Network analysis for **context** and **content** oriented wireless networking

Katia Jaffrès-Runser
University of Toulouse, INPT-ENSEEIHT,
IRIT lab, IRT Team

Ecole des sciences avancées de Luchon
Network analysis and applications
July 3rd, 2014

The smartphone phenomenon

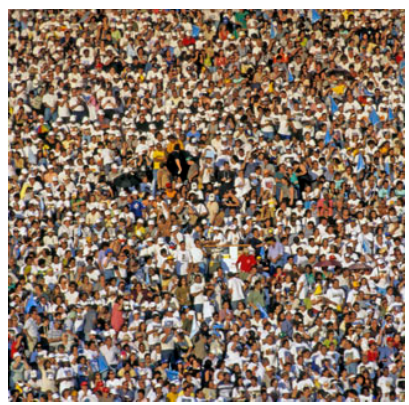
- Multiple sensing and communication capabilities
  - Sensors, camera, GPS, microphone
  - 3G, WiFi, Bluetooth, etc.
  - Storage capabilities (several Gbytes)
  - Computing power

Mobile Traffic is growing constantly

- Increasing volume of mobile data between 2014-2018
  - "…worldwide mobile data traffic will increase nearly **11-fold** over the next four years and reach **an annual run rate of 190 exabytes** (10^{18}) by 2018…"
  - 54% of mobile connections will be 'smart' connections by 2018

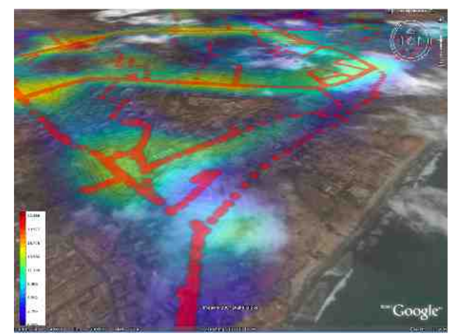
[Cisco VNI Global Mobile Data Traffic Forecast (2013-2018)]

+



In 2013, 4.1 billion users worldwide

=



Next Big Networking Challenge: meet traffic demand !

1. If data is not delay sensitive:
   – e.g. Videos, Application / system updates, music, podcasts, etc.

Leverage opportunistic encounters to route
or flood **delay tolerant** data hop by hop

Benefit: Reduce downloads from infrastructure wireless network

2. If several connectivity options exist:
   – e.g. 3G/4G, WiFi, Femto cells

Offload / Pre-fetch data using
the ‘best‘ available connectivity, at the best time and location

Benefit: Load balancing between available infrastructures

Smartphones are carried by humans

Opportunistic wireless networks

a.k.a. Pocket Switched Networks

1) Large scale and highly dynamic
2) Connections between the network entities are neither purely regular nor purely random
3) Evolve according to semi-rational decisions of entities ≠ random networks
   - Semi-rational decisions tend to be regular and to repeat themselves
   - Random decisions deviate from the regular pattern and are unlikely to repeat

Leverage **social interactions** to improve opportunistic networking, pre-fetching and offloading solutions

Outline

1. **Measure and classify** social interactions
   - RECAST algorithm

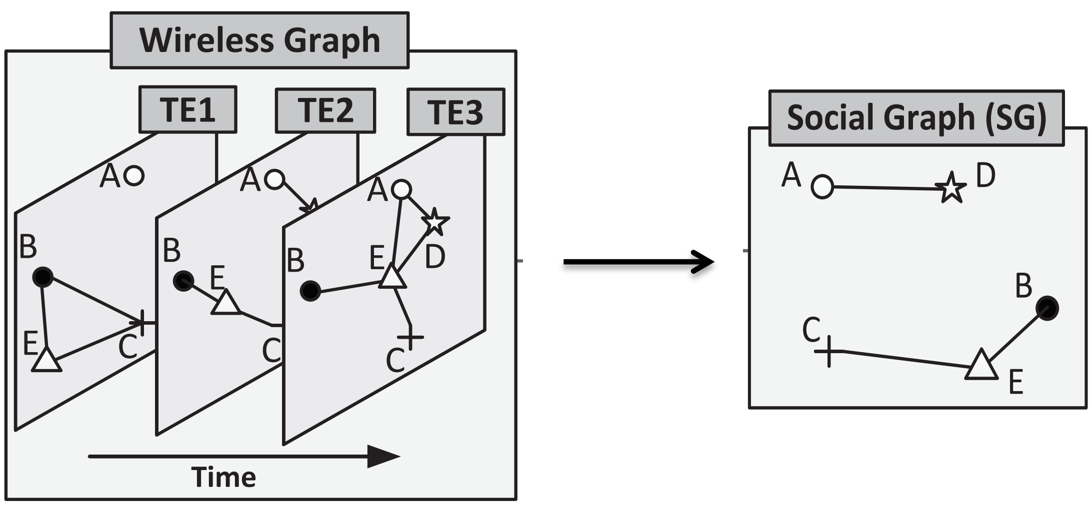
2. **Transfer information** in opportunistic wireless networks

3. **Context** and **content** wireless networking

1. **Measure and classify** social interactions

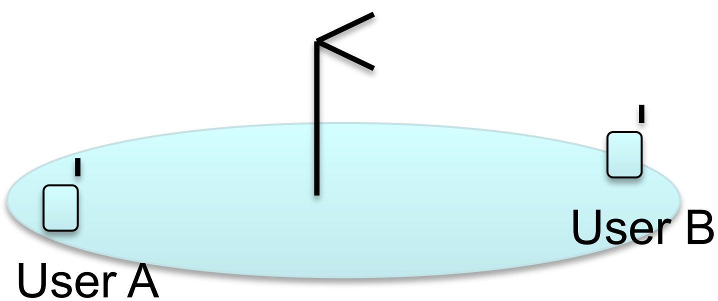
Objective: understand human interactions from measurements

- What we record: Intermittent physical wireless links
  - Intermittency originates from human mobility and habits
- Main problem:
  - Extract a social graph from measured physical interactions
  - Determine which intermittent link relates to regular vs. random interactions



Record interactions

- Open datasets exist (cf. Crawdad http://crawdad.cs.dartmouth.edu/)
- Different types of temporal contact measurements
  - Measure a **direct link between User A and B** (e.g. Bluetooth, WiFi Direct connectivity)
  - Assume a link exists between User A and User B if they are connected to the same WiFi access point
    - False positives !



  - Measure location of users (GPS): if users are close enough, assume they are connected
    - Distance-based threshold is unrealistic

Example 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

Dataset	Local	# entities	Duration	Type	Avg. # encounters/ node/day
Dartmouth [1]	campus	1156	2 months	Individuals	145.6
USC [2]	campus	4558	2 months	Individuals	23.8
San Francisco [3]	City	551	1 month	Cabs	834.7

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

[1] T. Henderson et al. "The changing usage of a mature campus-wide wireless network," in Proc. of ACM MobiCom 2004

[2] W. jen Hsu et al. "Impact: Investigation of mobile-user patterns across university campuses using wlan trace analysis," CoRR, vol. abs/cs/0508009, 2005

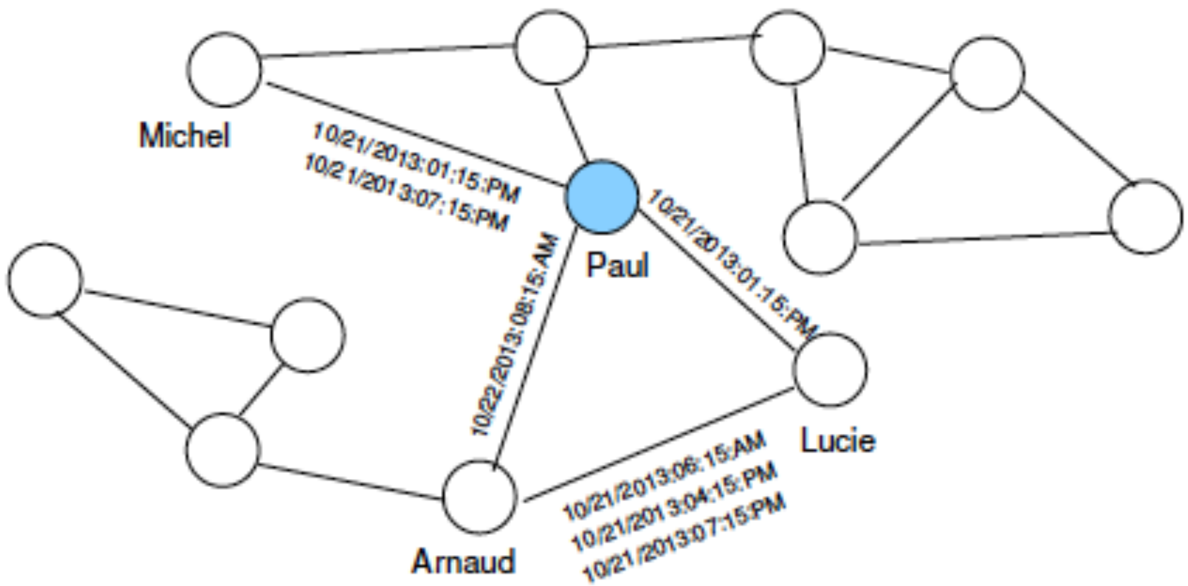
[3] A. Rojas et al. "Experimental validation of the random waypoint mobility model through a real world mobility trace for large geographical areas," in Proc. of the 8th ACM MSWiM 2005

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).

