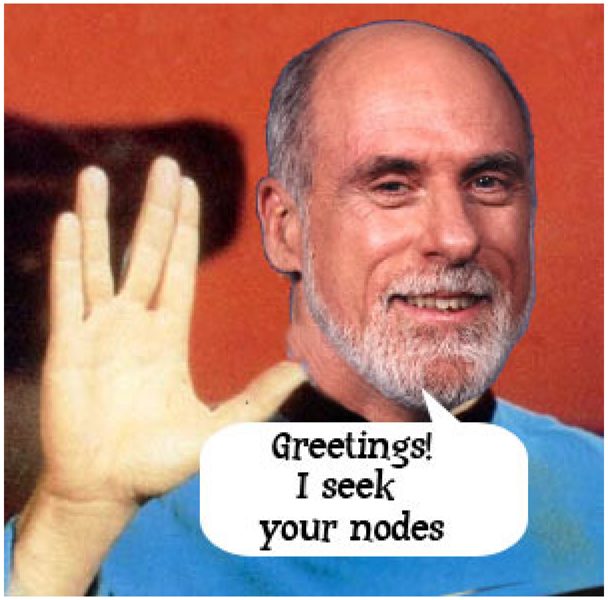


Telling apart social and random relationships in wireless networks

Aline Carneiro Viana
Hipercom research team

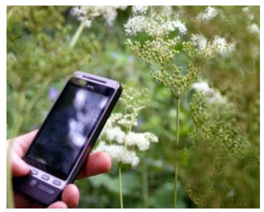
June 2012

P.O.S.V. de Melo, A.A.F. Loureiro from Federal Univ. of Minas Gerais
M. Fiore, K. Jaffrès-Runser, F. Le Mouel from INSA Lyon

The smartphone phenomena
and the ***culture of the small screen****…*

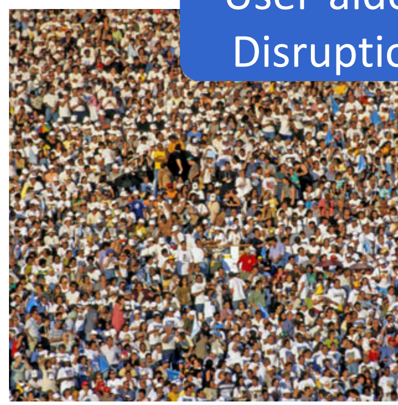
Smartphones have the potential to be:

visually-aware
sonically-aware
always-connected
directionally-aware
location-aware
motion-aware

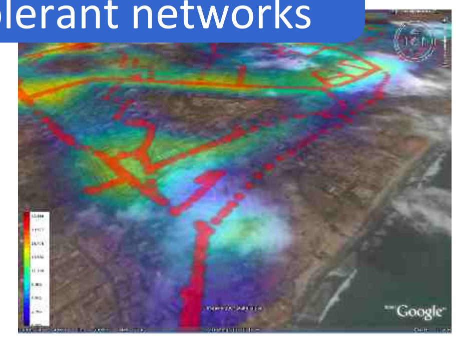


+

User-aided wireless networks/ Disruption-Tolerant networks



=

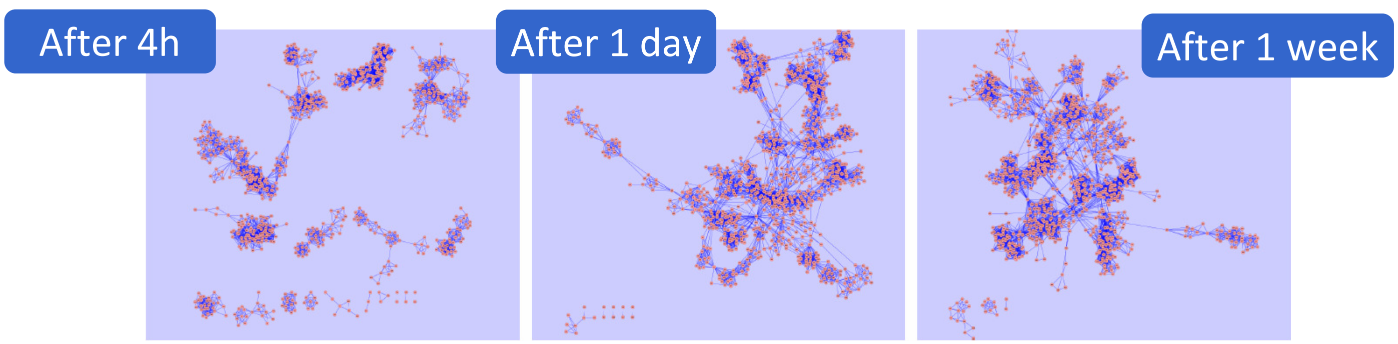


- New potential wireless and pervasive applications
  - Wireless Social networks, global sensing, content distribution
  - Increasing volume of mobile data between 2010-2015

…involving. devices carried by humans

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

- Decision-Based Wireless Networks (DbWN)
  - Have large number of vertices and edges that exhibit a pattern
  - Communities are naturally formed, reflecting social decisions of entities



