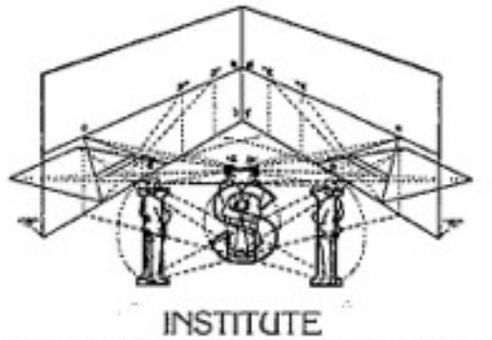
On- and off-line interaction networks

Alain Barrat

CPT, Marseille, France

ISI, Turin, Italy



Outline

- Online social networks: a case study
  - network analysis
  - measuring homophily
  - measuring selection and influence mechanisms
- Mining face-to-face interactions
- Online vs offline networks
  - comparison
  - predicting links?

Social networks

- Huge field of research
- Data: mostly small samples, surveys
- Multiplexity
- Longitudinal data

Issue of data mining

McPherson et al, Annu. Rev. Sociol. (2001)

New technologies

- Email networks
- Cellphone call networks
  - Mobility patterns
  - Interaction networks
- Real-world interactions
  - MIT reality mining
  - Sociopatterns.org
- Online networks/ social web

NEW DATASETS,

-longitudinal data

-on vs offline comparison

Case study: aNobii

(similar analysis done also for last.fm and flickr)

- Social network
- ~100 000-150 000 users
- "specialized/topical" content-sharing site
- Users expose profiles (content) and links:
  - Books read by user; Wishlist of books
  - Tags describing the books
  - Groups of discussion
  - Geographical information
- Communication between users