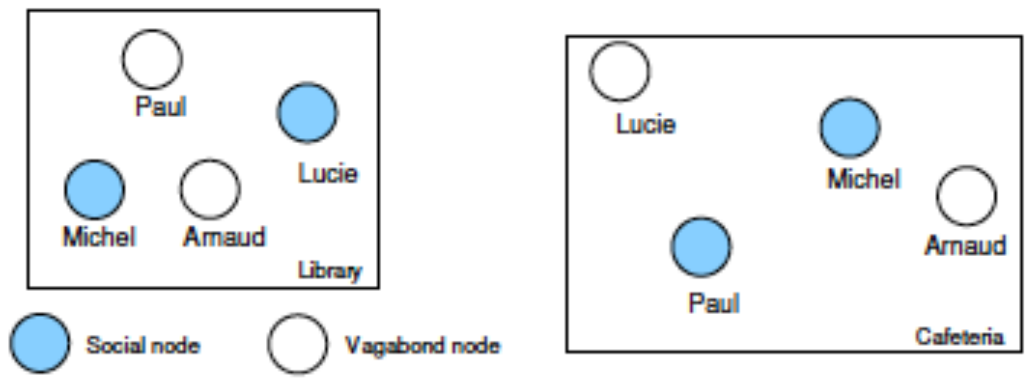


[4] A. G. Miklas et al., "Exploiting social interactions in mobile systems," in *Proceedings of the UbiComp '07*

Rationale and related initiatives

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.



- 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 → *geographical dependency*

[5] G. Zyba, G. Voelker, S. Ioannidis, and C. Diot, "Dissemination in opportunistic mobile ad-hoc networks: The power of the crowd, in *Infocom'11*

RECAST classifier [1]

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

Together with

Pedro O. Vaz de Melo, Antonio Loureiro – UMFG Brazil

Aline Viana - Inria, Marco Fiore - CNR Italy

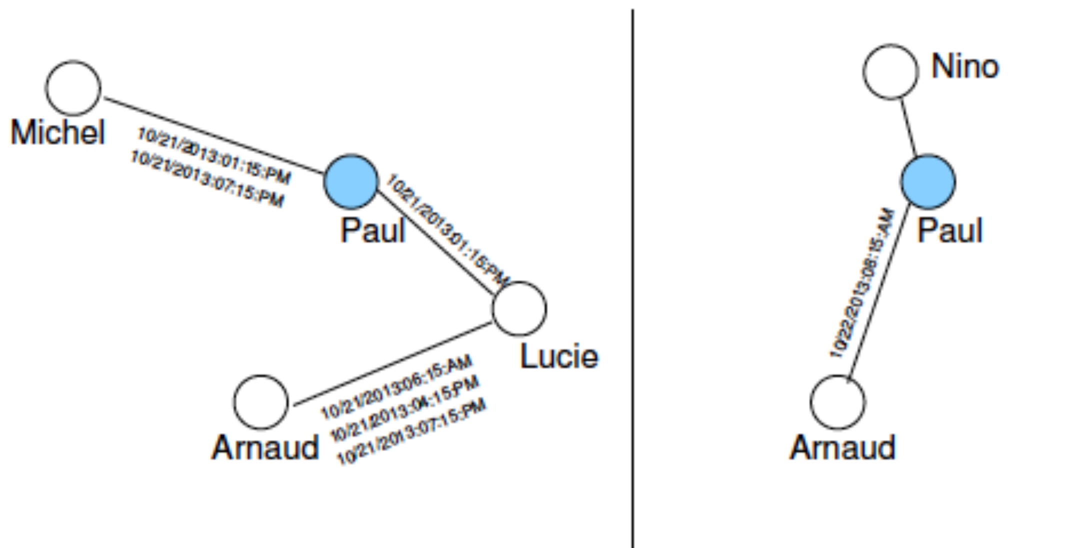
Frédéric Le Mouël – INSA Lyon

[1] RECAST: Telling Apart Social and Random Relationships in Dynamic Networks, P. Olmo Vaz de Melo, A. Viana, M. Fiore, K. Jaffrès-Runser, F. Le Moüel and A. A. F. Loureiro, 16th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (ACM MSWim 2013), Barcelona, Spain, 3-8 November 2013.

Graphs extracted from contact traces

Two possible representations

1. δ event graph: $\mathcal{G}_k(\mathcal{V}_k, \mathcal{E}_k)$
   There is an edge in $\mathcal{E}_k$ if contact within $\delta = 1$ day for instance.



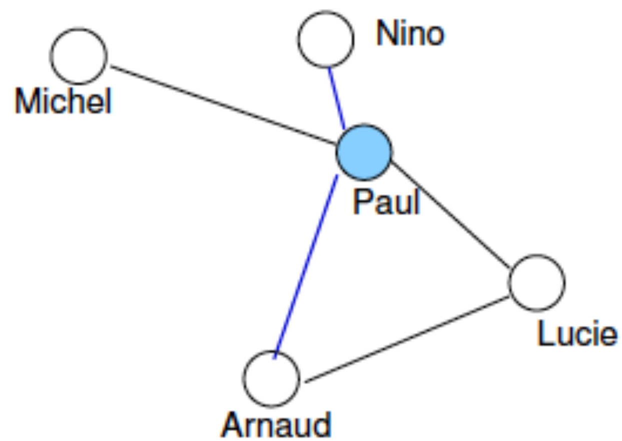
Day 1 event graph $\mathcal{G}_1(V_1, E_1)$ | Day 2 event graph $\mathcal{G}_2(V_2, E_2)$

2. Accumulative graph $G_t(V_t, E_t)$

Graphs extracted from contact traces

Two possible representations

1. δ event graph: $\mathcal{G}_k(\mathcal{V}_k, \mathcal{E}_k)$
   There is an edge in $\mathcal{E}_k$ if contact within $\delta = 1$ day for instance.
2. Accumulative graph $G_t(V_t, E_t)$: $G_t = \{\mathcal{G}_1 \cup \mathcal{G}_2 \cup \ldots \cup \mathcal{G}_t\}$



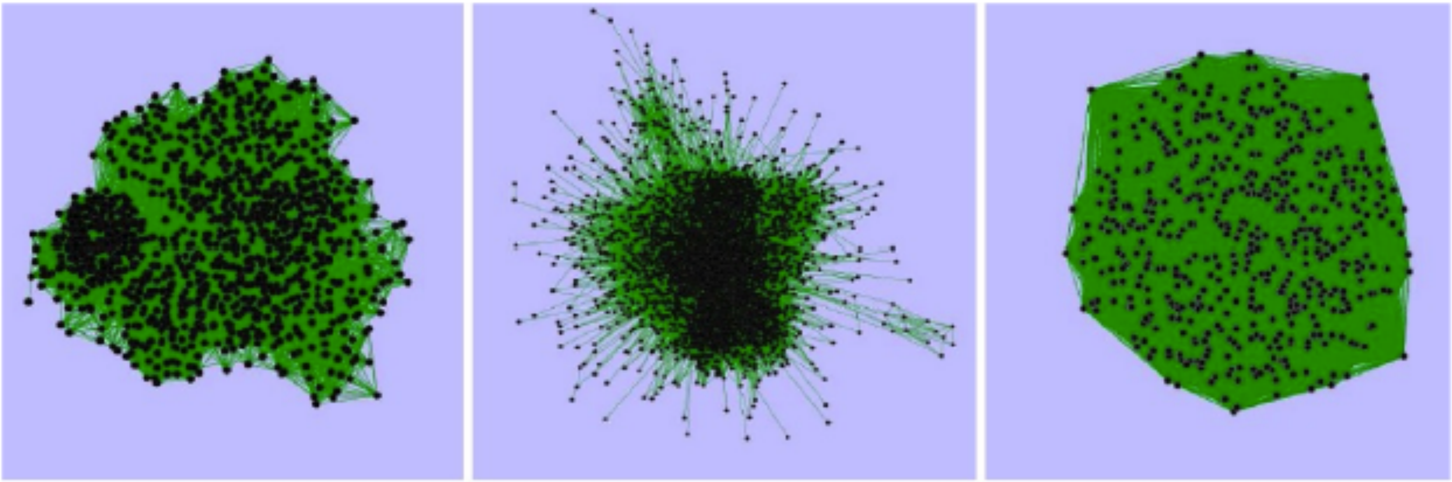
$G_2(V_2, E_2)$ Accumulative graph up to Day 2

Accumulates all event graphs up to time step t.

Graphs extracted from contact traces

Example accumulative graph G_t for $t = 2$ weeks

For $\delta = 1$ day and using force-direct layout algorithm for plotting



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

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

Social graph and 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[6]:

An edge is set between nodes of degree d_i and d_j with probability $p_{ij} = (d_i \times d_j)/\sum_{k=1}^{|V|} d_k$

Random accumulative graph G_t^R

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

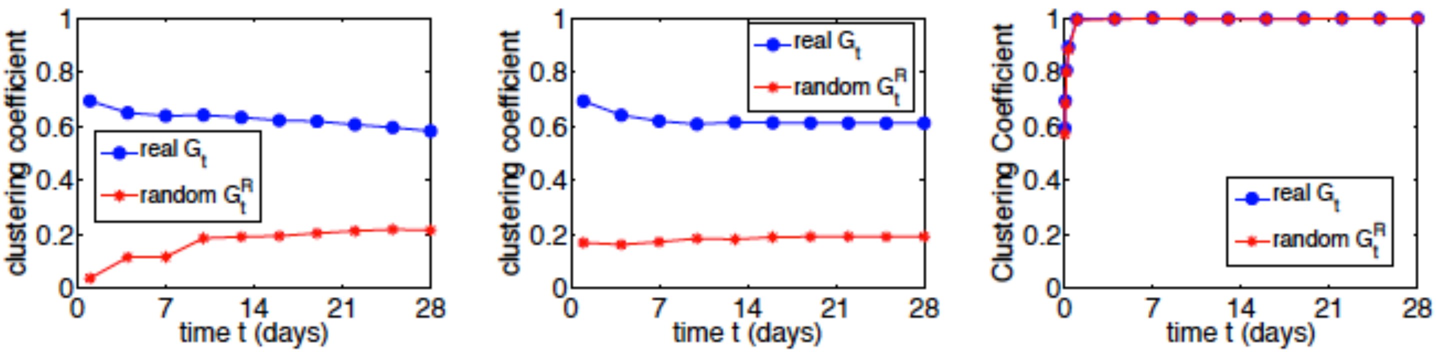
$$G_t^R = \{RND(\mathcal{G}_1) \cup RND(\mathcal{G}_2) \cup \ldots \cup RND(\mathcal{G}_t)\}$$

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

Comparison social vs. random graphs

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

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



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

- Dartmouth / USC traces have an order of magnitude higher $\bar{cc}$ than G^R → social decisions
- San Francisco: each individual taxi in the trace encounters most of the other taxis → closer to a random behavior

Social network features: Regularity and Similarity

Social nodes' behavior tend to

- repeat on a regular basis (because of daily activities for instance) $\rightarrow$ Regularity
- build persistent communities and generate common acquaintances $\rightarrow$ 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 $\{\mathcal{G}_i\}_{i\in[1,..,t]}$
- Topological overlap $to(i,j)$ [8]:
  Ratio of neighbors shared by two nodes calculated for the accumulative graph G_t.