  - Evolves according to semi-rational decisions of entities ≠ random networks
    - Semi-rational decisions tend to be regular and to repeat themselves

Random events are always possible in humans routines

- But…
  - …introduce significant amount of noise in predictable patterns
  - …make the process of knowledge discovery in datasets a complex task

- Proposal: Random relationship classifier strategy (RECAST)
  - Accurately identify random from social interactions (nodes wireless encounters) in large datasets

- Application scenarios:
  - Recommendation systems
  - Forwarding strategies
  - Ad-hoc message dissemination schemes (high coverage and limited number of messages)

Outline

1. Considered real-world datasets
2. Comparison with random graphs
   – Temporal graph generation
3. Random relationship classifier strategy (RECAST)
   – Identified features
   – Algorithm
4. Classification results
5. Case of study
   – Data dissemination
6. Conclusion

Considered real-world datasets

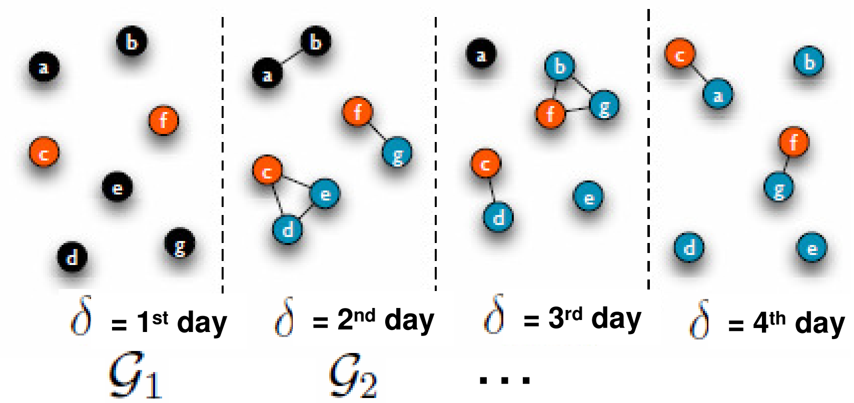
Dataset	Local	Number of entities	Duration	Entities type	Avg. # encounters/node/day
Dartmouth [30]	University campus	1156	2 months	Devices	145.6
USC [31]	University campus	4558	2 months	Devices	23.8
San Francisco [32]	City	551	1 month	Cabs	834.7

- [30] T. Henderson et al. "The changing usage of a mature campus-wide wireless network," in *Proc. of ACM* MobiCom 2004.
- [31] W. jen Hsu et al. "Impact: Investigation of mobile-user patterns across university campuses using wlan trace analysis," *CoRR*, vol. abs/cs/0508009, 2005.
- [32] 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.

Comparison with Random Graphs

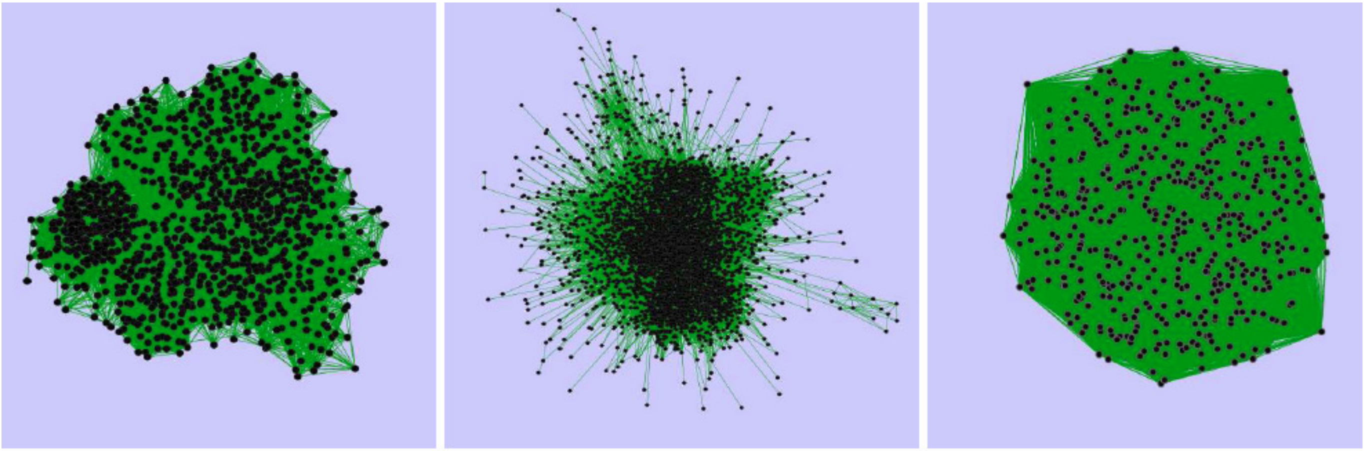
Temporal graph generation

- Time steps δ = 1 day
- Event graph: $\mathcal{G}_k(\mathcal{V}_k, \mathcal{E}_k)$
- Time accumulative graph:
  - $G_t = (V_t, E_t)$
  - $G_t = \{\mathcal{G}_1 \cup \mathcal{G}_2 \cup \ldots \cup \mathcal{G}_t\}$



- G_t for δ = 1 day and t = 2 weeks

Difficult to extract any knowledge!!



