Detecting spatio-temporal clusters in Reality Mining datasets

Matthew Orlinski
Fraunhofer IAIS

March 31, 2014
Opportunistic Networks (OppNets) are made up of mobile devices communicating directly with each other.
OppNets

- Opportunistic Networks (OppNets) are made up of mobile devices communicating directly with each other.

- They can be used for:
  - Pervasive sensing applications.
Opportunistic Networks (OppNets) are made up of mobile devices communicating directly with each other.

They can be used for:

- Pervasive sensing applications.
- Sharing resources.
OppNets

- Opportunistic Networks (OppNets) are made up of mobile devices communicating directly with each other.

- They can be used for:
  - Pervasive sensing applications.
  - Sharing resources.
  - Distributing locally relevant information.
Challenges for OppNets

- Connections are intermittent, making multi-hop communication between participants difficult.
Challenges for OppNets

- Connections are intermittent, making multi-hop communication between participants difficult.

- We can’t use many of the protocols developed for MANETS because of the high level of device mobility.

- Instead we use Delay or Disruption Tolerant (DTN) protocols.
What can we do?

- Analysing how humans interact may help to improve the performance of OppNets.
What can we do?

- Analysing how humans interact may help to improve the performance of OppNets.
  1. Outline the energy requirements for future Peer Aware Communications (PAC) standards such as IEEE 802.15.8 and Wi-Fi direct.
What can we do?

- Analysing how humans interact may help to improve the performance of OppNets.
  1. Outline the energy requirements for future Peer Aware Communications (PAC) standards such as IEEE 802.15.8 and Wi-Fi direct.
  2. Improve data routing protocols by taking advantage of human mobility to forward packets.
### Table: Comparison of some reality mining datasets

The daily encounter probability is the probability that an encounter with a particular other device will take place on any given day. Granularity is the time between peer discovery scans.

<table>
<thead>
<tr>
<th></th>
<th>Infocom5</th>
<th>Infocom6</th>
<th>Cambridge</th>
<th>Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environment</td>
<td>Conference</td>
<td>Campus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration (days)</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>246</td>
</tr>
<tr>
<td>Number of devices</td>
<td>41</td>
<td>78</td>
<td>36</td>
<td>97</td>
</tr>
<tr>
<td>Device type</td>
<td>iMote</td>
<td>iMote</td>
<td>iMote</td>
<td>Phone</td>
</tr>
<tr>
<td>Number of encounters</td>
<td>22459</td>
<td>128979</td>
<td>10641</td>
<td>102594</td>
</tr>
<tr>
<td>Daily encounter probability</td>
<td>0.78</td>
<td>0.73</td>
<td>0.24</td>
<td>0.01</td>
</tr>
<tr>
<td>Granularity (seconds)</td>
<td>120</td>
<td>120</td>
<td>600</td>
<td>300</td>
</tr>
<tr>
<td>Geographic location</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Cell ID</td>
</tr>
</tbody>
</table>
# Reality mining data

**Table:** Comparison of some reality mining datasets. The daily encounter probability is the probability that an encounter with a particular other device will take place on any given day. Granularity is the time between peer discovery scans.

<table>
<thead>
<tr>
<th></th>
<th>Infocom5</th>
<th>Infocom6</th>
<th>Cambridge</th>
<th>Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environment</td>
<td>Conference</td>
<td>Campus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration (days)</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>246</td>
</tr>
<tr>
<td>Number of devices</td>
<td>41</td>
<td>78</td>
<td>36</td>
<td>97</td>
</tr>
<tr>
<td>Device type</td>
<td>iMote</td>
<td>iMote</td>
<td>iMote</td>
<td>Phone</td>
</tr>
<tr>
<td>Number of encounters</td>
<td>22459</td>
<td>128979</td>
<td>10641</td>
<td>102594</td>
</tr>
<tr>
<td>Daily encounter probability</td>
<td>0.78</td>
<td>0.73</td>
<td>0.24</td>
<td>0.01</td>
</tr>
<tr>
<td>Granularity (seconds)</td>
<td>120</td>
<td>120</td>
<td>600</td>
<td>300</td>
</tr>
<tr>
<td>Geographic location</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Cell ID</td>
</tr>
</tbody>
</table>
Visually analysing the encounters with a graph

Figure: All of the encounters in the Cambridge dataset.
Figure: The encounters present in the $4^{th}$, $5^{th}$, and $6^{th}$ hourly time frames of the Cambridge dataset.