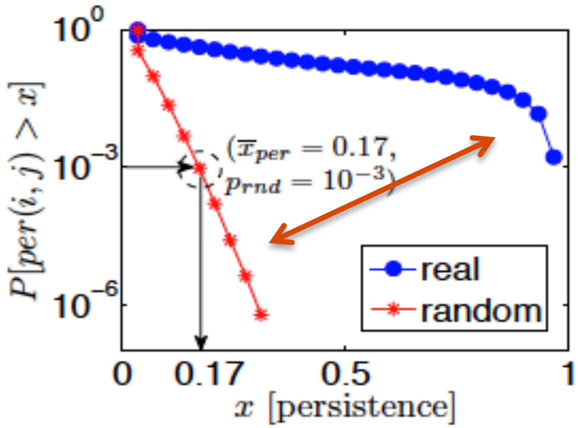
---

[7] N. Eagle et al., "From the Cover: Inferring friendship network structure by using mobile phone data," Proceedings of the National Academy of Sciences, Sept. 2009

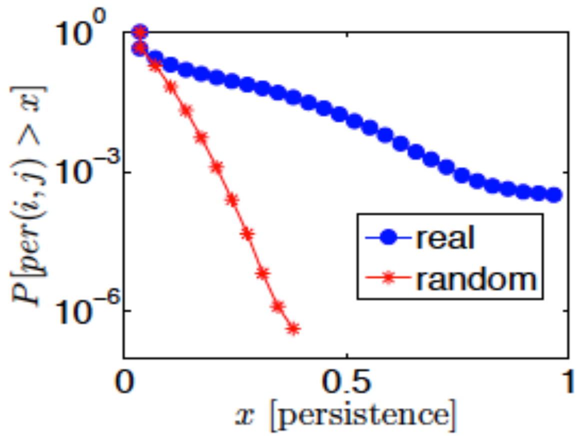
[8] J. P. Onnela et al., "Structure and tie strengths in mobile communication networks", Proc. of the National Academy of Sciences, May 2007

CCDF of edge persistence after 4 weeks

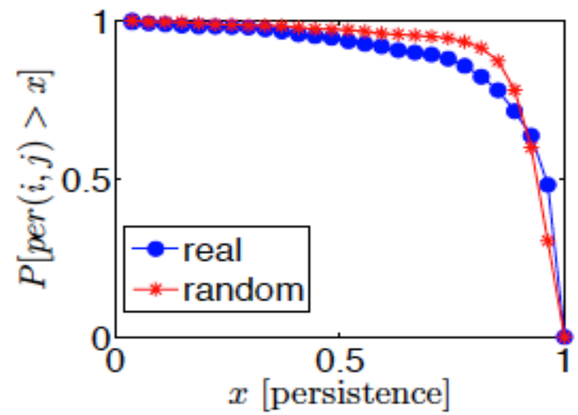
Individuals tend to see each other regularly



(d) Dartmouth



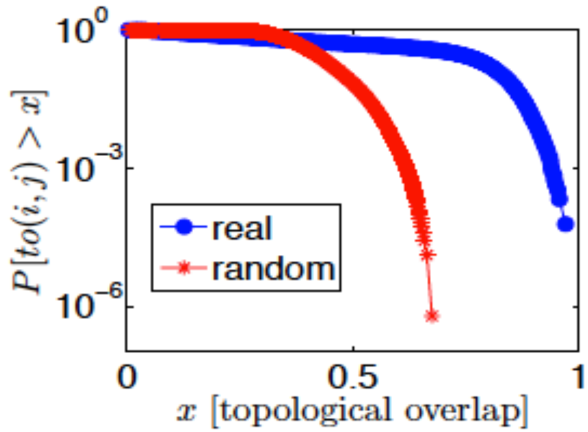
Encounters occur almost in a random fashion



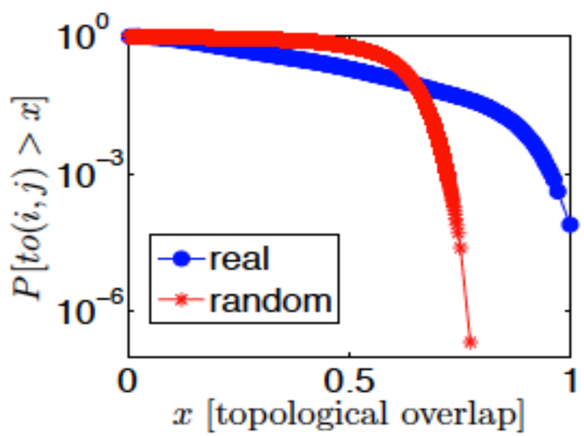
(f) San Francisco

CCFD of topological overlap after 4 weeks

Individuals of G_t have common neighbors

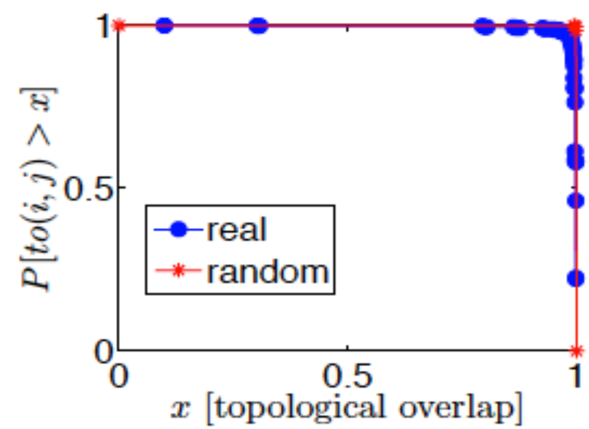


(g) Dartmouth



(h) USC

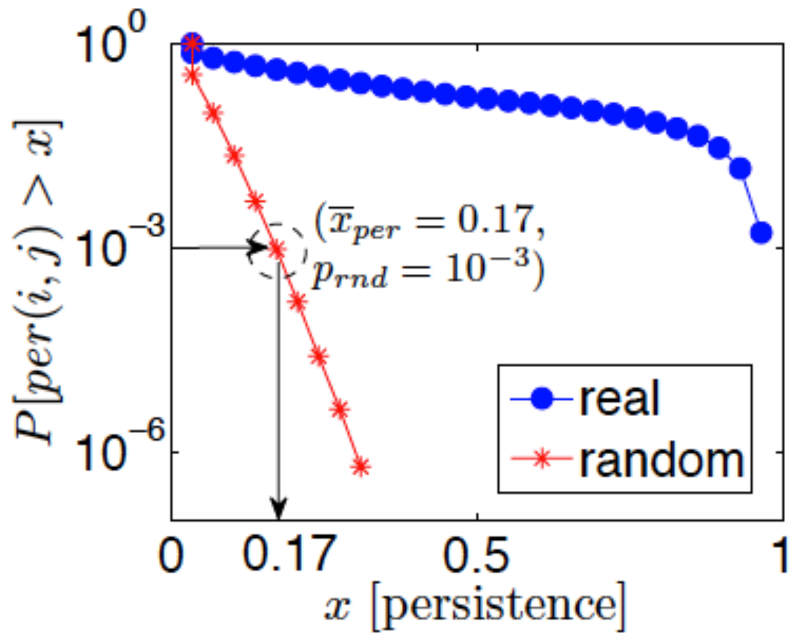
Common neighbors occur in a random fashion



(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 $\bar{x}_{per} = 0.17$.



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

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

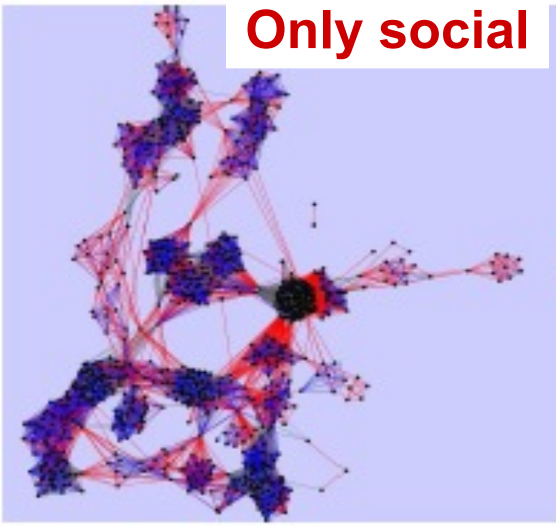
Only parameter of RECAST: p_{rnd}, the mis-classification error bound.