Network properties

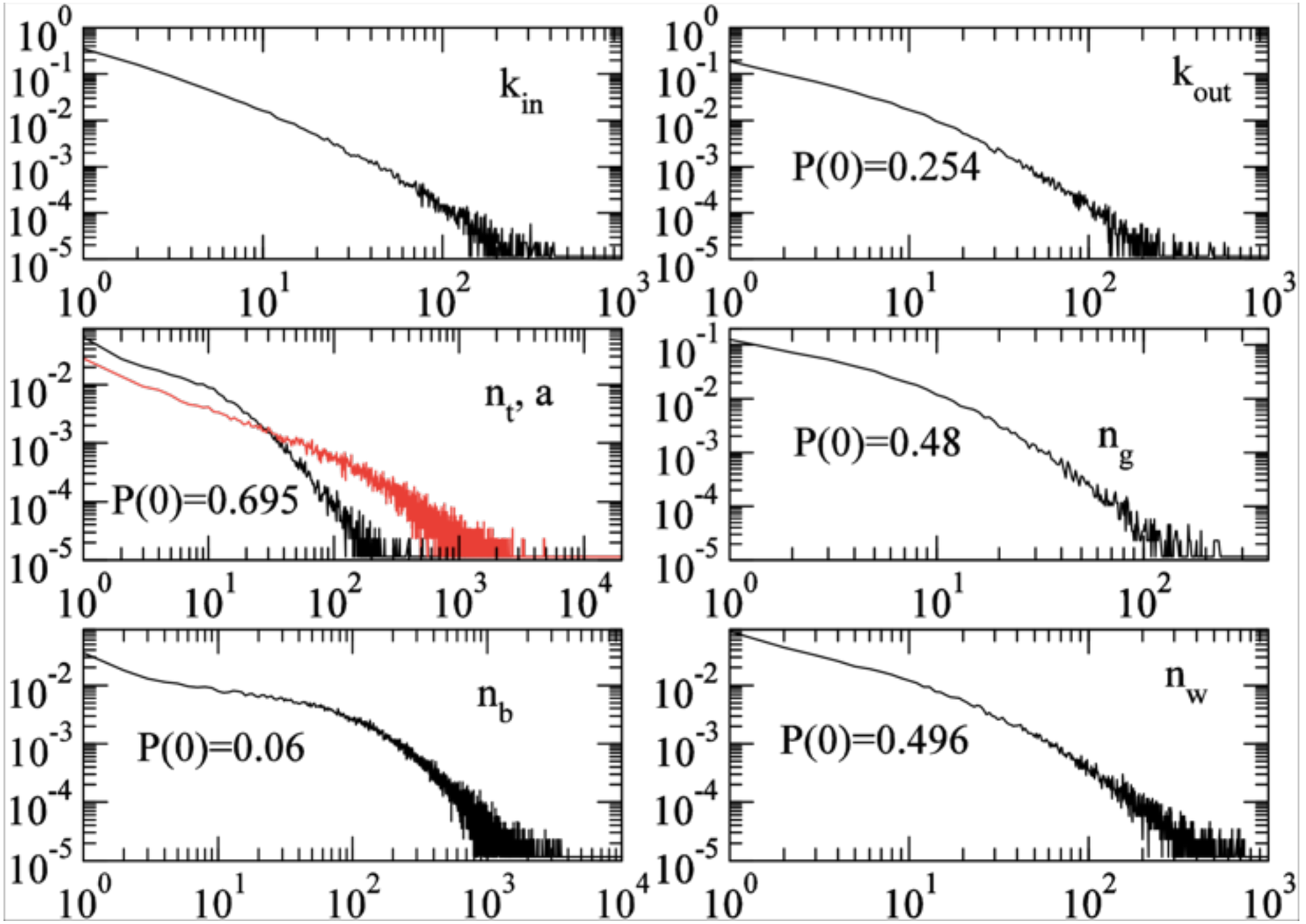
Activity measures

Heterogeneity of all users' activity amounts

Networking

Tagging/Groups

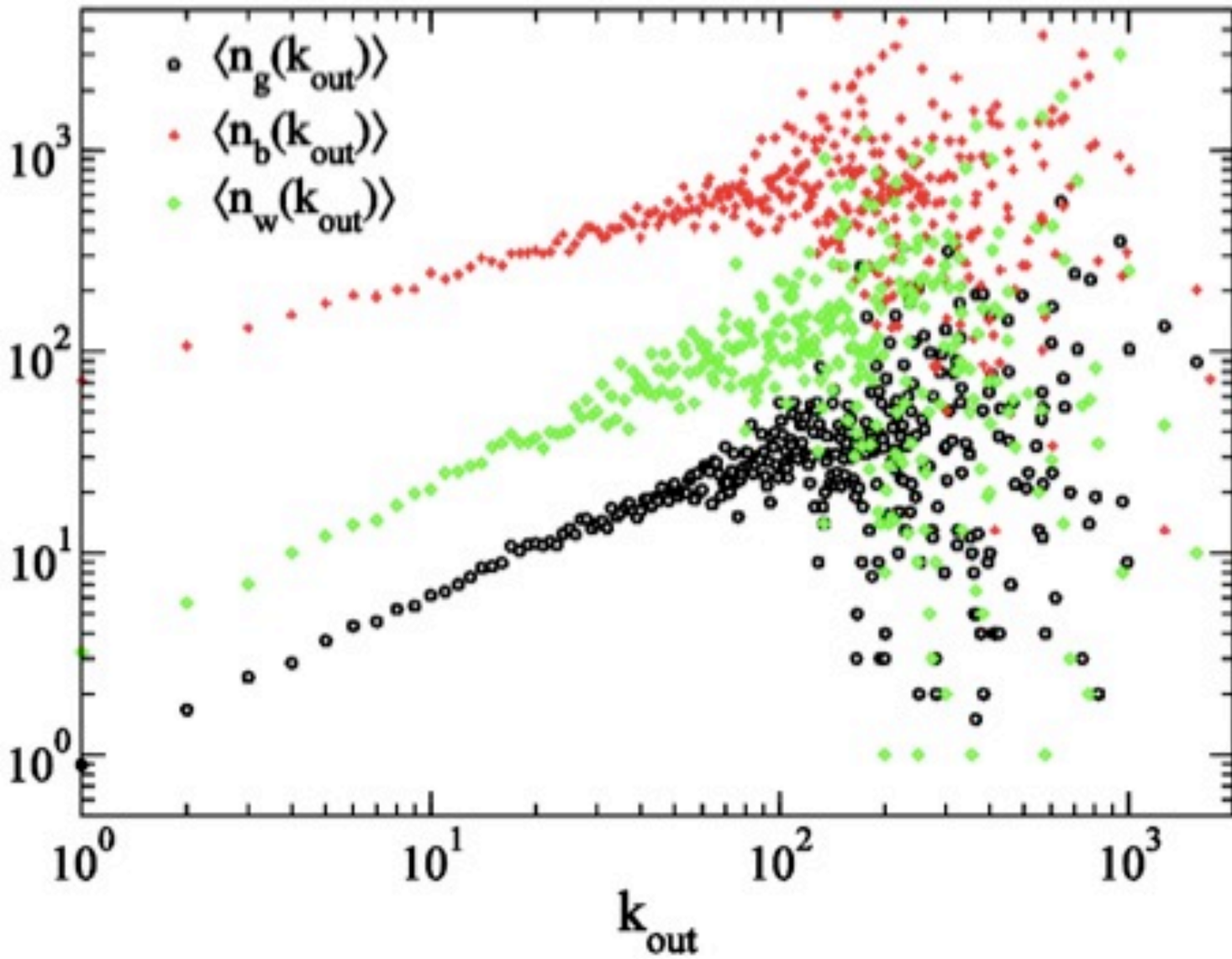
Books



Correlations

Correlation between user's activity types:

