Truth discovery in crowdsourced detection of spatial events

Robin Wentao Ouyang
Mani Srivastava
Alice Toniolo
Timothy J. Norman
Mobile crowdsourced event detection

- Potholes, graffiti, bike racks, flora, …
Truth discovery

• Given crowdsourced detection reports with time and loc tags, find which reported events are true and which are false
Challenges

• Detection reports are non-conflicting

• Uncertainty in both participants’ reliability and mobility
  ▫ Missing reports are ambiguous

• Supervision is difficult
### Possible solutions

<table>
<thead>
<tr>
<th></th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$u_2$</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$u_3$</td>
<td>1</td>
<td></td>
<td>-1</td>
</tr>
<tr>
<td>$u_4$</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$u_5$</td>
<td>-1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

(a) Location tracking

<table>
<thead>
<tr>
<th></th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$u_2$</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$u_3$</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$u_4$</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$u_5$</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

(b) Ignoring missing

<table>
<thead>
<tr>
<th></th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$u_2$</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$u_3$</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$u_4$</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$u_5$</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

(c) Missing as neg.

- **Severe privacy and energy issues**
- **Trivial conclusion**
- **Performance degradation**
Problem

Can we design an algorithm that can reliably discover true events in mobile crowdsourced event detection but \textit{without location tracking} and \textit{supervision}?
Proposed model

- **Graphical model**
- **A participant’s likelihood of reporting an event depends on**
  - 1) whether the participant visited the event location
  - 2) whether the event at that location is true or false
  - 3) how reliable the participant is
Proposed model

• Location popularity
  ▫ For each event at location $l_j$
    • Draw the location’s popularity
      \[ g_j \sim \text{Beta}(\lambda_{g_j,1}, \lambda_{g_j,0}) \]
Proposed model

• Participants Location visit indicators
  ▫ For participant \( u_i \) and event at location \( l_j \)
    • Draw a location visit indicator \( h_{i,j} \sim \text{Bernoulli}(g_j) \)

• A participant has a higher chance to visit more popular locations
Proposed model

- Event label
  - For each event at location $l_j$
    - Draw the event’s prior truth probability $s \sim \text{Beta}(\lambda_s, 1, \lambda_s, 0)$
    - Draw the event’s label $z_j \sim \text{Bernoulli}(s)$
Proposed model

- Three-way participant reliability
  - For each participant $u_i$
    - Draw her true positive rate while present (TPR)
      \[ a_i \sim \text{Beta}(\lambda_{a_i,1}, \lambda_{a_i,0}) \]
    - Draw her false positive rate while present (FPR)
      \[ b_i \sim \text{Beta}(\lambda_{b_i,1}, \lambda_{b_i,0}) \]
    - Draw her reporting rate while absent (RRA)
      \[ c_i \sim \text{Beta}(\lambda_{c_i,1}, \lambda_{c_i,0}) \]

- Concerns
  - A participant’s reliability depends on: whether she visited the event location and whether the event there is true or false
  - A participant’s TPR and FPR may be asymmetric (reliable vs. conservative participants)
  - A participant must conform to physical constraints (RRA)
Proposed model

• Reports (detection = 1, missing = 0)
  ▫ For participant \( u_i \) and event at location \( l_j \):
    - If \( h_{i,j} = 1 \) and \( z_j = 1 \), draw \( x_{i,j} \sim \text{Bernoulli}(a_i) \)
    - If \( h_{i,j} = 1 \) and \( z_j = 0 \), draw \( x_{i,j} \sim \text{Bernoulli}(b_i) \)
    - If \( h_{i,j} = 0 \), draw \( x_{i,j} \sim \text{Bernoulli}(c_i) \)
Analysis

1) Missing reports are well explained

\[ p(x_{i,j} = 0) = \sum_{k=0}^{1} \sum_{q=0}^{1} p(h_{i,j} = k)p(z_j = q)p(x_{i,j} = 0|h_{i,j} = k, z_j = q) \]

\[ = (1 - g_j)(1 - c_i) + g_j[(1 - s)(1 - b_i) + s(1 - a_i)]. \]

When location popularity \( g_j \rightarrow 1 \), we have

\[ p(x_{i,j} = 0) \rightarrow (1 - s)(1 - b_i) + s(1 - a_i) \]

When location popularity \( g_j \rightarrow 0 \), we have

\[ p(x_{i,j} = 0) \rightarrow 1 - c_i \]
Analysis

2) Location tracking is avoided.
   ▫ Location popularity is a collective rather than a personal measure.
   ▫ Its prior counts need to be estimated only once.
   ▫ It can be jointly learned with other parameters from data.