Main steps

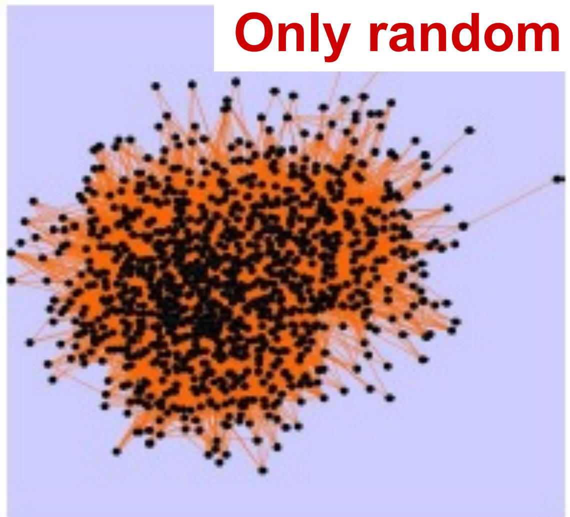
- Calculate the $per(i,j)$ and $to(i,j)$ for each edge
- Knowing p_{rnd}, calculate $\bar{x}_{per}$ and $\bar{x}_{to}$ from CCDF's
- For each edge,
  - if $per(i,j) > \bar{x}_{per} \rightarrow (i,j)$ is social for edge persistence
    else (i,j) is random for edge persistence
  - if $to(i,j) > \bar{x}_{to}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

Class	**Edge persistence**	**Topological overlap**
Friends	social	social
Acquaintances	random	social
Bridges	social	random
Random	random	random

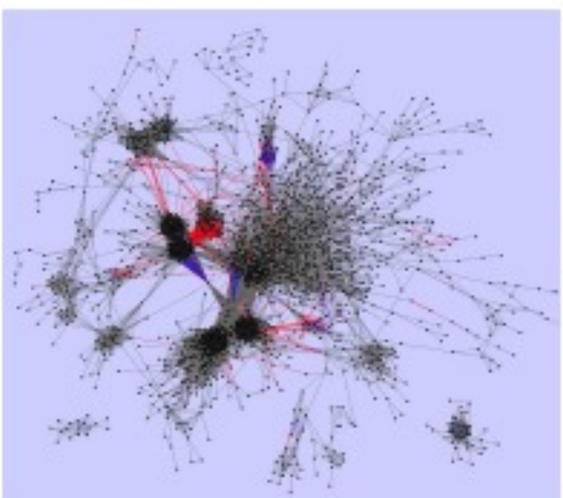
Classification after 2 weeks



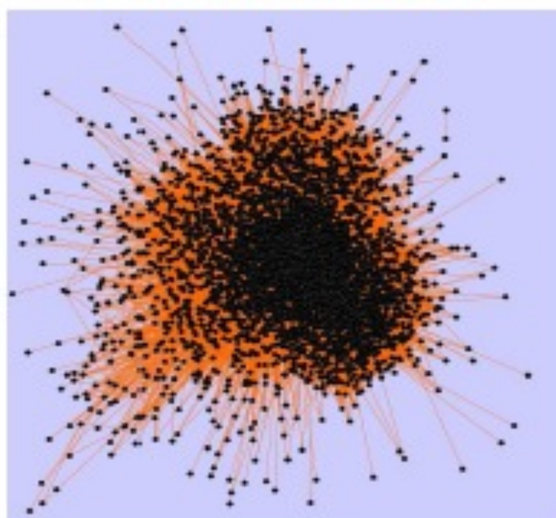
(a) Dartmouth, only social edges



(b) Dartmouth, only random edges



(c) USC, only social edges



(d) USC, only random edges

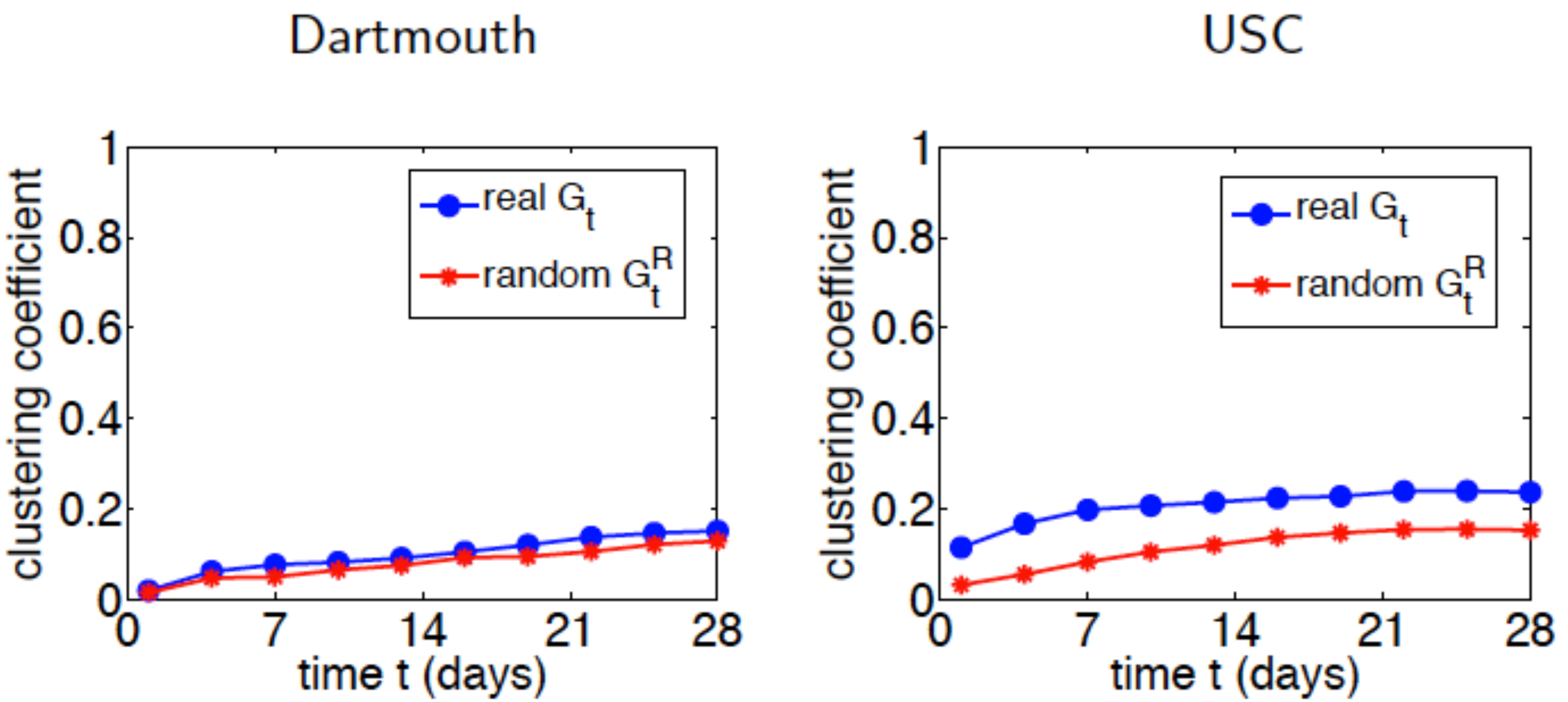
Friends edges are in blue
Bridges edges are in red
Acquaintance edges are in gray
Random edges are in orange

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

- Random-edges network
No structure appears, looking like random graphs

.uchon, 2014

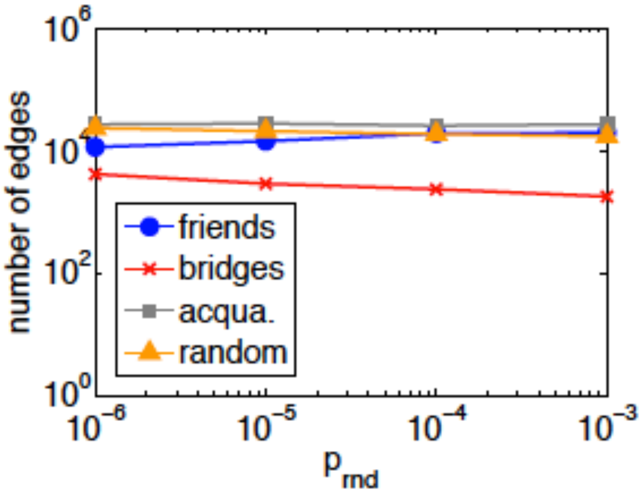
Cluster coefficient analysis for **random edges only**



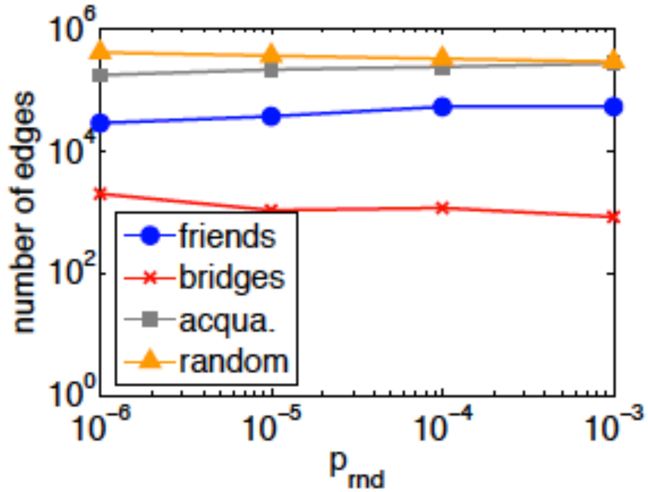
Validates the efficiency of RECAST to identify random edges for Dartmouth and USC

Impact of p_{rnd}

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



Dartmouth



USC

RECAST is not sensitive to p_{rnd} !