Sharing and annotating activities



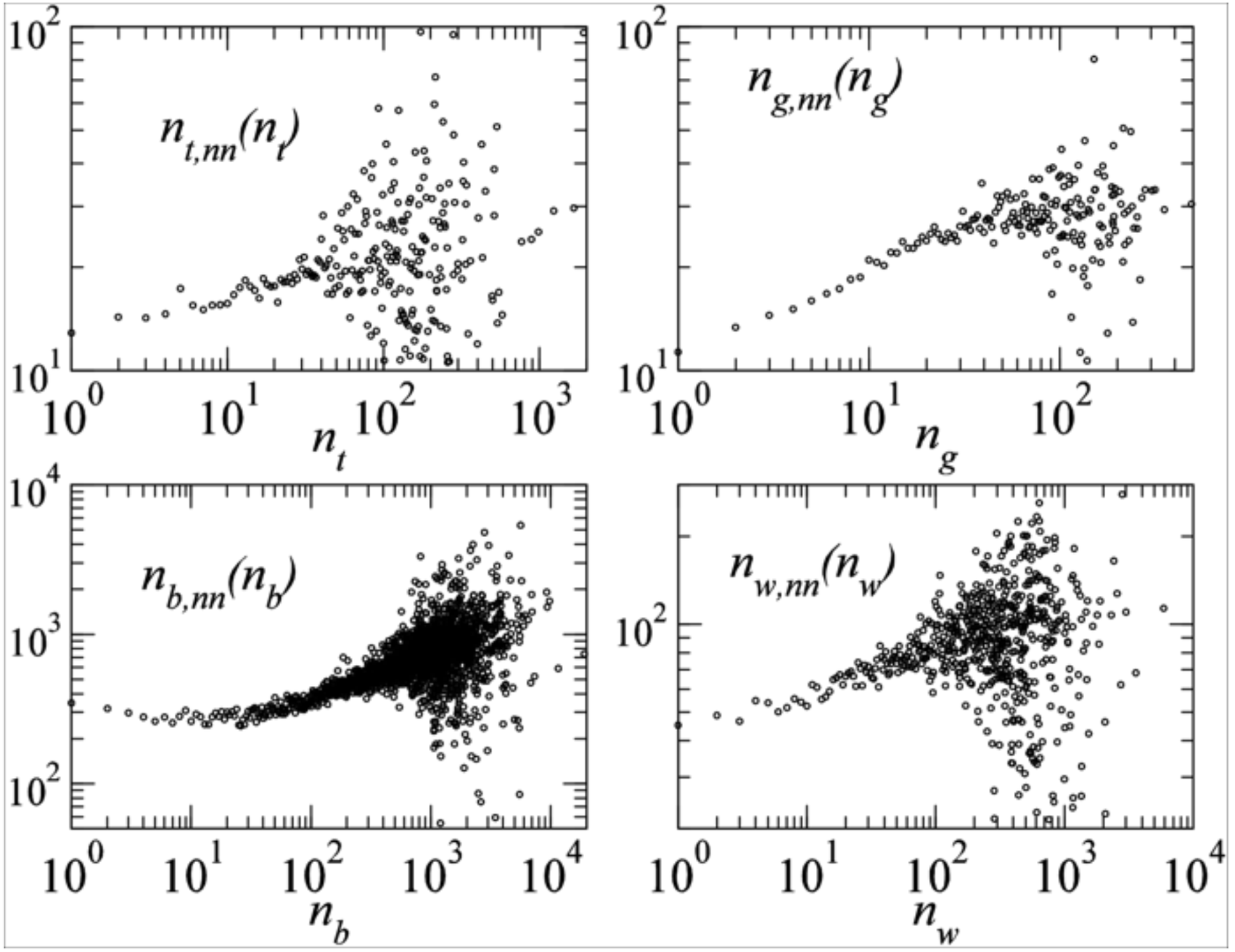
Social networking

Mixing patterns

average activity of nearest neighbors as a function of own activity

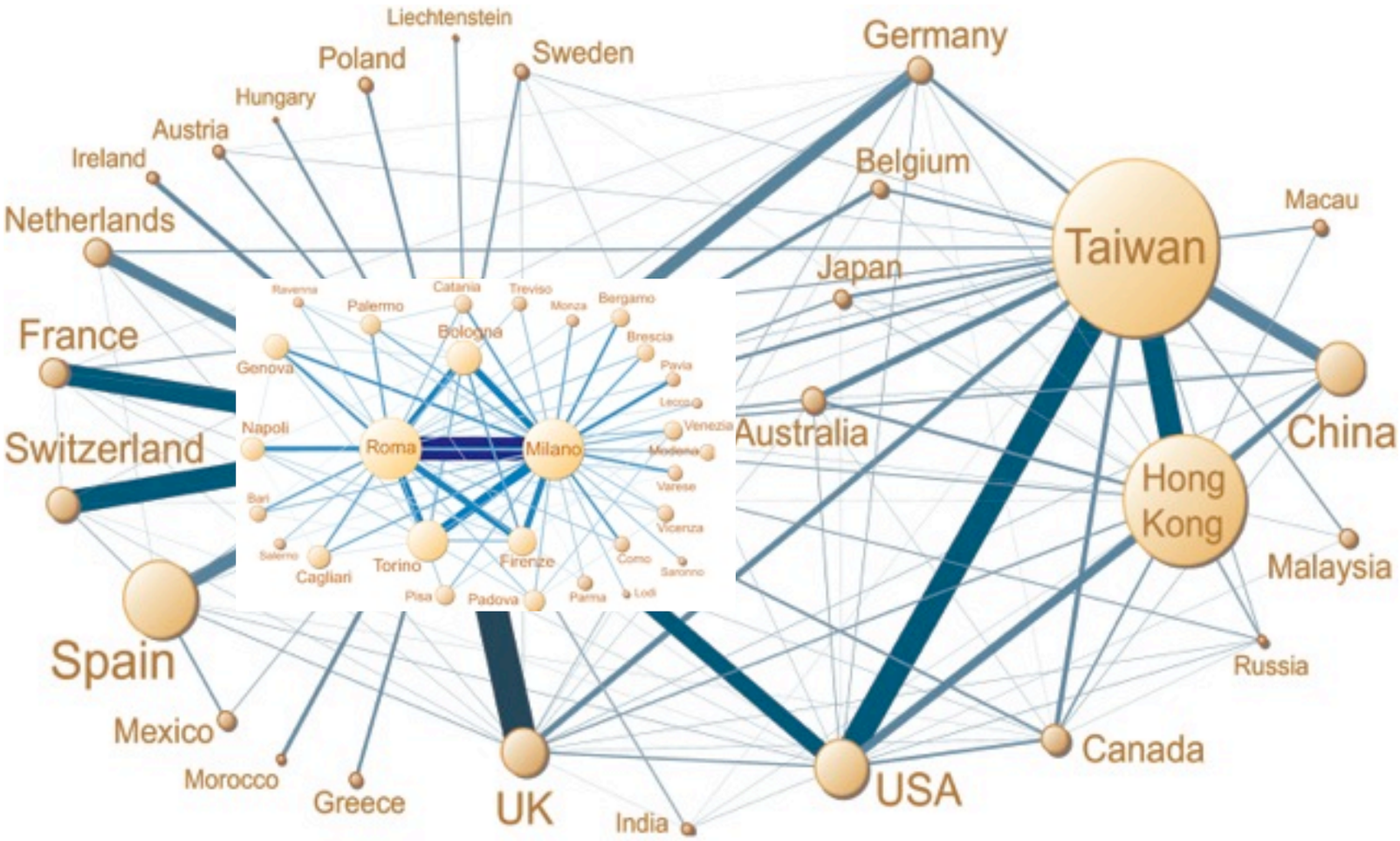
The more a user is active, the more his/her neighbours are active

Assortative mixing, usual in social networks



Homophily?

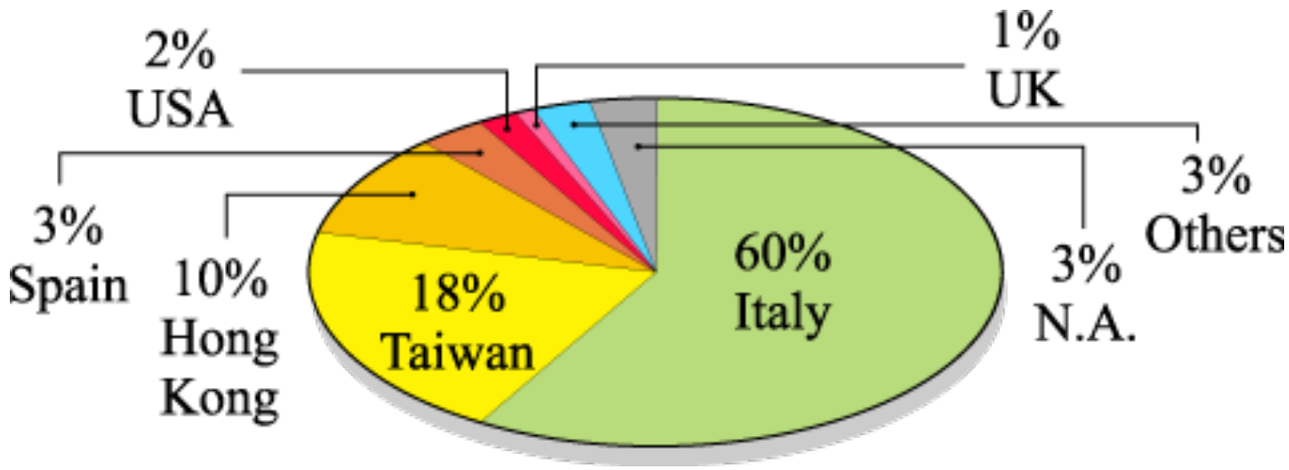
Is geography important?