3) Different aspects of participant reliability are handled.

4) Prior belief can be easily incorporated.
Experiments

• Methods in comparison
  ▫ MV (majority voting)
  ▫ TF (truth finder [1])
  ▫ GLAD (generative model of labels, abilities, and difficulties [2])
  ▫ LTM (latent truth model [3])
  ▫ EM (expectation maximization [4])
  ▫ TSE (truth finder for spatial events) – proposed

Experiments

- Traffic light detection
- A mobility dataset containing time-stamped GPS location traces for 536 taxicabs in SF
  - Spatial area of interest 3.5km x 4.4km – further divided into two subareas
  - Temporal span 25 days
- Detection reports
  - A participant waits for 15-120 seconds
Experiments

- Traffic light detection

<table>
<thead>
<tr>
<th></th>
<th>Area 1</th>
<th></th>
<th></th>
<th>Area 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>pre</td>
<td>rec</td>
<td></td>
<td>pre</td>
<td>rec</td>
</tr>
<tr>
<td>MV</td>
<td></td>
<td>1.000</td>
<td>0.098</td>
<td>0.179</td>
<td>1.000</td>
<td>0.167</td>
</tr>
<tr>
<td>TF</td>
<td></td>
<td>1.000</td>
<td>0.098</td>
<td>0.179</td>
<td>1.000</td>
<td>0.167</td>
</tr>
<tr>
<td>GLAD</td>
<td></td>
<td>1.000</td>
<td>0.098</td>
<td>0.179</td>
<td>1.000</td>
<td>0.167</td>
</tr>
<tr>
<td>LTM</td>
<td></td>
<td>0.960</td>
<td>0.423</td>
<td>0.587</td>
<td>1.000</td>
<td>0.398</td>
</tr>
<tr>
<td>EM</td>
<td></td>
<td>0.956</td>
<td>0.431</td>
<td>0.594</td>
<td>1.000</td>
<td>0.404</td>
</tr>
<tr>
<td>TSE</td>
<td></td>
<td>0.970</td>
<td>0.895</td>
<td>0.931</td>
<td>0.995</td>
<td>0.794</td>
</tr>
</tbody>
</table>

F1 scores for traffic light detection experiments.
Experiments

- Traffic light detection (Area 2)

(a) F1 versus $M$  
(b) F1 versus $N$
Experiments

- Image-based event detection

<table>
<thead>
<tr>
<th></th>
<th>Bike rack $\mathcal{B}_M$</th>
<th>Restaurant $\mathcal{C}_M$</th>
<th>Plant $\mathcal{P}_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$pre$ $rec$ $F1$</td>
<td>$pre$ $rec$ $F1$</td>
<td>$pre$ $rec$ $F1$</td>
</tr>
<tr>
<td>MV</td>
<td>1.000 0.468 0.638</td>
<td>0.836 0.468 0.600</td>
<td>0.908 0.488 0.602</td>
</tr>
<tr>
<td>TF</td>
<td>1.000 0.493 0.661</td>
<td>0.836 0.468 0.600</td>
<td>0.908 0.488 0.602</td>
</tr>
<tr>
<td>GLAD</td>
<td>1.000 0.545 0.706</td>
<td><strong>0.954</strong> 0.463 0.623</td>
<td><strong>1.000</strong> 0.510 0.675</td>
</tr>
<tr>
<td>LTM</td>
<td>0.954 0.630 0.758</td>
<td>0.921 0.569 0.704</td>
<td>0.983 0.550 0.677</td>
</tr>
<tr>
<td>EM</td>
<td>0.955 0.636 0.763</td>
<td>0.917 0.568 0.702</td>
<td>0.973 0.583 0.685</td>
</tr>
<tr>
<td>TSE</td>
<td>0.963 0.874 0.916</td>
<td>0.916 0.708 0.799</td>
<td>0.865 0.671 0.756</td>
</tr>
</tbody>
</table>
Experiments

- Simulation (F1 score on event labels)
Experiments

- Simulation (MAE on TPRs a and FPRs b)
Discussion

- Sequential mobility modeling
- Dependent sources
- Cross-domain truth discovery
Conclusion

• Our proposed model integrates location popularity, location visit indicators, truth of events and three-way participant reliability in a unified framework.

• It can efficiently handling both unknown participants’ reliability and mobility.

• It can efficiently discover true events in mobile crowdsourced event detection without any supervision and location tracking.
Q & A

Thank you!