Iterate
Evaluation: Traffic Regular Mapping from Social Sensing Data
Method: Intentionally **Simple** “Sensors”

- **Unreliable “Sensors”!**
  - **Claim Traffic Light**
    - 15-90 s Stop
  - **Claim Stop Sign**
    - 2-10 s Stop
Traffic Regulator Mapping: OtO vs Regular EM

**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)

Average Source Reliability Estimation Error
- Regular EM: 10.19%
- OtO EM: 7.74%

Number of Correctly Identified Traffic Lights
- Regular EM: 31
- OtO EM: 36

Number of Mis-Identified Traffic Lights
- Regular EM: 2
- OtO EM: 3
Traffic Regulator Mapping: OtO vs Regular EM

Source Reliability Estimation

Stop Sign Location Detection

<table>
<thead>
<tr>
<th></th>
<th>Regular EM</th>
<th>OtO EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Source Reliability Estimation Error (Full Dataset)</td>
<td>25.34%</td>
<td>16.75%</td>
</tr>
<tr>
<td>Number of Correctly Identified Stop Signs (Full Dataset)</td>
<td>127</td>
<td>139</td>
</tr>
<tr>
<td>Number of Mis-Identified Stop Signs (Full Dataset)</td>
<td>25</td>
<td>24</td>
</tr>
</tbody>
</table>
Traffic Regulator Mapping: OtO vs Regular EM

Source Reliability Estimation

Stop Sign Location Detection

[Graph showing source reliability estimation over source ID]

<table>
<thead>
<tr>
<th>Source Reliability Estimation Error (Full Dataset)</th>
<th>Regular EM</th>
<th>OtO EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Correctly Identified Stop Signs (Full Dataset)</td>
<td>127</td>
<td>139</td>
</tr>
<tr>
<td>Number of Mis-Identified Stop Signs (Full Dataset)</td>
<td>25</td>
<td>24</td>
</tr>
</tbody>
</table>
Traffic Regulator Mapping: Regular vs DV, DV+OtO EM

<table>
<thead>
<tr>
<th>A = stop sign 1 exists; B = stop sign 2 exists</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(A,B)</td>
<td>36%</td>
</tr>
<tr>
<td>p(not A, not B)</td>
<td>49%</td>
</tr>
<tr>
<td>p(A,not B) = p(not A, B)</td>
<td>7.5%</td>
</tr>
</tbody>
</table>

![Diagram showing traffic signs and arrows]
Smart Road Detection: Regular vs DV, DV+OtO EM

<table>
<thead>
<tr>
<th></th>
<th>Regular EM</th>
<th>OtO EM</th>
<th>DV EM</th>
<th>DV+OtO EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Source Reliability Estimation Error (Full Dataset)</td>
<td>25.34%</td>
<td>16.75%</td>
<td>15.99%</td>
<td>11.98%</td>
</tr>
<tr>
<td>Number of Correctly Identified Stop Signs (Full Dataset)</td>
<td>127</td>
<td>139</td>
<td>141</td>
<td>146</td>
</tr>
<tr>
<td>Number of Mis-Identified Stop Signs (Full Dataset)</td>
<td>25</td>
<td>24</td>
<td>29</td>
<td>25</td>
</tr>
<tr>
<td>Average Source Reliability Estimation Error (75% Dataset)</td>
<td>36.44%</td>
<td>18.2%</td>
<td>18.0%</td>
<td>15.29%</td>
</tr>
<tr>
<td>Number of Correctly Identified Stop Signs (75% Dataset)</td>
<td>92</td>
<td>101</td>
<td>111</td>
<td>116</td>
</tr>
<tr>
<td>Number of Mis-Identified Stop Signs (75% Dataset)</td>
<td>18</td>
<td>23</td>
<td>30</td>
<td>29</td>
</tr>
</tbody>
</table>
Limitations

• Sources are assumed to be independent
  – *Incorporate source dependency in the framework*

• Claims are assumed to be binary
  – *Extend the framework to handle non-binary claims*
Future Work

• Time dimension is an interesting direction to follow up
  – Accommodate dynamic states and dependencies of observed variables
• Claims are not always equal
  – Generalize the model to handle “hardness” of claims
• 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
The Apollo Fact-finder

http://apollo.cs.illinois.edu/

Apollo is a new sensor information processing tool for uncovering likely facts in noisy social (human-centric) sensing data.

Social sensing, where users proactively document and share their observations, has received significant attention in recent years as a paradigm for crowd-sourcing observation tasks. However, it poses interesting challenges in assessing confidence in the information received.

By borrowing clustering and ranking tools from data mining literature, we show how to group data into sets (or claims), corroborating specific events or observations, then iteratively assess both claim and source credibility, ultimately leading to a ranking of described claims by their likelihood of occurrence. Apollo belongs to a category of tools called fact-finders. It is the first fact-finder designed and implemented specifically for social sensing.

This is a collaborative work of
Summary

• Study the *reliable social sensing* problem by exploiting the physical world constraints
• Proposed several truth estimation approaches that consider both constraints on sources and claims
• Evaluated through a real world social sensing application
Q&A

Thank You!