Geography

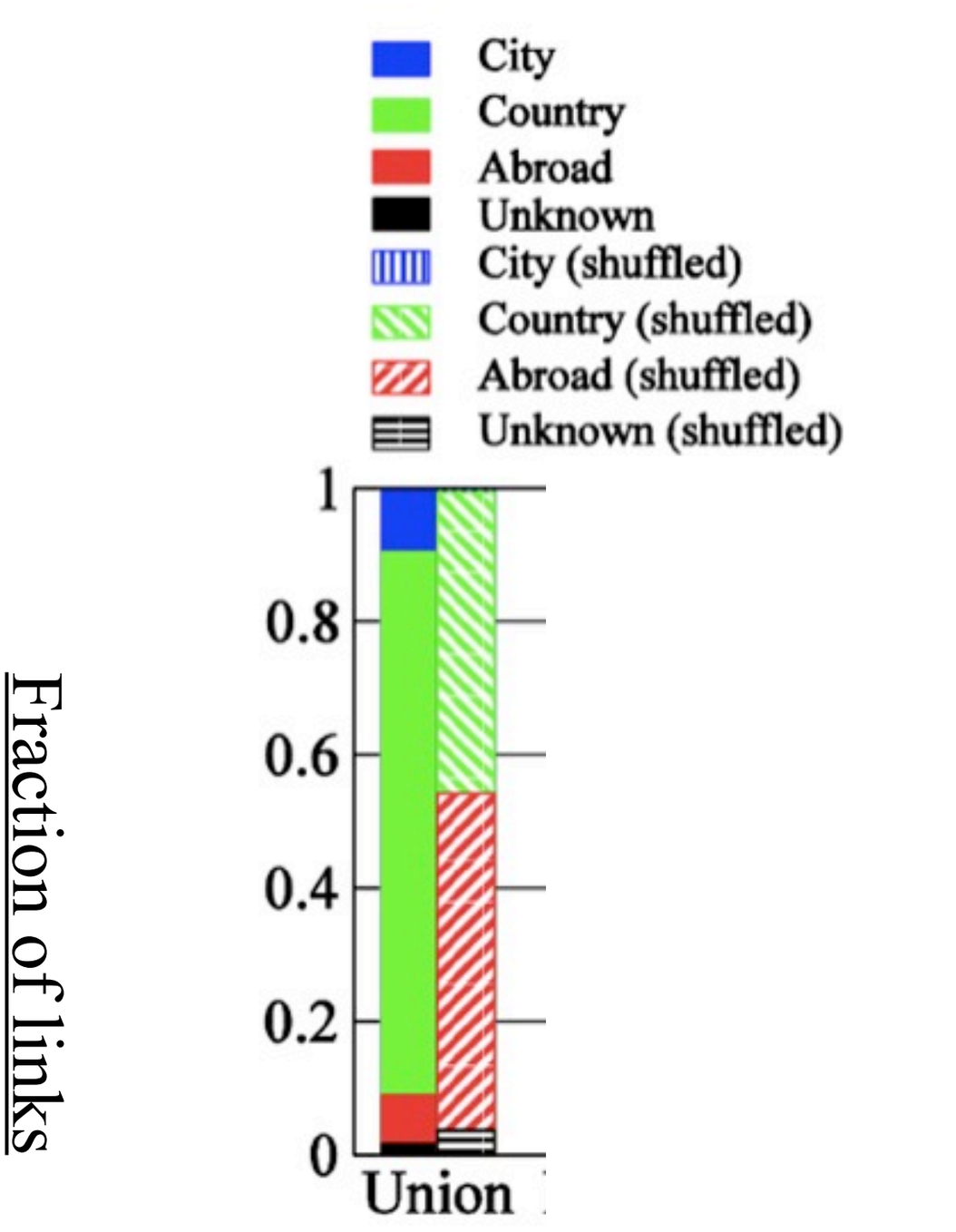
Dataset peculiarities

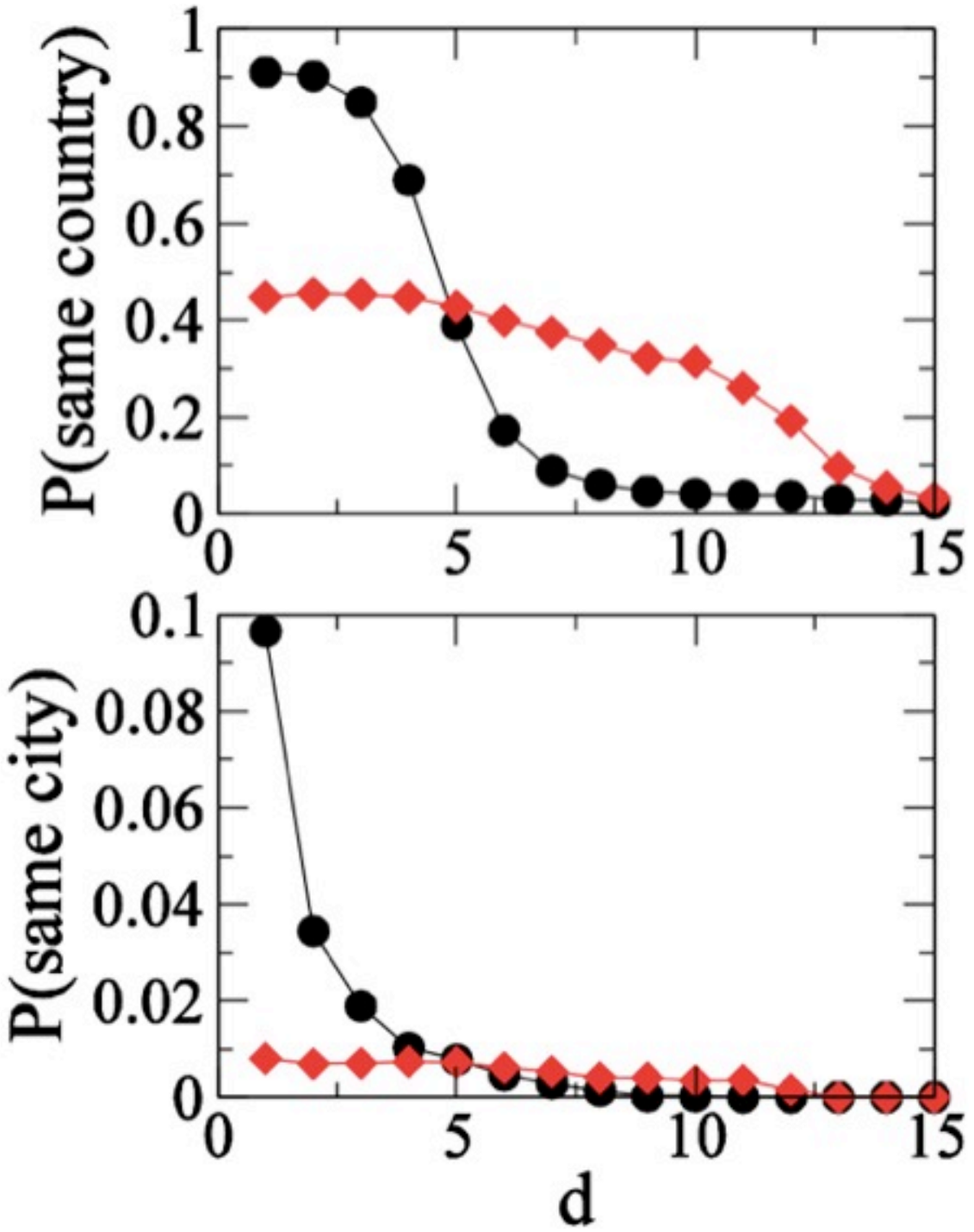
- Many users specify their home country (97%) or town (38%)
- Particular geographical distribution



=> need to compare with null model

Geography


City
Country
Abroad
Unknown
City (shuffled)
Country (shuffled)
Abroad (shuffled)
Unknown (shuffled)
1
0.8
0.6
0.4
0.2
0
Union
Fraction of links


P(same country)
1
0.8
0.6
0.4
0.2
0
0
5
10
15
P(same city)
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
d
Distance on network

Topical alignment of users' profiles?

- Measure: common books, tag usage patterns, shared groups
- global?
- local? (between neighbors on the social network)
- dependence on distance on the social network?

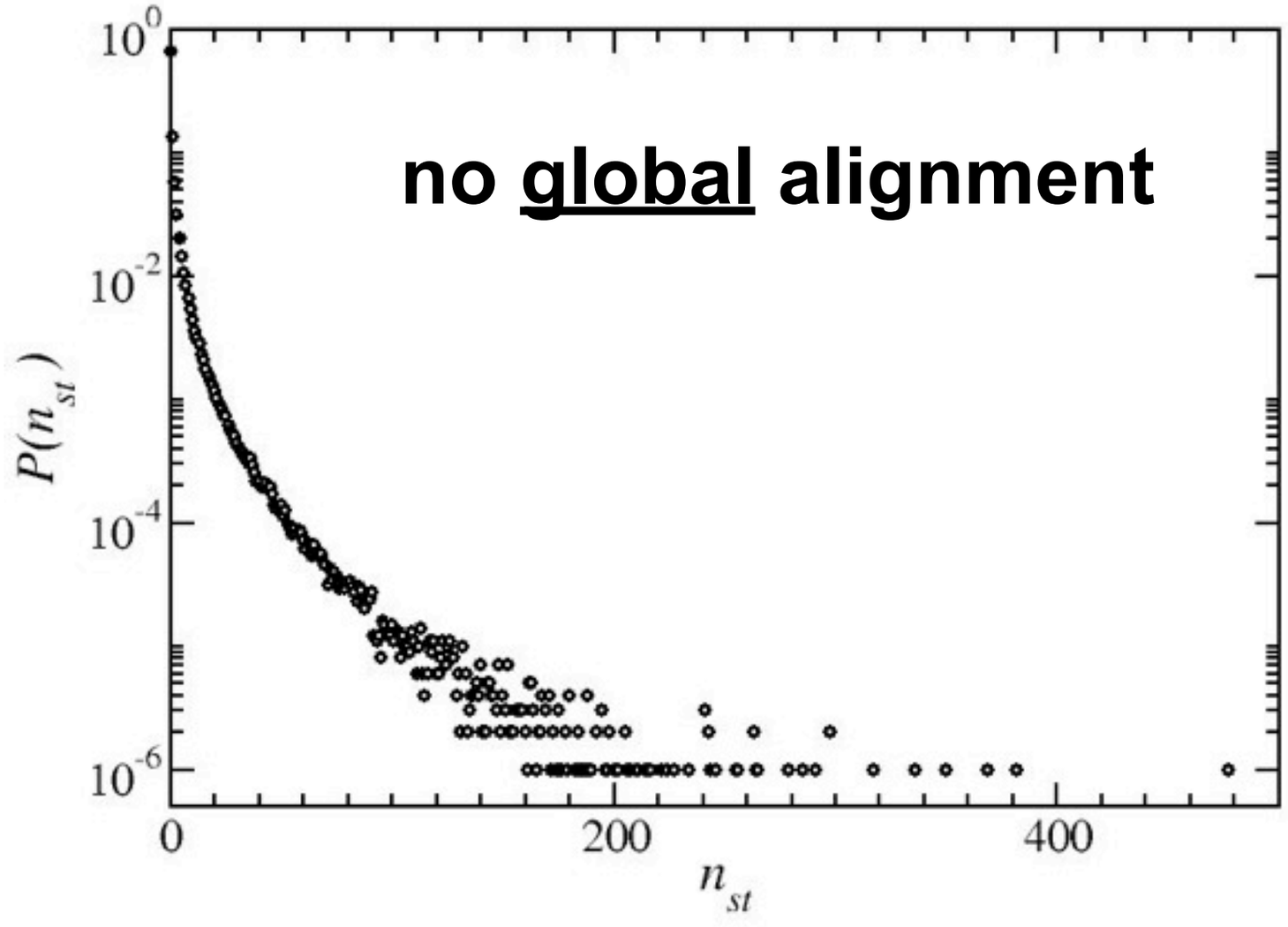
measures of alignment:

