RECAST: Telling Apart Social and Random Relationships in Dynamic Networks

Pedro Olmo Vaz de Melo¹, Aline C. Viana², Marco Fiore³, Katia Jaffrès-Runser⁴, Frédéric le Mouël⁵, Antonio A.F. Loureiro¹

¹ Universidade Federal de Minas Gerais, Brazil
² Inria, France
³ IEIIT - CNR, Italy
⁴ University of Toulouse, IRIT / ENSEEIHT, Toulouse, France
⁵ INSA Lyon, France

13ème Forum des Jeunes Mathématicien-ne-s 2013,
November 13-15, 2013
Smartphones have the potential to be:
- visually-aware
- sonically-aware
- always-connected
- directionally-aware
- location-aware
- motion-aware

User-aided wireless networks/Disruption-Tolerant networks

Creates a **dynamic network** with high potential for wireless and pervasive applications

- Wireless social networking, global sensing, content distribution,
- More and more data in transit: 3/4G network offloading...
Internetworking human beings!

Real-world mobility scenarios create neither purely regular nor purely random connections among the entities composing the network.

Dynamic Complex Wireless Networks (DCWN)

- Have large number of vertices and edges that exhibit a pattern
- Evolves according to semi-rational decisions of its entities
- Semi-rational decisions:
  - are mostly regular and repeat themselves
  - but may be influenced by random events
1. Rationale and related initiatives

2. Random graphs vs. social graphs

3. RECAST: Random rElationship ClAssifier STrategy

4. Classification results

5. Conclusion
1. Rationale and related initiatives

2. Random graphs vs. social graphs

3. RECAST: Random rElationship ClAssifier STrategy

4. Classification results

5. Conclusion
Data sets

Data collection to build *contact traces*

- Log the contact time and duration of a node to an access point
- Log the GPS coordinates of mobile nodes regularly

Derive a time-varying contact graph

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Local</th>
<th># entities</th>
<th>Duration</th>
<th>Type</th>
<th>Avg. # encounters/node/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dartmouth</td>
<td>campus</td>
<td>1156</td>
<td>2 months</td>
<td>Individuals</td>
<td>145.6</td>
</tr>
<tr>
<td>USC</td>
<td>campus</td>
<td>4558</td>
<td>2 months</td>
<td>Individuals</td>
<td>23.8</td>
</tr>
<tr>
<td>San Francisco</td>
<td>City</td>
<td>551</td>
<td>1 month</td>
<td>Cabs</td>
<td>834.7</td>
</tr>
</tbody>
</table>

- Dartmouth and USC collect connection dates/durations to WiFi APs,
- San Francisco collects GPS locations of taxi cabs.

---

3. A. Rojas et al. “Experimental validation of the random waypoint mobility model through a real world mobility trace for large geographical areas,” in Proc. of the 8th ACM MSWiM 2005
Rationale and related initiatives

Characterize interactions, i.e. edges of contact graph

- Regularity of contacts: How often did Arnaud and Paul meet per day? during the whole trace?

Miklas et al.\(^4\) determine whether 2 nodes are friends or strangers using an empirical threshold (friends encounter 10 times or more within 14 weeks).

---

\(^4\) A. G. Miklas et al., “Exploiting social interactions in mobile systems,” in *Proceedings of the UbiComp ’07*.  

**KJR**  
Forum des Jeunes Mathématicien-ne-s 2013
Characterize node’s behavior, i.e. vertices of contact graph

Using localization information, Zyba et al.\(^5\) differentiate *social* from *vagabond* nodes. Socials appear regularly in a given area while vagabonds visit an area rarely and unpredictably.

\[^5\text{G. Zyba, G. Voelker, S. Ioannidis, and C. Diot, “Dissemination in opportunistic mobile ad-hoc networks: The power of the crowd, in Infocom’11}\]

- Monitor the total appearance and regularity of appearance

Paul is social at the cafeteria but vagabond at the library: a per node/per area approach $\rightarrow$ *geographical dependency*
Rationale and related initiatives