We experimented with discrete time frames of 1, 6, and 12 hours.
Strongly connected subgraphs

- Strongly connected subgraphs are sections of a directed graph where there is a path from every vertex to every other vertex.
Multi-frame spatio-temporal clusters (MEBS)

- MEBS are strongly connected subgraphs which are present across multiple consecutive time frames.
- Stable structures which could form a backbone for the network.

![Diagram](image)

(a) $x^{th}$ hour  
(b) $x + 1^{th}$ hour

**Figure:** A MEBS cluster formed over 2 time frames.
Single-frame Expectation-Based Spatio-temporal clusters (SEBS)

- SEBS are strongly connected subgraphs in a single frame where the value of some metric is higher than expected for every node.
- Used to show when people come together for a short amount of time.
- The expected values are called the baselines.

Figure: SEBS example in the $x^{th}$ hour
The baseline for a single device $i$ during the current time frame $t_{(n)}$. 

$$b_{i}^{t_{n}} = \frac{m_{i}^{t_{(n-1)}} + m_{i}^{t_{(n-2)}} + \cdots + m_{i}^{t_{(n-w)}}}{w}$$
Baselines for devices

\[ b_{i}^{tn} = \frac{m_{i}^{t(n-1)} + m_{i}^{t(n-2)} + \cdots + m_{i}^{t(n-w)}}{w} \]

- Baselines are calculated using the average cumulative connection time with other devices in the previous \( w \) hours.
- We looked at the cumulative connection time because single connections between devices are unstable (the Parking Lot Problem).
- Only storing the average cumulative connection time for each time frame also cuts down on the amount of storage and processing needed for the algorithm.
Baselines for devices

\[ b_i^{tn} = \frac{m_{i}^{t(n-1)} + m_{i}^{t(n-2)} + \cdots + m_{i}^{t(n-w)}}{w} \]

- We experimented with \( w \) values of 2, 4, 12, 18, and 24.
A short cluster analysis...
Size of SEBS clusters

- Mean size is around 5-6 devices.
- With frame sizes of 12 hours there are large clusters detected at the start of Infocom6, and there are no clusters detected at all in Infocom5.
- This is due to the **premature clustering problem** which is caused by high metric values and mixing of participants at the start of the experiment, with ever decreasing participation thereafter.

![Figure: Mean size of SEBS clusters](image)
Size of MEBS clusters

- MEBS clusters at conferences tend to be smaller than in campus experiments.

Figure: Size of MEBS clusters spanning 2 time frames when using 1 hour long frames.
Duration of MEBS clusters

- MEBS clusters which last longer than 2 hours are rare in each of the datasets.
- MEBS clusters at conferences tend to last longer than in campus experiments.

![Bar chart showing MEBS clusters which last for more than 2 hours across different datasets.](chart)

- Reality
- Cambridge
- Infocom5
- Infocom6

MEBS clusters which last for more than 2 hours
Short analyses summary

- We’ve outlined a simple method to detect spatio-temporal clusters.
- Can SEBS clusters meet our requirements? (next slide)
Cluster based data delivery

$V'_k$'s cluster
End-to-end data delivery based on SEBS clusters

- Low data delivery probability because SEBS clusters tend to be too small.
End-to-end data delivery based on SEBS clusters

- SEBS clusters are more efficient than Bubble’s because SEBS clusters are \textbf{non-monotonic}.

![Graph showing efficiency vs. time for different protocols](image)

- **Bubble-kclique**
- **Bubble-simple**
- **PRoPHETv2**
- **Nomads**
- **SEBS**

The graph plots efficiency (packets delivered / relayed) against time (days) for various protocols, illustrating the performance differences over time.
Data delivery overall

<table>
<thead>
<tr>
<th></th>
<th>Data delivery probability (TTL 1 hour)</th>
<th>Overheads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bubble</td>
<td>11.4%</td>
<td>20.697</td>
</tr>
<tr>
<td>PRoPHETv2</td>
<td>14.5%</td>
<td>25.683</td>
</tr>
<tr>
<td>DEBTT [1]</td>
<td>12.1%</td>
<td>19.146</td>
</tr>
<tr>
<td>DRAFT [2]</td>
<td>14.7%</td>
<td>25.348</td>
</tr>
</tbody>
</table>


Why is data delivery so low?

(a) New encounters per hour  
(b) Hourly total encounter durations
Why are overheads so high?

► All of the protocols use message duplication to increase data delivery chance.

► Some messages are duplicated up to 50 times because we can’t predict when the next “useful” encounter will take place.
Future work

- Working toward human encounter forecasting to predict when encounters will occur.
- We need bigger human encounter datasets that better reflect daily life.
Thank you for listening

Matthew Orlinski
me@shigs.co.uk