- # common books of two users
- # distinct tags shared between two users
- # groups shared
- similarity measures (normalized)

Alignment of users' profiles

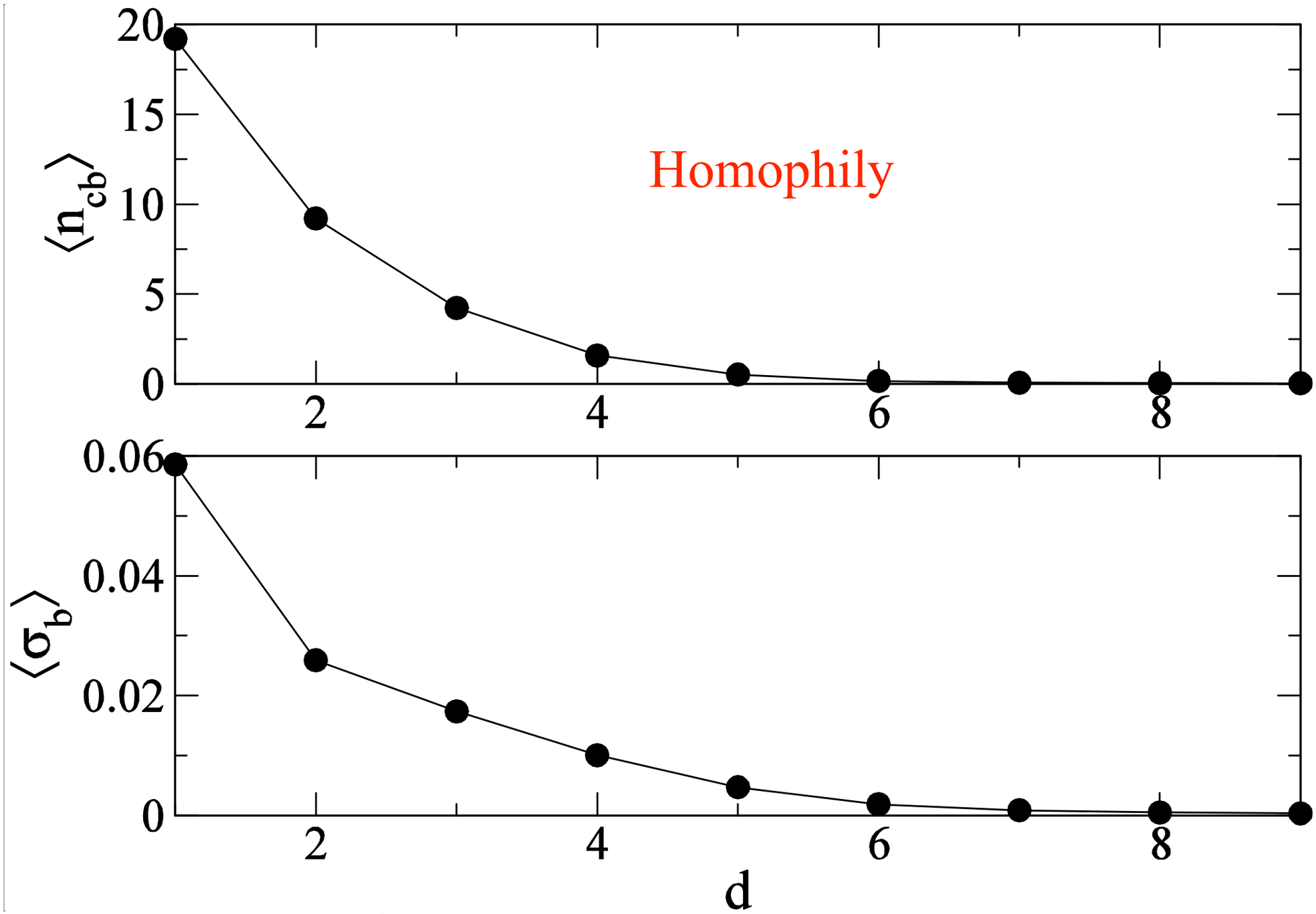
random pairs of users:

- no alignment (small average # of common tags/groups/books)
- most likely case: no shared tags/groups/books



no <u>global</u> alignment

Alignment along the network



Average number of common books of two users

Average normalized similarity measure between two users

Distance between users on social network

Real effect, or due to assortativity?