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

- **1st step:** from $\mathcal{G}_k(\mathcal{V}_k, \mathcal{E}_k)$ generates its random version $G^R(V, E^R) = \texttt{RND}\ (G)$ [1]
  - with the same number of nodes, edges, and empirical degree distribution
  - assigns edges with probability

$$p_{i,j} = (d_i \times d_j) / \sum_{k=1}^{|V|} d_k$$

  - the only difference is in the connections among nodes
    - G : nodes connect in a "semi-rational" way
    - G^R: the connections happen in a purely random fashion

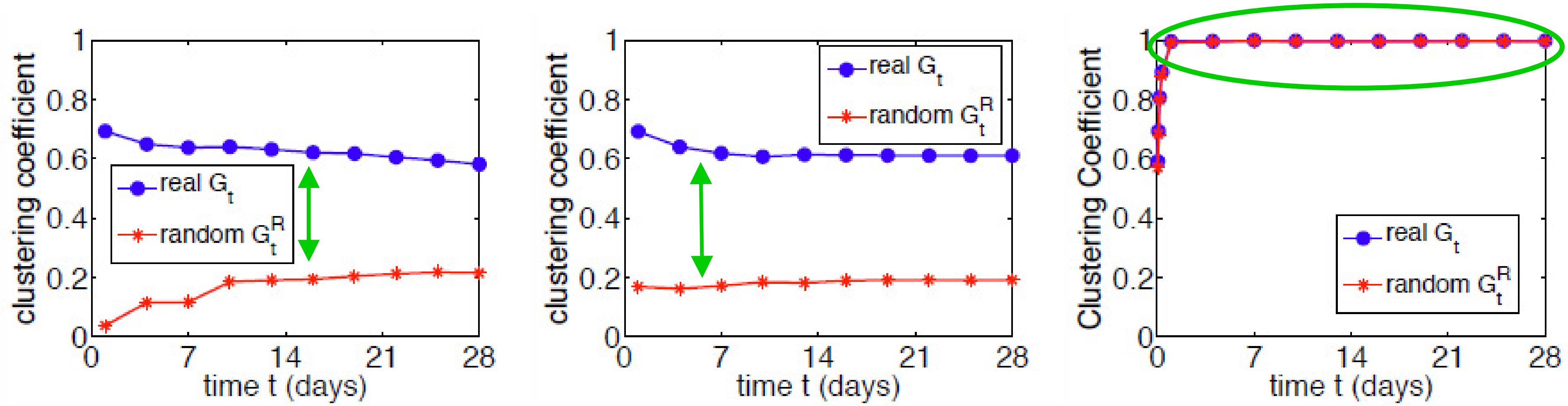
- **2nd step:** generates the temporal random version of G_t : G_t^R
  - T-RND algorithm
  - $G_t^R = \texttt{T-RND}\ (\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_t) = \{\texttt{RND}(\mathcal{G}_1) \cup \texttt{RND}(\mathcal{G}_2) \cup \ldots \cup \texttt{RND}(\mathcal{G}_t)\}$

[1] F. Chung and L. Lu, "Connected Components in Random Graphs with Given Expected Degree Sequences," *Annals of Combinatorics.* Nov. 2002.

Comparison with Random Graphs (2)

- Clustering coefficient (cc): probability of two neighbors of a node to be directly connected
  - good metric to differentiate social networks from random networks
  - when cc $_G$ >> cc $_{G^R}$ $\Rightarrow$ (part of) the decisions made by the agent of G are non-random

Each individual taxi encounters most of the other taxis $\Rightarrow$ similar to a random network



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

RECAST classifier

Social relationships

- Two main features:
  - Regularity [2]
    - Encounters between "friends" repeat often
  - Similarity [3]
    - two "friends" share common "friends"

- How to represent them mathematically?
  - Edge persistence
  - Topological overlap

[2] N. Eagle, A. Pentland, and D. Lazer, "From the Cover: Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences*, Sept. 2009.
[3] J. P. Onnela, J. Saram¨aki, J. Hyvoonen, G. Szab´o, D. Lazer, K. Kaski, J. Kert´esz, and A. L. Barab´asi, "Structure and tie strengths in mobile communication networks," *Proc. of the National Academy of Sciences*, May 2007.