Outline

1. **Measure and classify** social interactions
   - RECAST algorithm

2. **Transfer information** in opportunistic wireless networks

3. **Context** and **content** wireless networking

2. **Transfer information** in opportunistic wireless networks

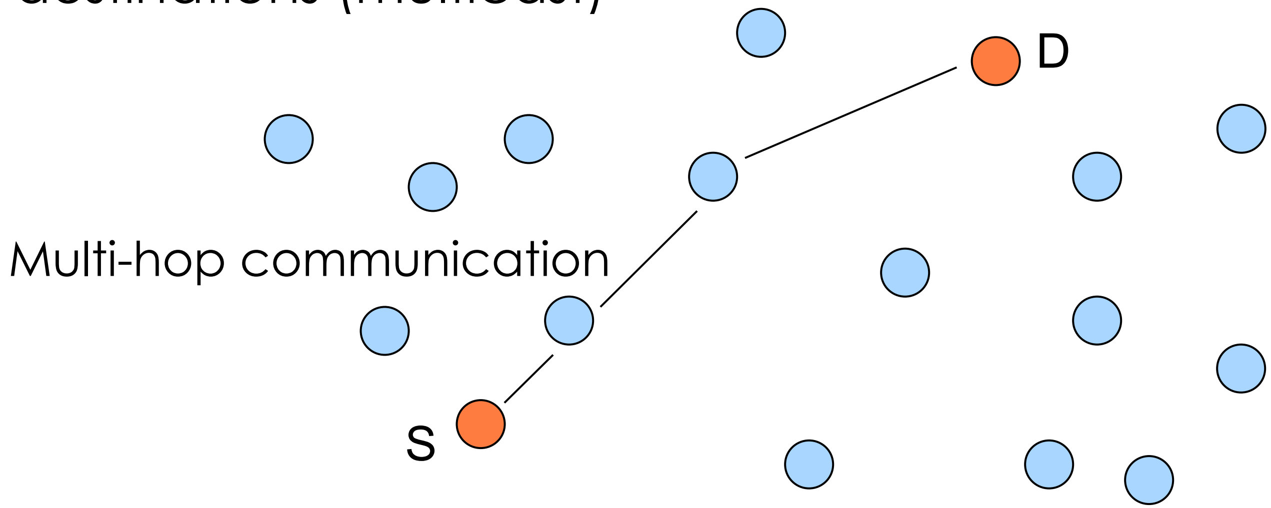
Two different problems exist in wireless networking:

- Information **dissemination** (i.e. broadcast)

Transfer a set of messages to **all nodes** of the network

- Information **routing** (unicast or multicast)

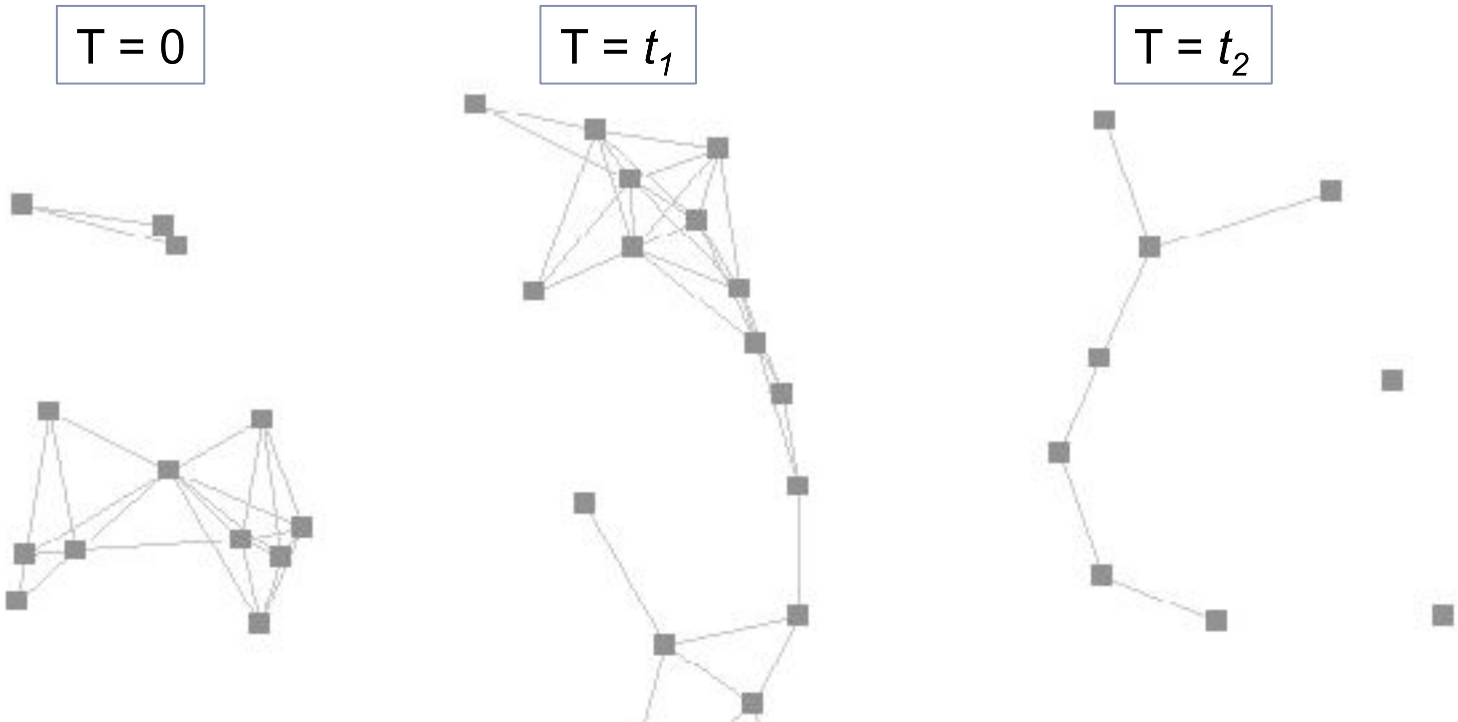
Transfer a set of messages to **a unique destination** (unicast) or a set of destinations (multicast)



2. **Transfer information** in opportunistic wireless networks

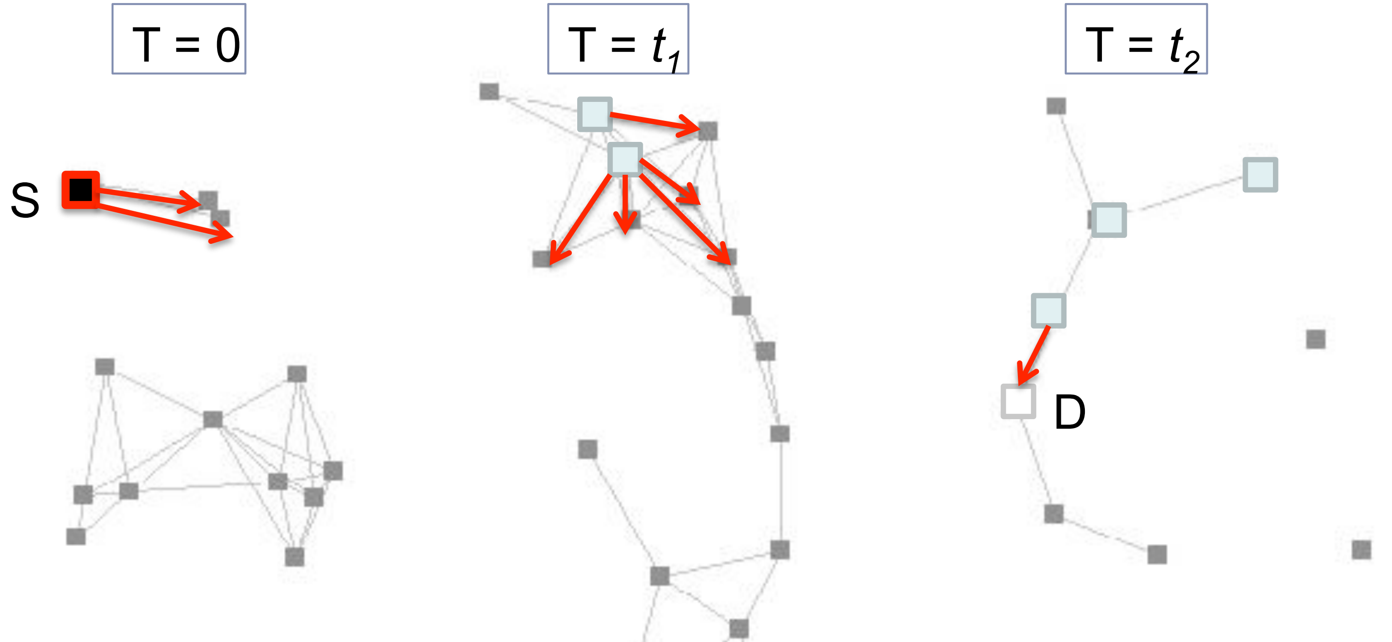
BUT in opportunistic wireless networks,

- there is no end-to-end path available **at all times**
- only **delay tolerant data** can be forwarded in such conditions



‘Social agnostic’ opportunistic routing protocols

- Direct delivery: the source node carries its data until it meets the destination, eventually
  - The slowest but no overhead
  - Lowest delivery ratio
- Epidemic (flooding)
  - The fastest but highest overhead (i.e. nb of replicates)
  - Best delivery ratio for infinite buffers



‘Social-agnostic’ opportunistic routing protocols

Objective : Keep the same delivery ratio than epidemic, but with as little replicates as possible

Best solution known so far: **Spray and Wait**

- Source emits L copies of the message: **Spray phase**
  - Gives a copy to the L first encountered nodes.
- All message carriers wait to deliver their copy to D: **Wait phase**

- Alternative binary spray phase:
  - The source gives L/2 copies to the 1st encountered node.
  - Then, at each encounter, a carrier node gives the half of its copies to be new carrier.
  - Wait phase start once a node has only one copy left