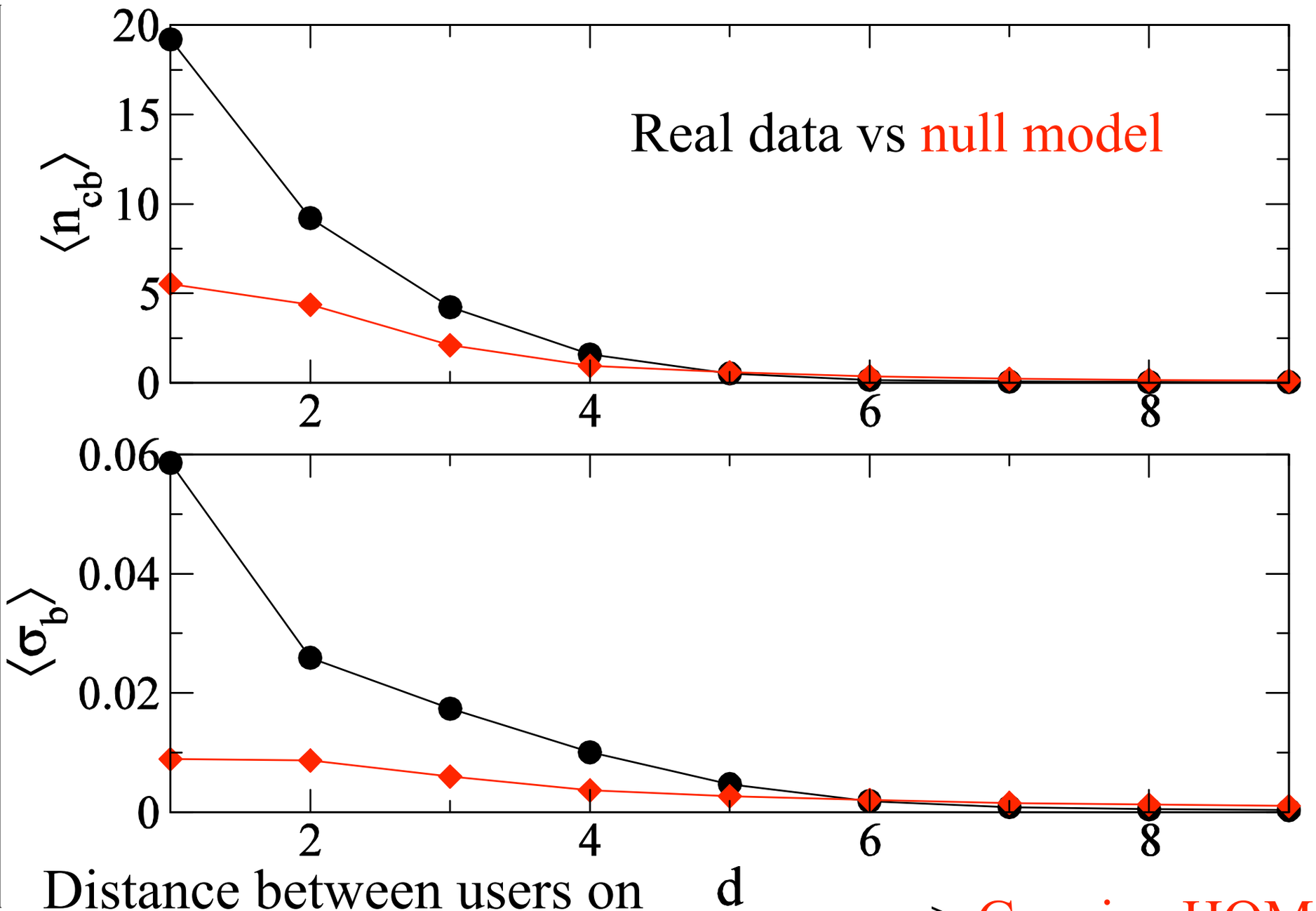
Lexical/topical alignment: building a null model

- conserve the structure of the social graph

- keep unchanged the statistical properties
  - tag frequencies
  - activity of users
  - correlations between activities
  - mixing patterns