Edge persistence

- Percentage of times an edge occurred over the past discrete time steps 1,2, …, t
- Applied at the event graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_t\}$

day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
encounter between Smith and Johnson	x		x		x		

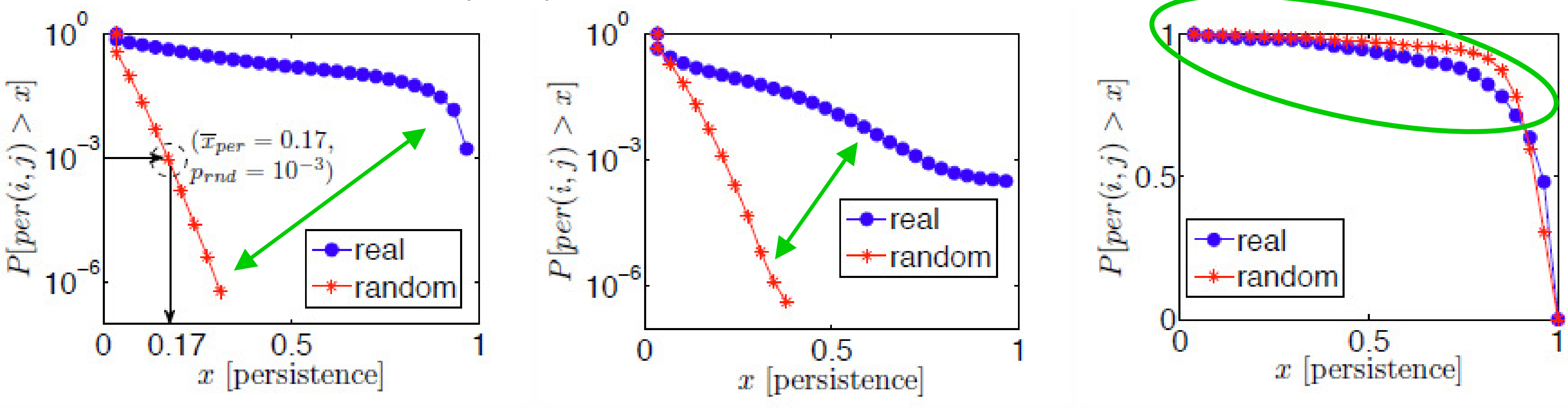
Edge Persistence: 3/7

Edge persistence

- Complementary cumulative distribution function of *per (i, j)*
- 4 weeks of contacts of each dataset

Individuals tend to see each other regularly, for reasons beyond pure randomness

Encounters occur almost in a random fashion

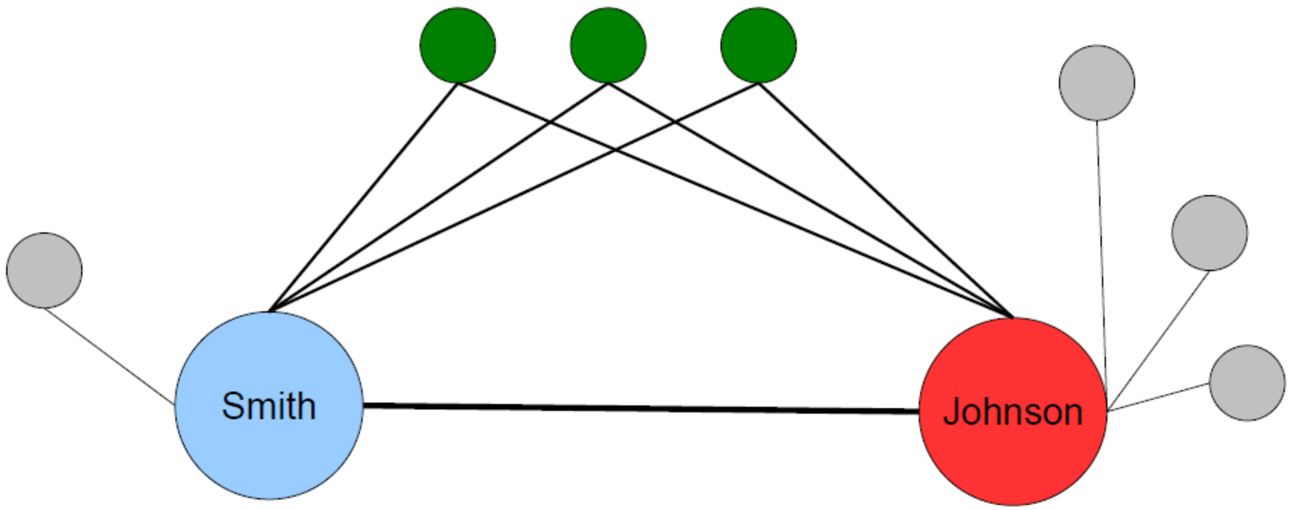


(a) Dartmouth (b) USC (c) San Francisco

Feature values > x are very unlikely to occur in a random network ⇒ are most probably due to actual social relationships

Topological overlap

- Ratio of neighbors shared by two nodes
- Extracted from the aggregated temporal graph $G_t = \{\mathcal{G}_1 \cup \mathcal{G}_2 \cup \ldots \cup \mathcal{G}_t\}$