**RECAST**

- Characterizes the *interactions of nodes* based on their probability to originate from a random or social behaviour
- Identify different kinds of social interactions (friends, acquaintances, bridges or random)
- No geographical dependency, i.e., is of general validity
Outline

1. Rationale and related initiatives

2. Random graphs vs. social graphs

3. RECAST: Random rElationship ClAssifier STStrategy

4. Classification results

5. Conclusion
Temporal social graphs from contact traces

Two possible representations

1. $\delta$ event graph: $G_k(V_k, E_k)$
   There is an edge in $E_k$ if contact within $\delta = 1$ day for instance.

   \begin{align*}
   \text{Day 1 event graph } G_1(V_1, E_1) \\
   \text{Day 2 event graph } G_2(V_2, E_2)
   \end{align*}

2. Accumulative graph $G_t(V_t, E_t)$
Two possible representations

1. $\delta$ event graph: $G_k(V_k, E_k)$
   There is an edge in $E_k$ if contact within $\delta = 1$ day for instance.

2. Accumulative graph $G_t(V_t, E_t)$: $G_t = \{G_1 \cup G_2 \cup \ldots \cup G_t\}$

$G_2(V_2, E_2)$ Accumulative graph up to Day 2

Accumulates all event graphs up to time step $t$. 
Temporal graphs generation from contact traces

Example accumulative graph $G_t$ for $t = 2$ weeks
For $\delta = 1$ day and using force-direct layout algorithm for plotting

Seems difficult to extract any knowledge from these social graphs:
→ gathers all social AND random interaction!

(a) Dartmouth  
(b) USC     
(c) San Francisco

KJR  
Forum des Jeunes Mathématicien-ne-s 2013
Comparing a social graph to its random counterpart

Random graph equivalent of $G$

Calculate a random graph $G^R$ from a graph $G(V,E)$:

- Keep same number of vertices and edges,
- Randomly assign edges to keep the same node degree distribution using $RND$ algorithm:

A edge is set between nodes of degree $d_i$ and $d_j$ with probability

$$p_{ij} = \frac{d_i \times d_j}{\sum_{k=1}^{\mid V \mid} d_k}$$

Random accumulative graph $G_t^R$

Random accumulative graph derived from event graphs $\{G_i\}_{i \in [1,..,t]}$

$$G_t^R = \{ RND(G_1) \cup RND(G_2) \cup \ldots \cup RND(G_t) \}$$

---

Comparison social vs. random graphs

Network clustering coefficient can identify a network with an elevated number of clusters (i.e. communities).

- If $\bar{c}(G) \gg \bar{c}(G^R)$, parts of the decisions of the nodes of $G$ are NOT random

- Dartmouth / USC traces have an order of magnitude higher $\bar{c}$ than $G^R \rightarrow$ social decisions

- San Francisco: each individual taxi in the trace encounters most of the other taxis $\rightarrow$ closer to a random behavior
1. Rationale and related initiatives

2. Random graphs vs. social graphs

3. RECAST: Random rElationship ClAssifier STTrategy

4. Classification results

5. Conclusion
Social network features: Regularity and Similarity

Social nodes’ behavior tend to

- repeat on a regular basis (because of daily activities for instance) → Regularity
- build persistent communities and generate common acquaintances → Similarity

Mathematical metrics

- **Edge persistence** $per(i, j)$ $^7$: Percentage of time steps an edge exists over the past discrete time steps in the event graphs $\{G_i\}_{i \in [1, \ldots, t]}$
- **Topological overlap** $to(i, j)$ $^8$: Ratio of neighbors shared by two nodes calculated for the accumulative graph $G_t$.

---


CCDF of edge persistence $\text{per}(i,j)$ after 4 weeks

Individuals tend to see each other regularly

Encounters occur almost in a random fashion
CCDF of topological overlap $to(i,j)$ after 4 weeks

Individuals of $G_t$ have common neighbors

![Graph showing CCDF for reality and random](image)

Common neighbors occur in a random fashion

![Graph showing CCDF for reality and random](image)

(g) Dartmouth

(h) USC

(i) San Francisco
Social vs. random edges

In the random network, we only have a probability of $10^{-3}$ to have edges with a persistence of more than $\bar{x}_{per} = 0.17$.

\[ P[\text{per}(i, j) > \bar{x}_{per}] \]

$\rightarrow$ Thus, in the social graph $G_t$:

- edges with $\text{per}(i, j) > \bar{x}_{per}$ can be classified as social edges
- edges with $\text{per}(i, j) < \bar{x}_{per}$ can be classified as random edges

Note that there is a $p_{rnd}$ chance that a social edge is actually random (mis-classification)
RECAST classification algorithm

Only parameter of RECAST: $p_{\text{rnd}}$, the mis-classification error bound.