- but: **remove assortativity-related alignment**

Alignment along the network



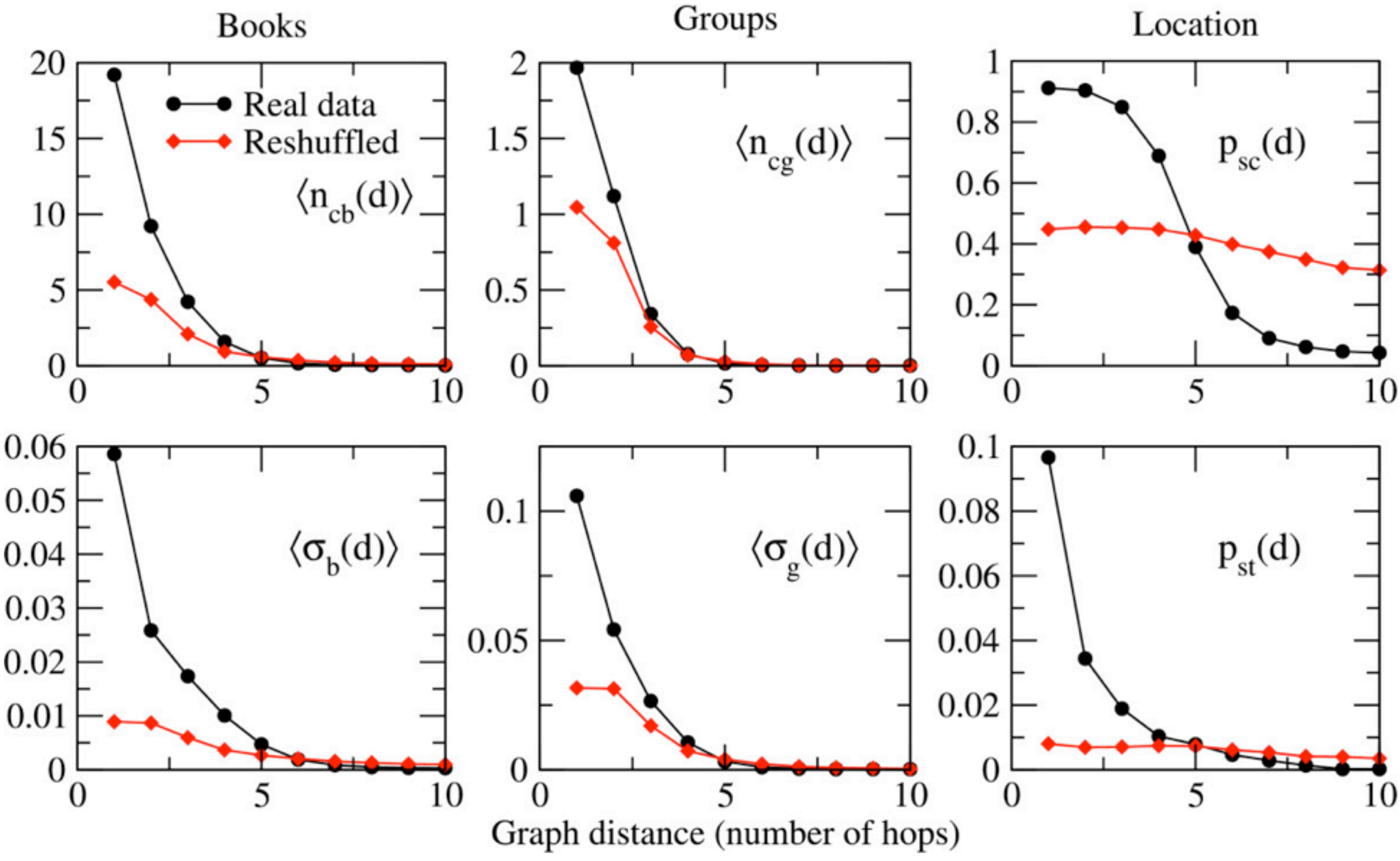
Average number of common books

Average normalized similarity measure

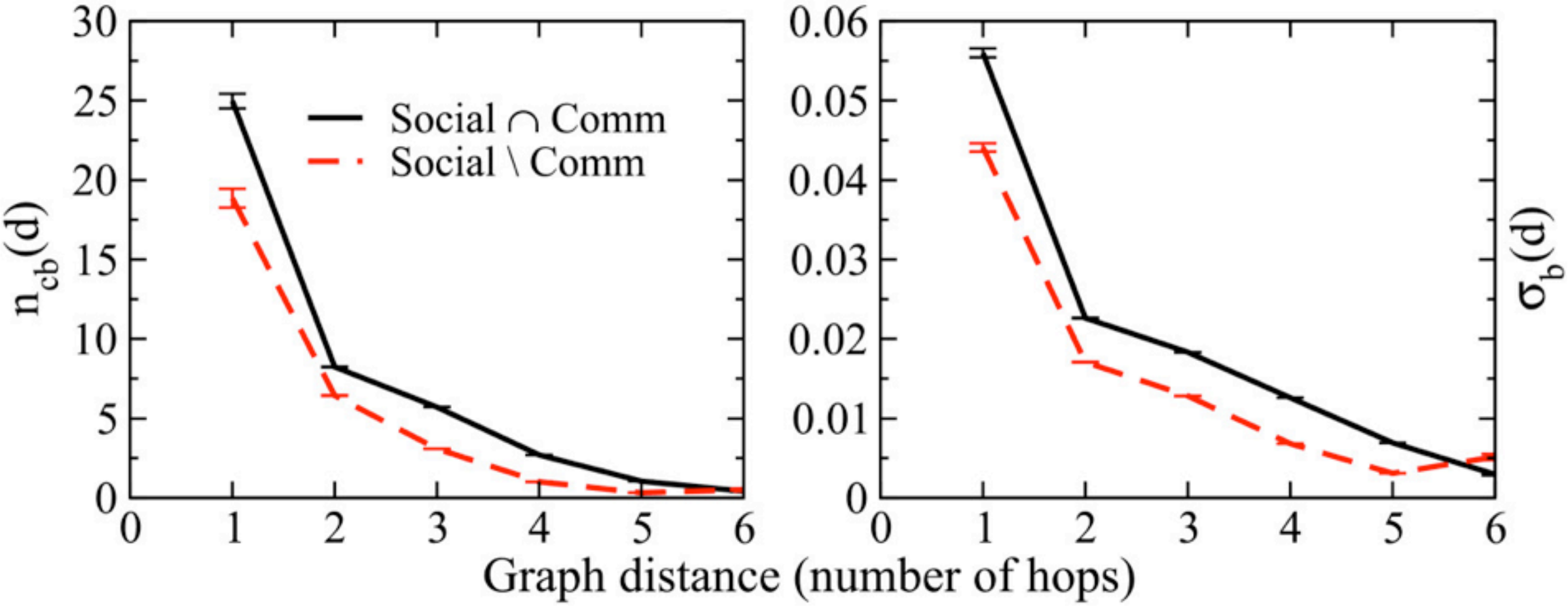
Distance between users on social network

=> Genuine HOMOPHILY effect, not only due to assortativity w.r.t. amount of activity

Alignment along the network



Alignment along the network and the communication network



Stronger effect along the communication network

Homophily:
Selection or influence?

Dynamics

Successive snapshots at intervals of 15 days

- New nodes
- New links from new to old nodes

Every 2 weeks:

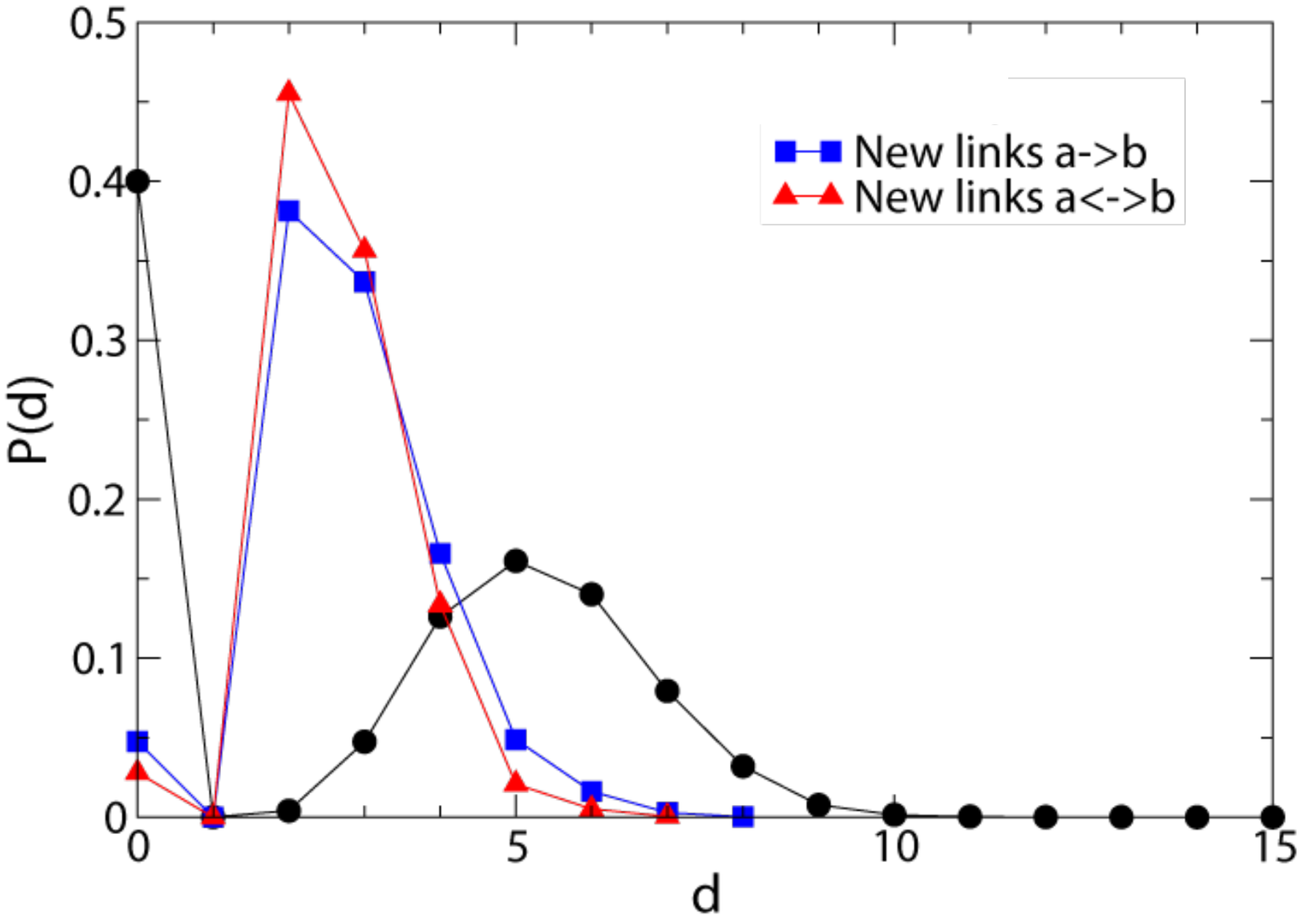
– 2000 to 3000 new users
– 20000 to 30000 new links

However: all statistical properties remain stationary

- New links between old nodes
- Evolution of users' profiles

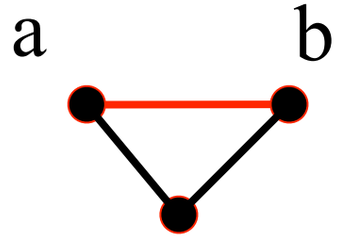
Measure: homophily because of
•Selection?
•Influence?

Dynamics: new links



Distance between a and b on social network before creation of link (a,b)

Triangle closure (many new links between users who were at distance 2)

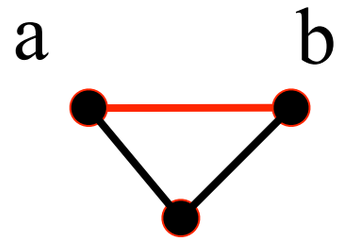


hence many new links between users more similar than random

Dynamics: selection or influence?