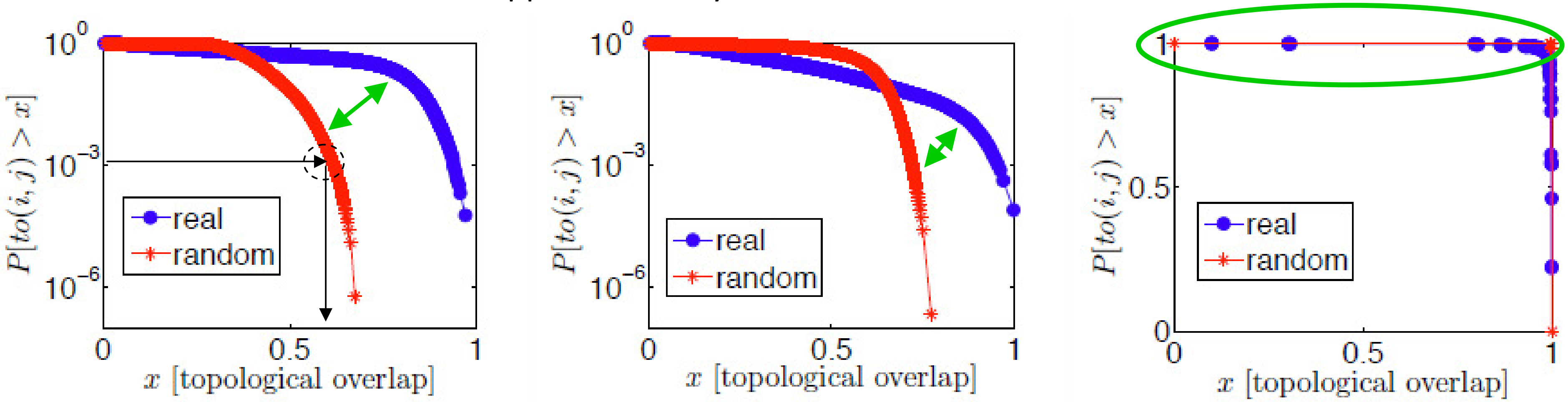
Topological Overlap = 3 / [(5-1) + (7-1) – 3] = 3/7

Topological overlap

- Complementary cumulative distribution function of *to (i, j)*
- 4 weeks of contacts of each dataset

Individuals share common neighbors in a way that could not happen randomly

Common neighbors occur in a random fashion



(d) Dartmouth (e) USC (f) San Francisco

Feature values > x are very unlikely to occur in a random network ⇒ are most probably due to actual social relationships

RECAST algorithm

- For each edge (i,j)
  - compute *per(i,j)* using the event graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_t\}$
  - compute *to(i,j)* using the aggregated temporal graph $G_t = \{\mathcal{G}_1 \cup \mathcal{G}_2 \cup \ldots \cup \mathcal{G}_t\}$

- Compare these values with the ones from the random graph
  - prnd can be seen as the expected classification error percentage

$$\text{Get } \bar{x}_{to} \mid \overline{F}_{to}(\bar{x}_{to}) = p_{rnd} \text{ and } \bar{x}_{per} \mid \overline{F}_{per}(\bar{x}_{per}) = p_{rnd}$$

- Classify edges into classes of relationships

3 types of social relationships

Class	Edge persistence	Topological overlap
Friendship	social	social
Acquaintanceship	random	social
Bridges	social	random
Random	random	random

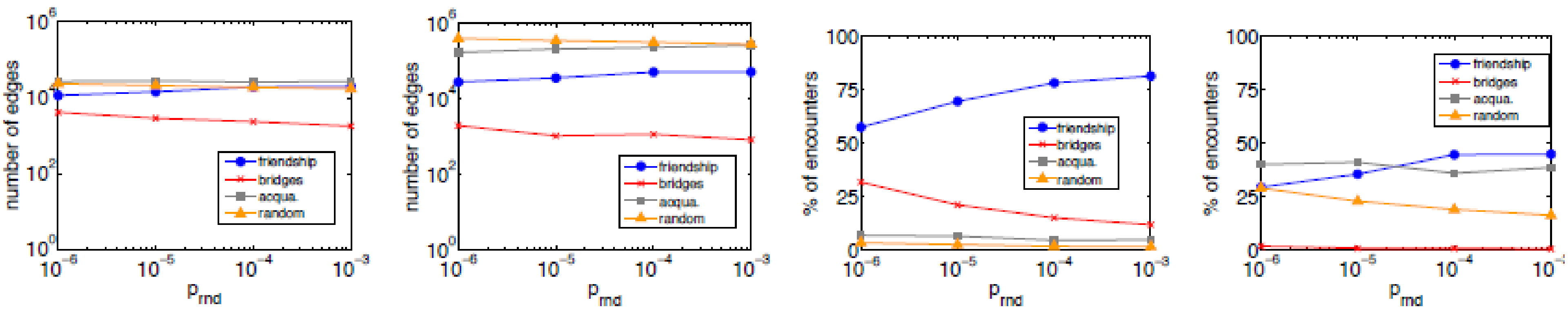
Classification results

Number/Percentage of edges per class *vs* p_{rnd} value

- 4 weeks of contacts of each dataset

Similar dynamics in tight relationships among individuals in the two campuses

USC has a significantly higher tendency to evolve to a random topology than the Dartmouth