Spray and Wait performance

Spray and Wait beats Epidemic because of **limited buffer** size

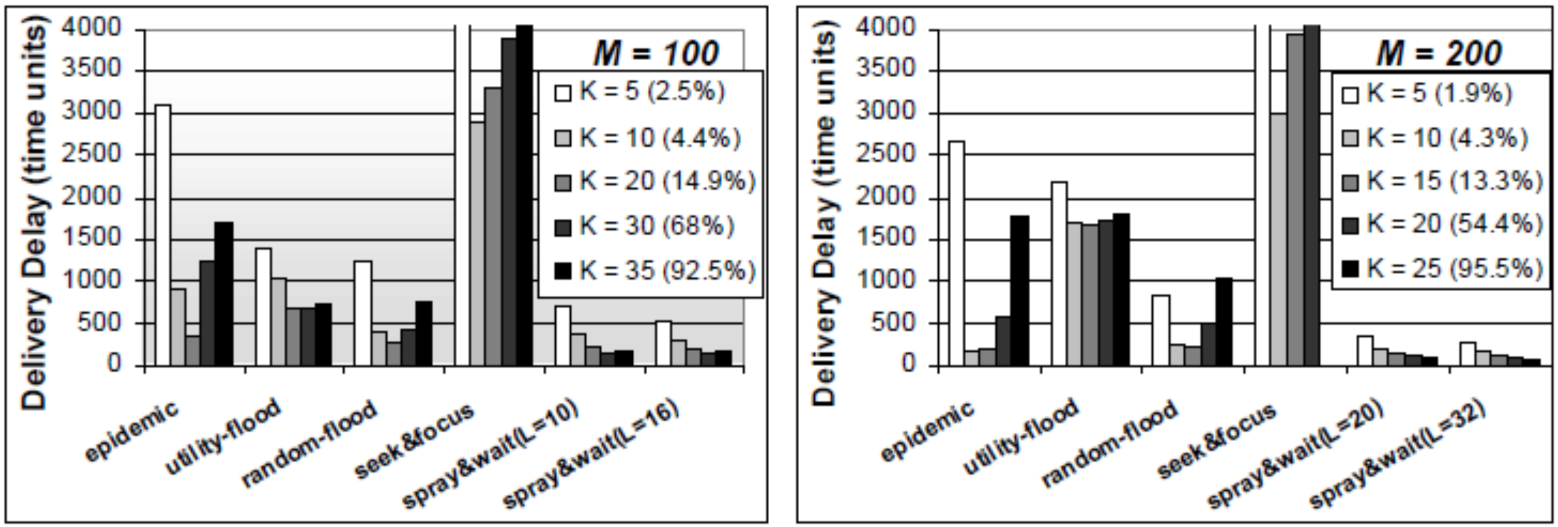


Figure 7: Scenario B - Delivery delay as a function of number of nodes M and transmission range K.

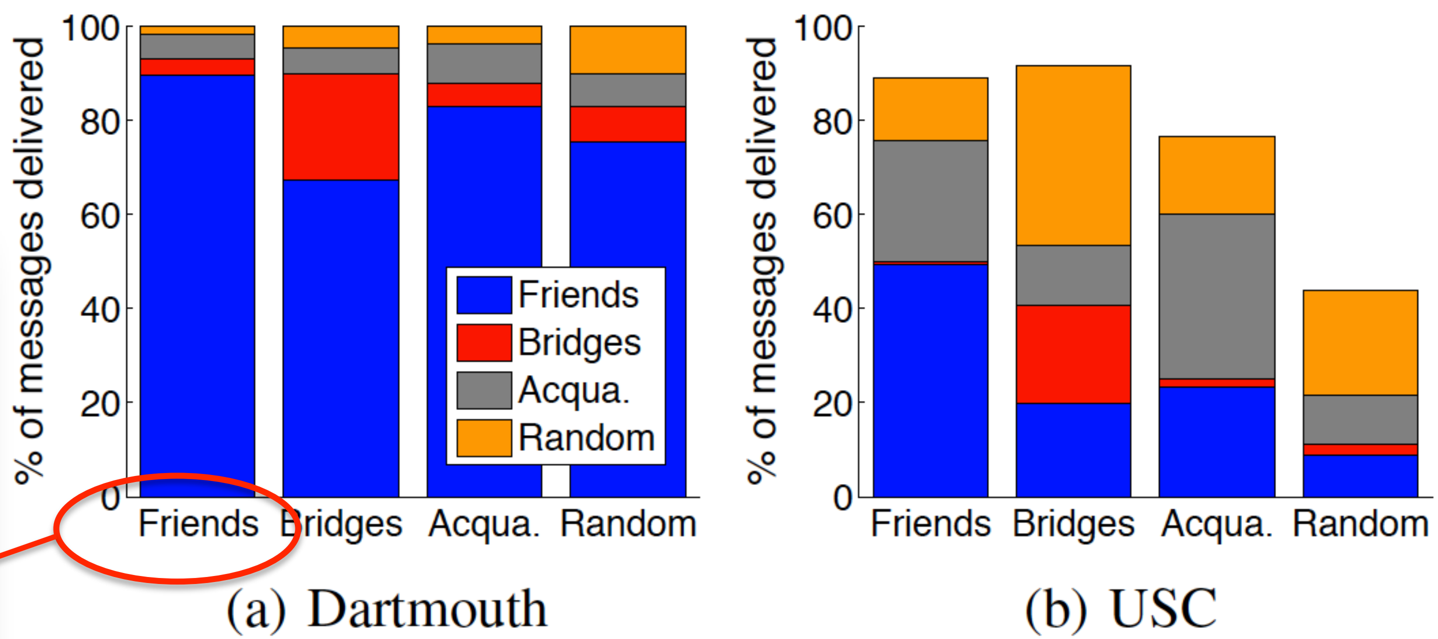
Social-aware routing

Is it worth accounting for the social graph ?

Let's assume we start an epidemic transmission between a source and a destination that share a edge in the social network.

(Social graph calculated with 4 first weeks of data set)

Which edges participate in the forwarding in the following 2 weeks?

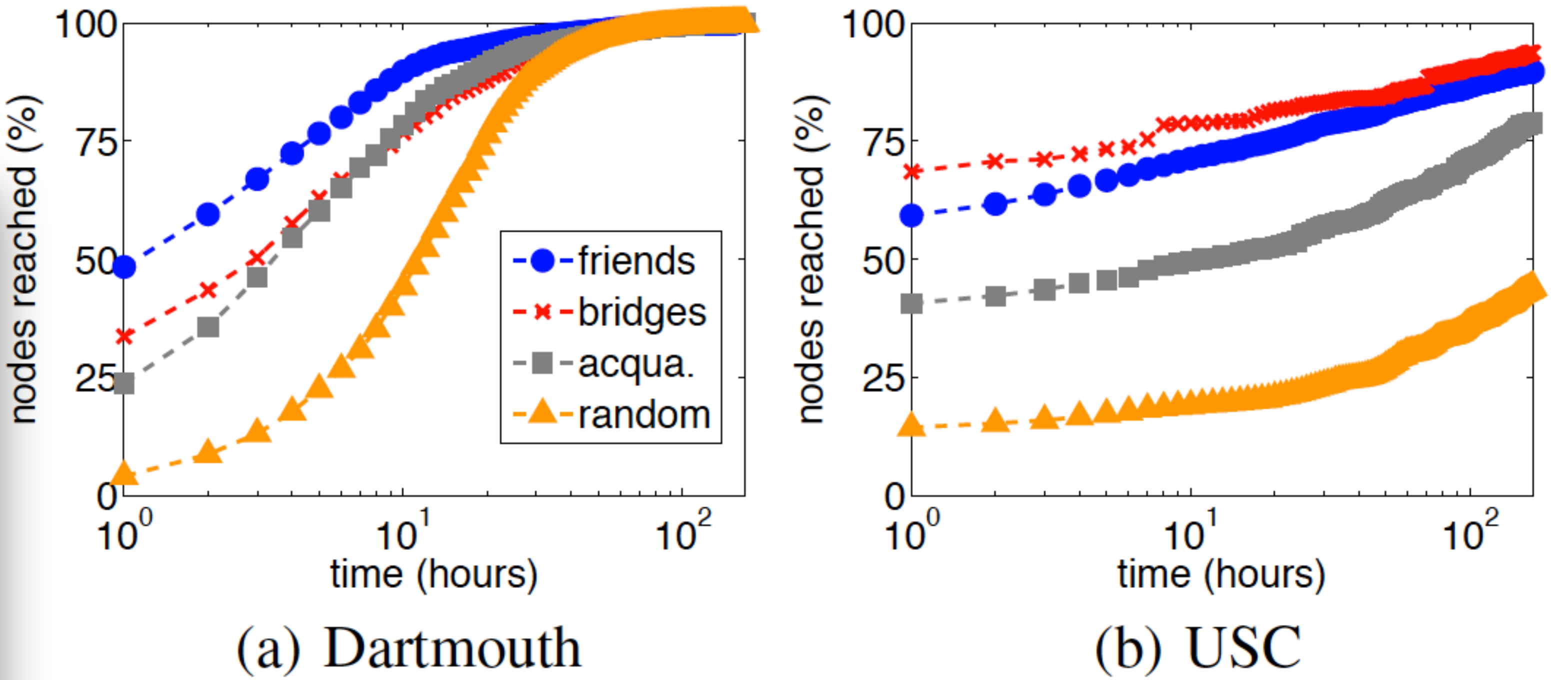


S and D are friends

Social-aware routing

Is it worth accounting for the social graph ?

- The routing is much faster between nodes that share a social relationship
- Edge persistence has a strong impact on the routing efficiency.
- But random help as well…



How to design a social-aware routing protocol?

- Rely on centrality metrics and community detection
  - Betweenness
  - Similarity
  - Persistence
  - K-cliques,...
- State of the art solutions
  - SimBetTS [1]
  - BUBBLE Rap [2]
  - Peoplerank [3]
- Key issues :
  1. How to calculate these metrics in a distributed manner?
  2. How to use them to route data efficiently?

[1] E. Daly and M. Haahr, "Social Network Analysis for Information Flow in Disconnected Delay-Tolerant MANETs," IEEE Transactions on Mobile Computing, vol. 8, no. 5, pp. 606 –621, May 2009.

[2] P. Hui, J. Crowcroft, and E. Yoneki, "BUBBLE Rap: Social-based Forwarding in Delay Tolerant Networks," IEEE Transactions on Mobile Computing, Dec. 2010.