Main steps

- Calculate the $\text{per}(i,j)$ and $\text{to}(i,j)$ for each edge
- Knowing $p_{\text{rnd}}$, calculate $\bar{x}_{\text{per}}$ and $\bar{x}_{\text{to}}$ from CCDF’s
- For each edge,
  - if $\text{per}(i,j) > \bar{x}_{\text{per}} \rightarrow (i,j)$ is social for edge persistence
    else $(i,j)$ is random for edge persistence
  - if $\text{to}(i,j) > \bar{x}_{\text{to}} \rightarrow (i,j)$ is social for topological overlap
    else $(i,j)$ is random for topological overlap
- Classify edges into classes of relationships according to:

<table>
<thead>
<tr>
<th>Class</th>
<th>Edge persistence</th>
<th>Topological overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friends</td>
<td>social</td>
<td>social</td>
</tr>
<tr>
<td>Acquaintances</td>
<td>random</td>
<td>social</td>
</tr>
<tr>
<td>Bridges</td>
<td>social</td>
<td>random</td>
</tr>
<tr>
<td>Random</td>
<td>random</td>
<td>random</td>
</tr>
</tbody>
</table>
Outline

1. Rationale and related initiatives

2. Random graphs vs. social graphs

3. RECAST: Random rElationship ClAssifier STTrategy

4. Classification results

5. Conclusion
Snapshots after 2 weeks

Dartmouth

Social edges

Random edges

USC

Social edges

Random edges

Friends edges are in blue, Bridges edges are in red
Acquaintance edges are in gray, Random edges are in orange
Cluster coefficient analysis for only random edges

Validates the efficiency of RECAST to identify random edges for Dartmouth and USC
Classification results after 4 weeks

Number of edges of a each class that appear in the first 4 weeks vs. $p_{\text{rnd}}$

RECAST is not sensitive to $p_{\text{rnd}}$!
Outline

1. Rationale and related initiatives

2. Random graphs vs. social graphs

3. RECAST: Random rElationship CIAssifier STrategy

4. Classification results

5. Conclusion
Conclusions and future works

RECAST

▶ requires a unique parameter
▶ combines user encounter frequency with their 2-hop social network ties
▶ identifies different kinds of social interactions: friendship, acquaintanceship and bridges

Different mobility traces may have completely different behaviors (San Francisco vs. USC)

Future works

▶ Provide a distributed RECAST classification
▶ Assess RECAST using data sets with ground truth
▶ Study spatio-temporal correlations of data sets
Thank you for your attention

First beta-version of RECAST classifier available on:
Epidemic data dissemination scenario

Studied scenario

- Edges are classified with RECAST after 4 weeks
- A dissemination stage is simulated for the following 2 weeks:
  - For each user $i$ and destination $j$, pick a random time in dissemination week 1 to start the epidemic transfer
  - If destination not reached after 1 week, transfer has failed.
Epidemic data dissemination results

- Each bar represents one of the relationships the source and the destination can share
- Within each bar, we depict the % of edges of a given class that have participated to the dissemination

Dartmouth

USC

- USC is a bigger network: harder to reach its destination
- Dartmouth rarely uses Random edges
- Although the majority of the edges in USC are Random, most of the hops are made by users who share a social relationship
Epidemic data dissemination: Path length

- Histogram of the path lengths of messages between users $i$ and $j$ that share the same class.

The expected time to reach a Random contact is significantly higher than the time needed to reach a social contact.

For both data sets, a majority of routes to social edges have a path length $\leq 3$, while only a few percents of routes to random edges do.