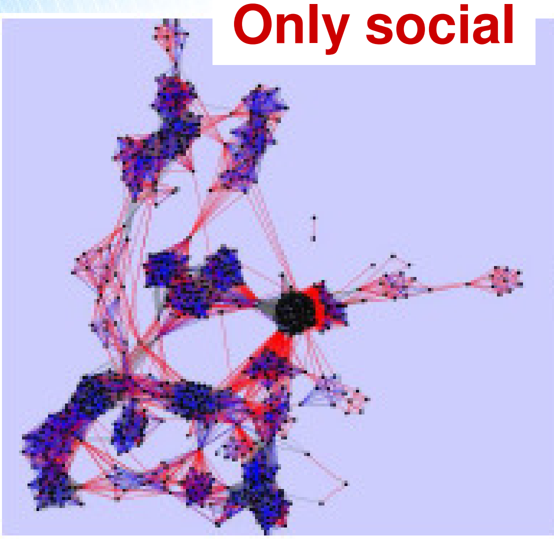


(a) Dartmouth (b) USC (c) Dartmouth (d) USC

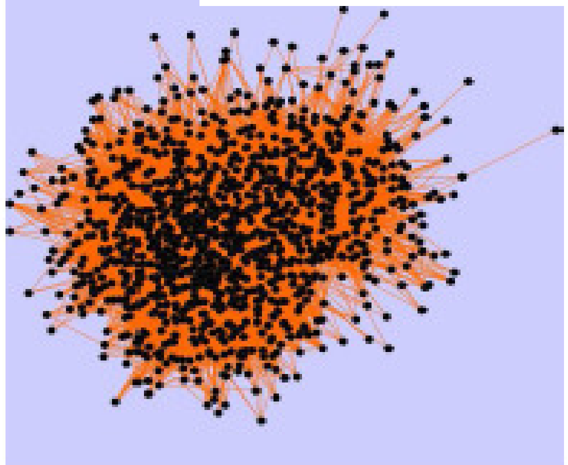
RECAST does not need a fine calibration of p_{rnd} to return a consistent edge classification

Snapshots after two weeks of interactions

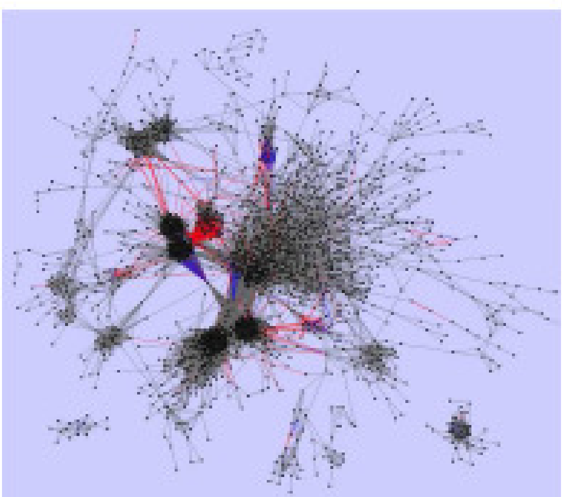
Only social **Only random**



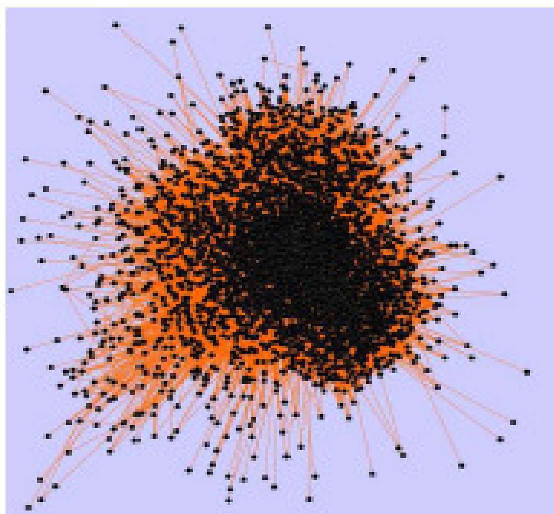
(a) Dartmouth, only social edges



(b) Dartmouth, only random edges



(c) USC, only social edges



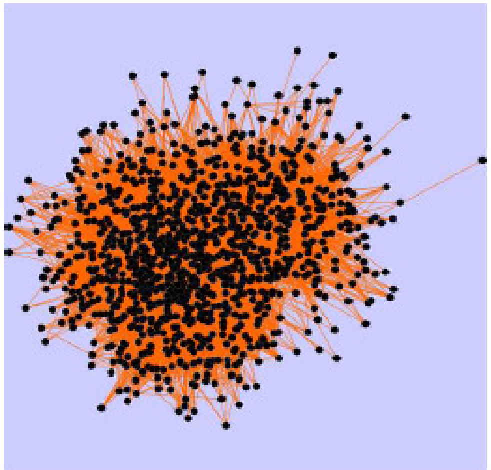
(d) USC, only random edges

- **Social-edges net.**: Complex structure of *Friendship* communities, linked to each other by *Bridges* and *Acquaintanceship*