[3] Mtibaa, A., May, M., Diot, C., Ammar, M.: Peoplerank: social opportunistic forwarding. In: Proceedings of the 29th conference on Information communications, INFOCOM'10, pp. 111–115. IEEE Press, Piscataway, NJ, USA (2010)

SimBetTS

- Social metrics considered
  - Similarity (~ topological overlap)

  Number of common neighbors between two nodes
  - Betweenness

  Number of times a node lies on the shortest path between any source-destination pair of the network
  - Tie strength = Frequency + Intimacy + Recency

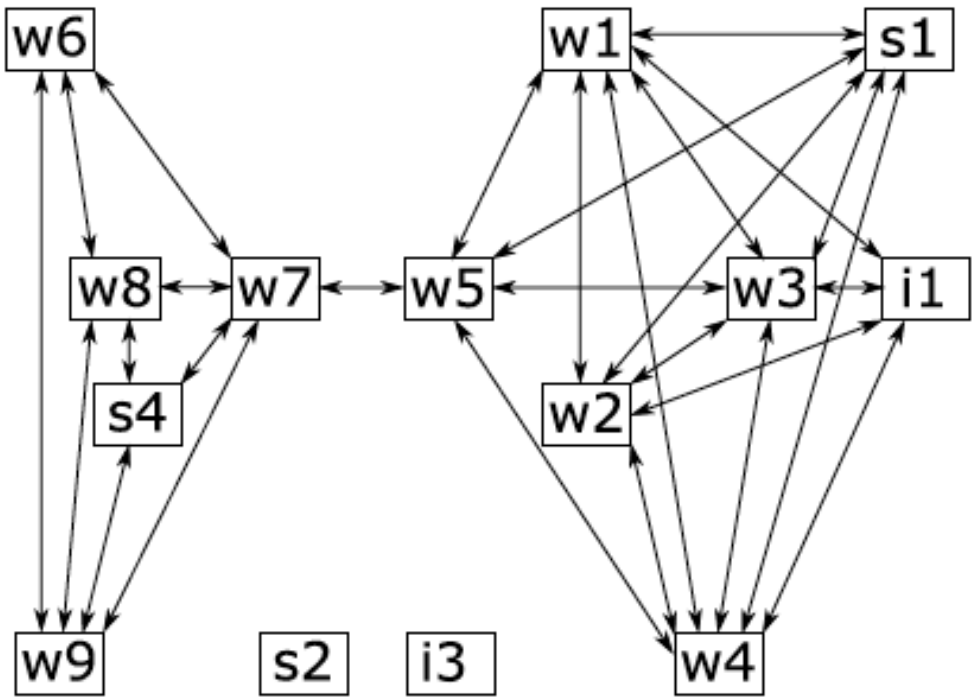
  (how frequent, how long and how recent)

- Decentralized computation
  - Use of an ego-network [1]
  - Each node stores the adjacency matrix relative to the contacts encountered made together

[1] P. V. Marsden. Egocentric and sociocentric measures of network centrality. Social networks, 24(4):407–422, October 2002

SimBetTS

- Egocentric computation [1]



Node	Sociocentric betweenness	Egocentric betweenness
w1	3.75	0.83
w2	0.25	0.25
w3	3.75	0.83
w4	3.75	0.83
w5	30	4
w6	0	0
w7	28.33	4.33
w8	0.33	0.33
w9	0.33	0.33
s1	1.5	0.25
s2	0	0
s4	0	0
i1	0	0
i3	0	0

[1] E. Daly and M. Haahr, "Social Network Analysis for Information Flow in Disconnected Delay-Tolerant MANETs," IEEE Transactions on Mobile Computing, vol. 8, no. 5, pp. 606 –621, May 2009.

SimBetTS

- Routing with SimBetTS metrics

- As two nodes *n* and *m* encounter, each node calculates for each destination the sum of these three utilities:

$$SimUtil_n(d) = \frac{Sim_n(d)}{Sim_n(d) + Sim_m(d)} \tag{14}$$

$$BetUtil_n = \frac{Bet_n}{Bet_n + Bet_m} \tag{15}$$

$$TSUtil_n(d) = \frac{TieStrength_n(d)}{TieStrength_n(d) + TieStrength_m(d)} \tag{16}$$

The message is kept or transferred to the node with the highest utility.

- An initial replication value R is assigned to a message. If R>1, the message is replicated and R is divided between the two nodes dependent on the SimBetTS utility value.

BUBBLE Rap

- Social metrics considered
  - Node centrality (betweenness, degree…)
  - Community detection:
    - k-clique community detection
    - Newman's weighted network analysis

- Decentralized computation
  - For node centrality: (no betweenness approx.)
    - number of encoutered nodes in the last 6 hours
    - average of the number of encountered nodes in the last 4 periods of 6 hours (last day)
  - For community detection
    - A variation of Clauset's[1] community detection with local modularity
    - Detection accuracy can be up to 85% of centralized K-clique algorithm.

[1] A. Clauset. Finding local community structure in networks. Physical Review E, 72:026132, 2005.

BUBBLE Rap

- Routing with centrality metrics only
  - Forward data only to nodes with higher centrality metric
  - Hierarchical path issue
    - Shortest end-to-end paths see an increase of node centralities, then a decrease for final delivery
    - Consequence: **in large networks**, messages may get stuck in a high degree node with no edge to the destination node.
- Routing with community labels only
  - Achieves bad performance if people of different communities do not mix together

- Main idea of BUBBLE Rap:

Use a Label per community and 2 centrality metrics

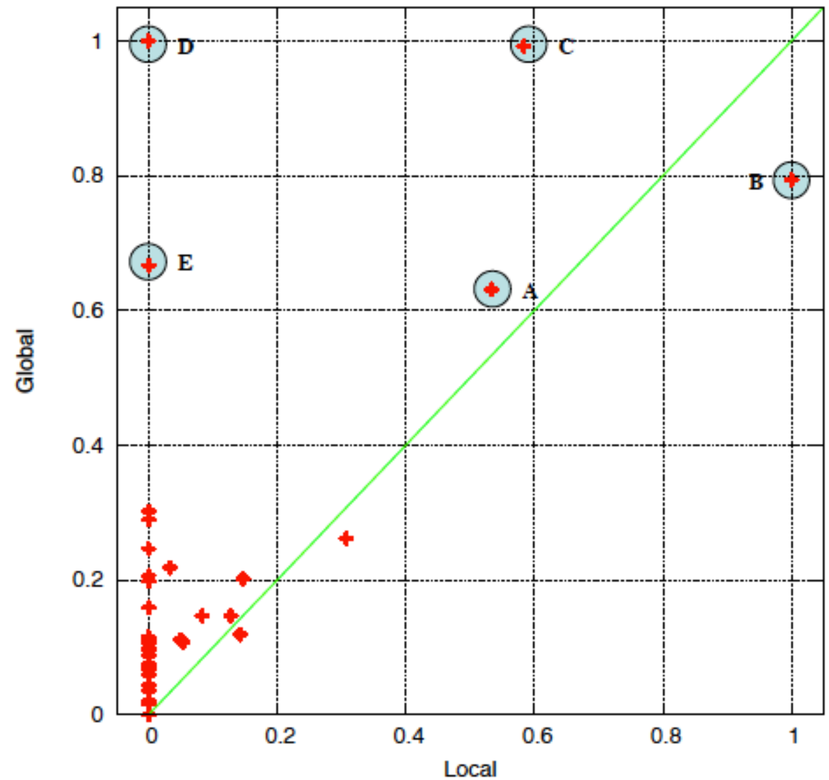
  - **Global** centrality metric – calculated **for the whole network**
  - **Local** centrality metric – calculated only **per community**

BUBBLE Rap

- Correlation of global centrality and local centrality of a given community A

- If you choose D or E, which are outside community A -> Never get to a destination of community A

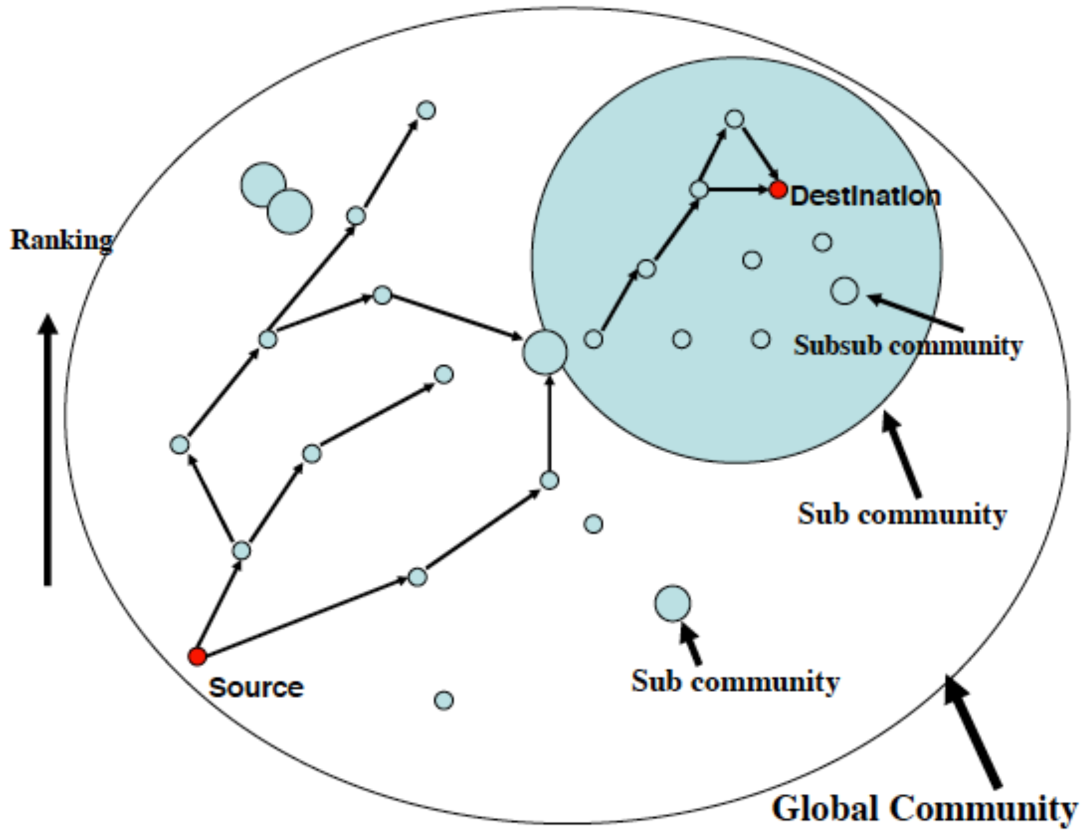
You are more lucky if you pick A,B or C.