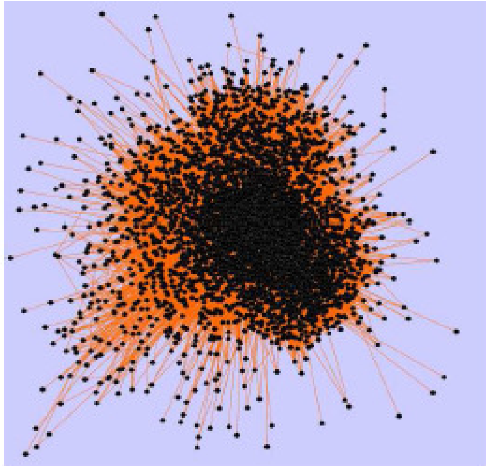
- **Random-edges net.**: No structure appears, looking like random graphs

Friendship edges are in **blue**
Bridges edges are in **red**
Acquaintance edges are in **gray**
Random edges are in **orange**

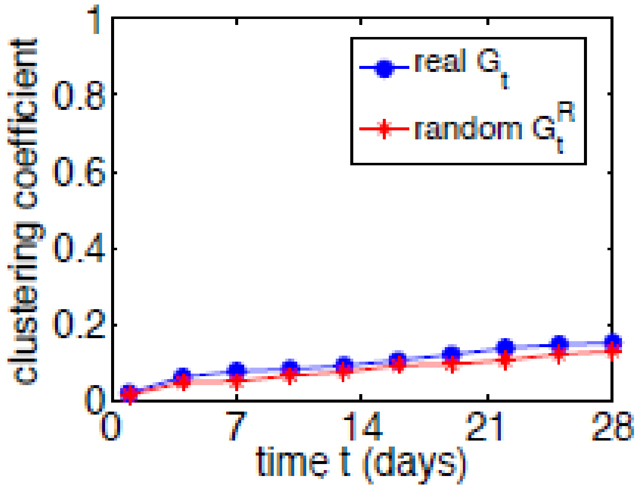
Cluster coefficient analysis, **only** Random edges



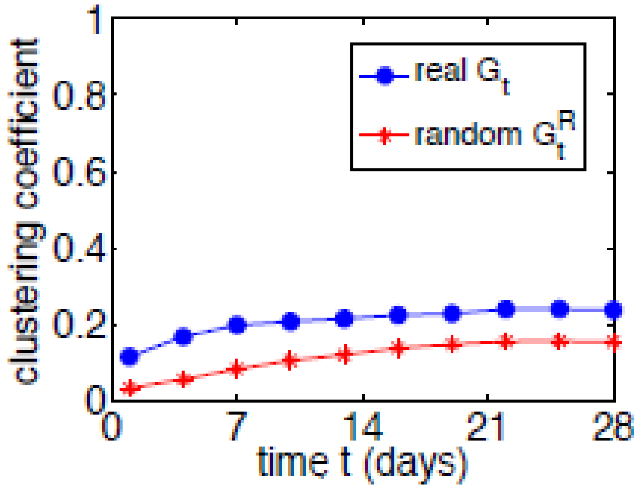
(b) Dartmouth, only random edges



(d) USC, only random edges



(a) Dartmouth



(b) USC

Validates the efficiency of RECAST classification!

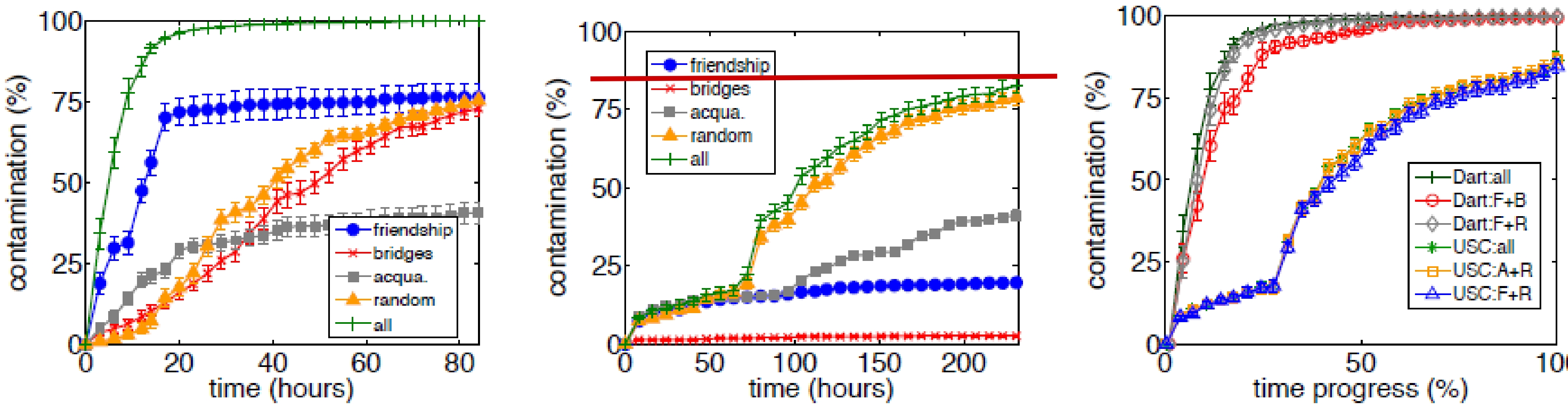
Case of Study

Data dissemination: when only edges of each class are used

Dartmouth dataset:

Training set of 4 weeks

Test set at the 5th week



(a) Dartmouth (b) USC (c) Mix of Classes

USC dataset:

Training set of 6 weeks

Test set at the 7th, 8th, 9th weeks

Data dissemination: results summary

- Efficient contamination needs:
  - edges that provide a high number of encounters inside communities (*Friendship* in Dartmouth and *Acquaintanceship in* USC)
  - edges that provide a high number of connections among individuals in different communities (*Random* and *Bridges* in Dartmouth and *Random in* USC)

- Contamination when *Bridge + Friendship* edges in the Dartmouth ≅ *Random + Friendship*
  - Number of *Bridge* edges ≅ 12% the number of *Random* edges
  - Using *Bridge* edges help to save computational resources

Related initiatives

- Users classification into social and vagabonds [Zyba et al., Infocom 2011]
  - regularity of appearance and duration of visits in a given area
  - only works on a per-individual per-area basis

- Links classification into friends and strangers [Miklas et al., UbiComp 2007]
  - pairs of users meeting 10 days or more out of 101 days are friends
  - otherwise are strangers

- A. G. Miklas et al., "Exploiting social interactions in mobile systems," in *Proc. of the UbiComp '07*.
- G. Zyba et al. "Dissemination in opportunistic mobile ad-hoc networks: The power of the crowd," in *Proc. of IEEE INFOCOM 2011*.

Summary and outlook

- RECAST
  - has no geographical dependency
  - combines user encounter frequency with their 2-hop social network ties
    - periodic behaviors can explain **50% to 70%s** of the human movement patterns
    - but a non-negligible percentage of mobility (**about 10% to 30%**) is due to social relationships
  - identifies different kinds of social interactions
    - friendship, acquaintanceship and bridges

- Different mobility traces may have completely different behaviors

- Researchers should not generalize their results based on the analysis of a single trace

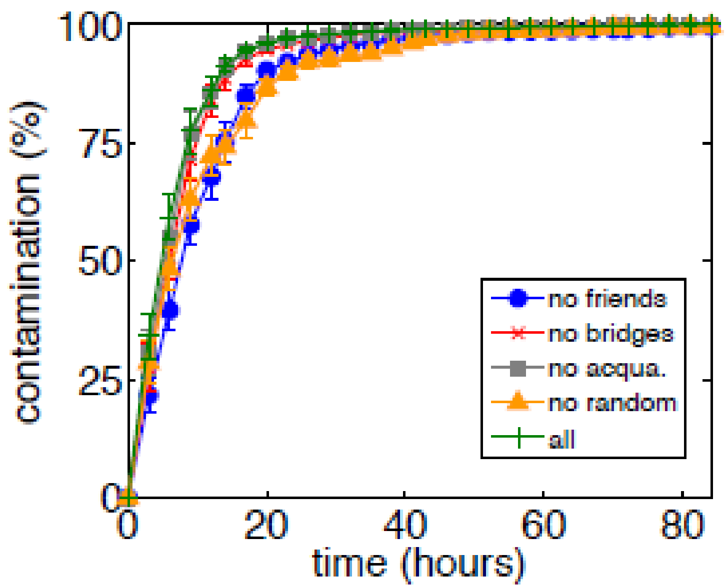
Thanks for your attention!

Questions?

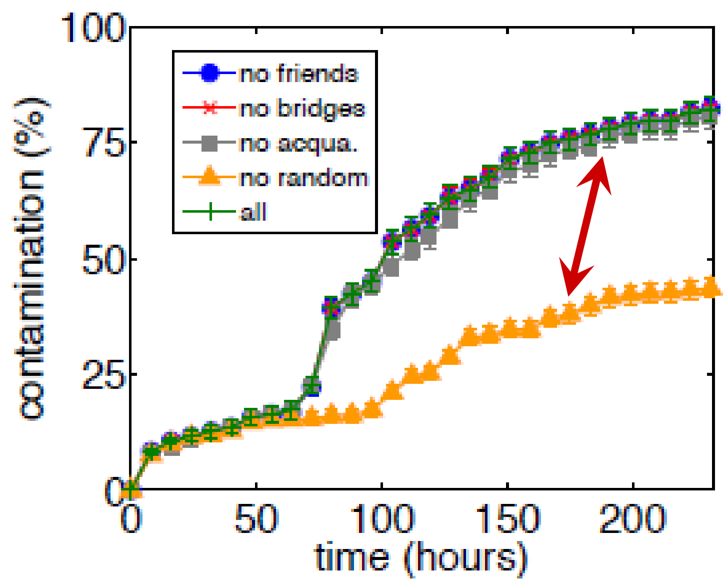
Data dissemination: when only edges of each class are used

Dartmouth dataset:
Training set of 4 weeks
Test set at the 5th week

USC dataset:
Training set of 6 weeks
Test set at the 7th, 8th, 9th weeks



(a) Dartmouth



(b) USC

Link prediction (training set = 4 weeks/test set = 5th week)