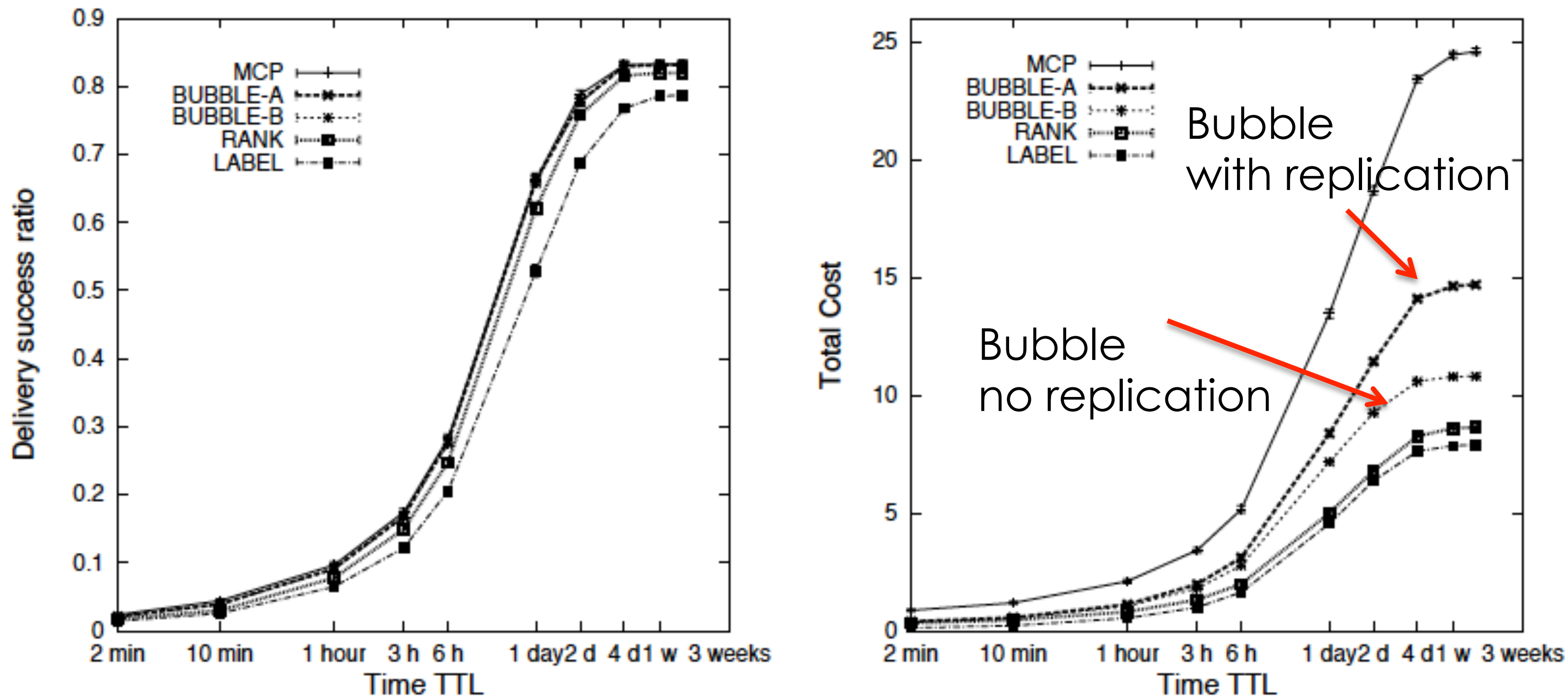
As soon as you reach the community of your destination, use local centrality

BUBBLE Rap

- Illustration of Bubble rap forwarding



BUBBLE Rap



Comparisons of several algorithms on *Cambridge* dataset

Social opportunistic routing

- Conventional routing fails in opportunistic wireless networks
- The knowledge of social dynamics improves data forwarding performance

- But **only considering social edges** for data forwarding is not enough
  - Non socially connected edges can bring connectivity
  - Random edges in RECAST could thus be leveraged as well

- Most of the solutions do not investigate the daily routines of nodes
  - It would be good to learn and then forecast future encounter periods of nodes
  - Maybe have several social graphs depending on the time of the day ?

Outline

1. **Measure and classify** social interactions
   - RECAST algorithm

2. **Transfer information** in opportunistic wireless networks

3. **Context** and **content** wireless networking

3. Context and **content** wireless networking

- In wireless networking
  - Previous research has leveraged CONTEXT information
    - Mobility,
    - Spectrum,
    - Available wireless technologies
    - …

- Now, what can be do if we can predict a portion of the content users will look for?
  - Content can be linked to a community's interests

  So I can push data to a community (implicit multicast)
  - If there are several networks available (WiFi, 3G, ..)

  I can '**pre-load**' data in the network using the less expensive technology

  …

3. Context and content wireless networking

- MACACO project
  - EU FP7, CHIST-ERA call, started Nov. 2013
- Our focus : a more intelligent data offloading strategy
  - Build data offloading mechanisms that take advantage of **context** and **content information**

Users Interest

Mobility information: Time and location

- Intuitions:
  - to **extract** and **forecast** the behaviour of mobile users in the three-dimensional space of **time, location** and **interest**
    - **'what', 'when' and 'where'** users are pulling data from the network
  - to **pre-fetch** the identified data and **cache** it at an earlier time
    - at the mobile terminals or at the edge nodes of the network

Project contributions

1. **To acquire real world data sets** to model mobile node behavior in the three-dimensional space

2. To derive appropriate **social models** for the **correlation between user interests and their mobility.**

3. To derive simple and efficient prediction algorithms to forecast the **node's mobility and interests**

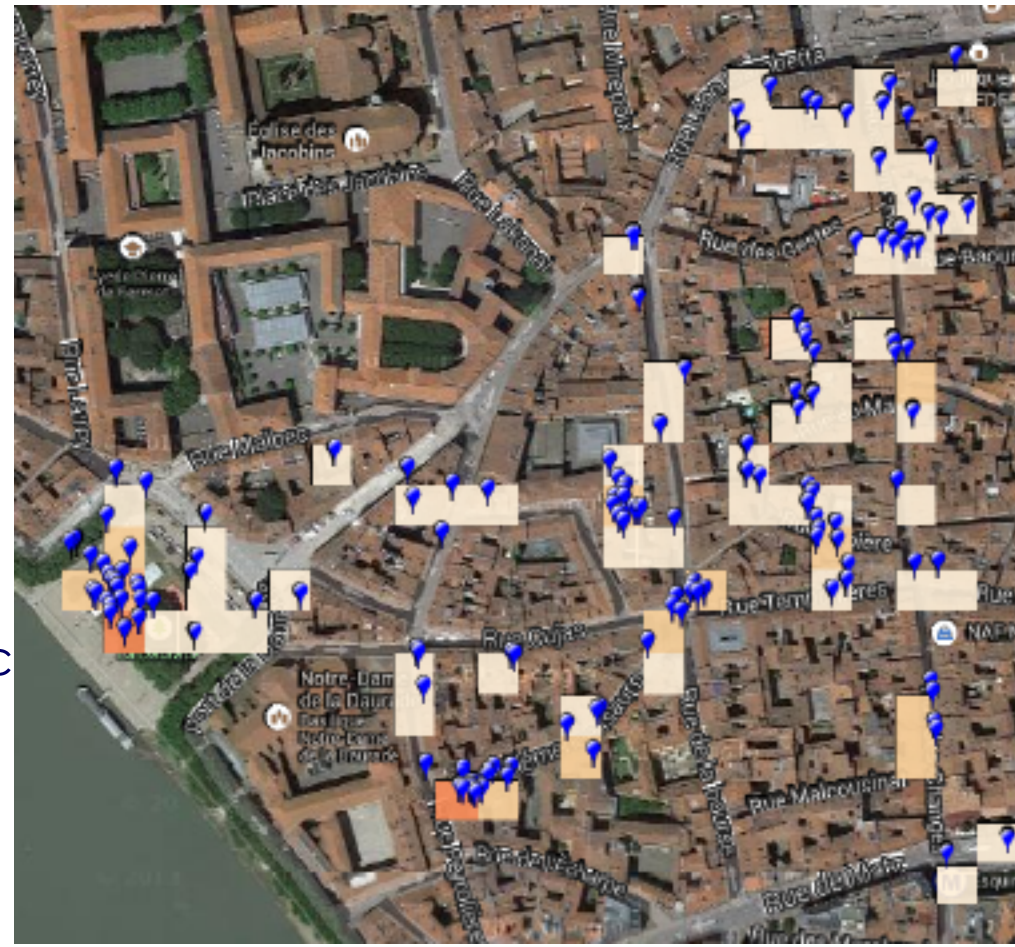
4. To output data pre-fetching mechanisms
   1. To integrate content-centric caching approach with social context awareness and opportunistic resource availability

5. To design a federated testbed for (no commercial interest):
   1. Content and context data collection
   2. Assessment of off-loading solutions

Gather context and content data

A smartphone application that measures:

- Context data
  - Location (GPS, Internet)
  - WiFi connectivity
  - Bluetooth connectivity
  - Cellular network towers
- Content data
  - Name of applications that have generated traffic
  - Browser history
  - Facebook network



Next…

- Having this data, exhibit the correlations between content and context
  - Do users have regular habits in data usage?
  - If yes, is it possible to model these networks with the content plane in mind?
- Using network models, deriving data pre-fetching strategies to adjust the load off available networks

….