	$<n_{cb}>$	σ_b	$<n_{cg}>$	σ_g
All a,b such that $d_{ab}=2$	9.5 (0.2)	0.02	1.12 (0.61)	0.05
Simple closure (a->b with $d_{ab}=2$)	18.2 (0.09)	0.04	1.81 (0.45)	0.1
Double closure (a <-> b with $d_{ab}=2$)	23.4 (0.03)	0.05	2.2 (0.36)	0.12

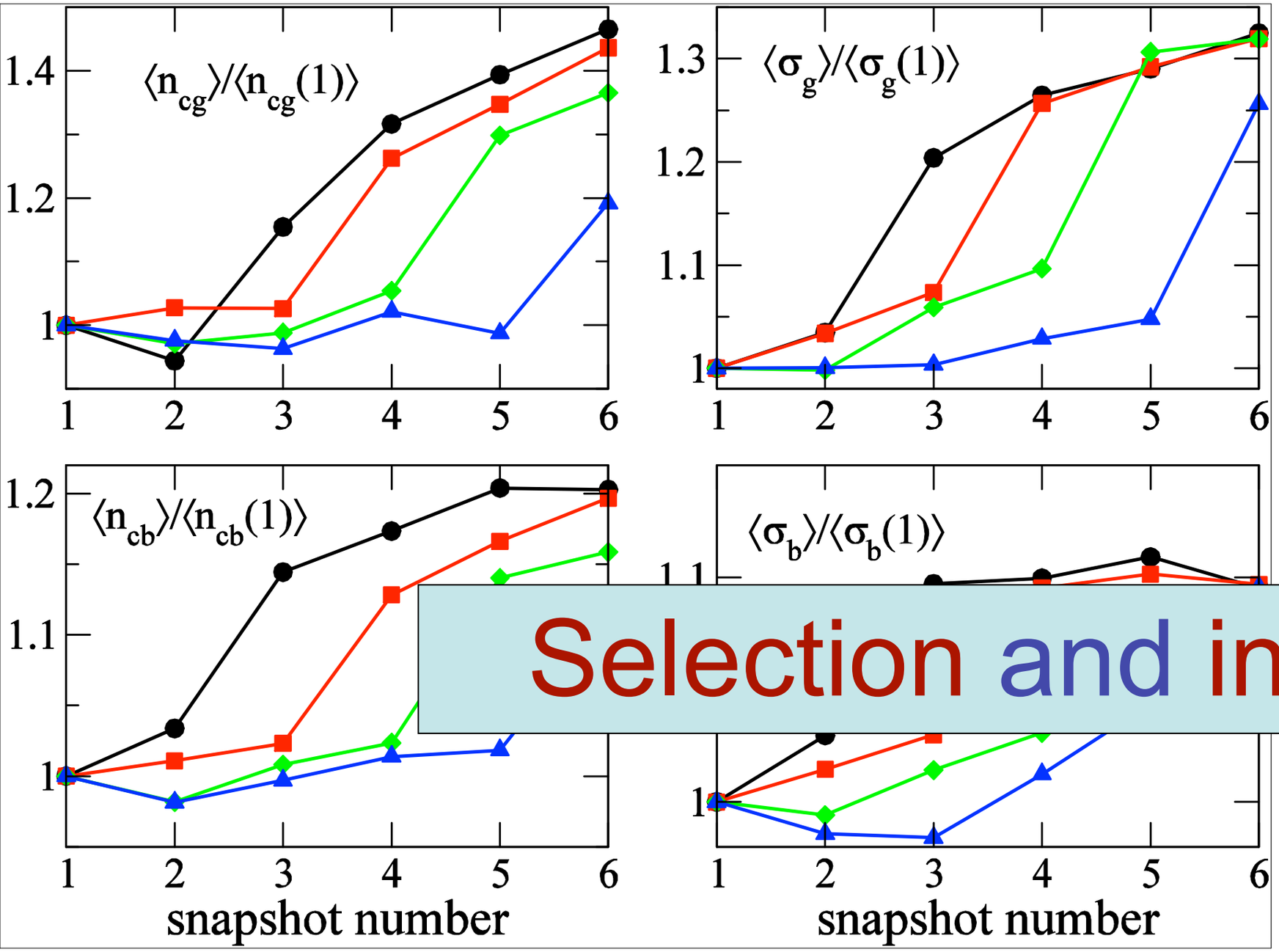
New links between already present users



Selection

Larger average similarity at t for pairs which become linked between t and t+1
(and smaller proba to have 0 similarity)

Dynamics: selection or influence?

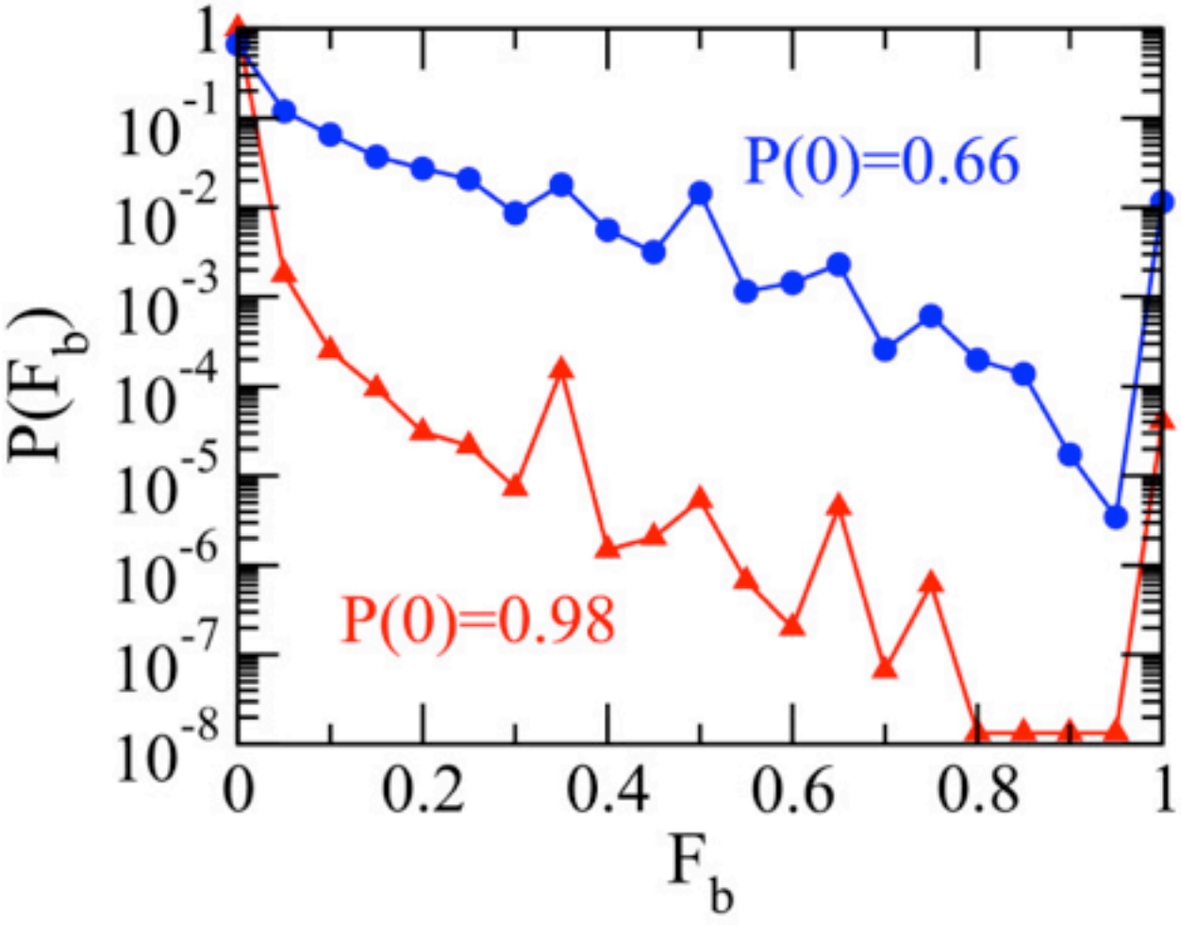


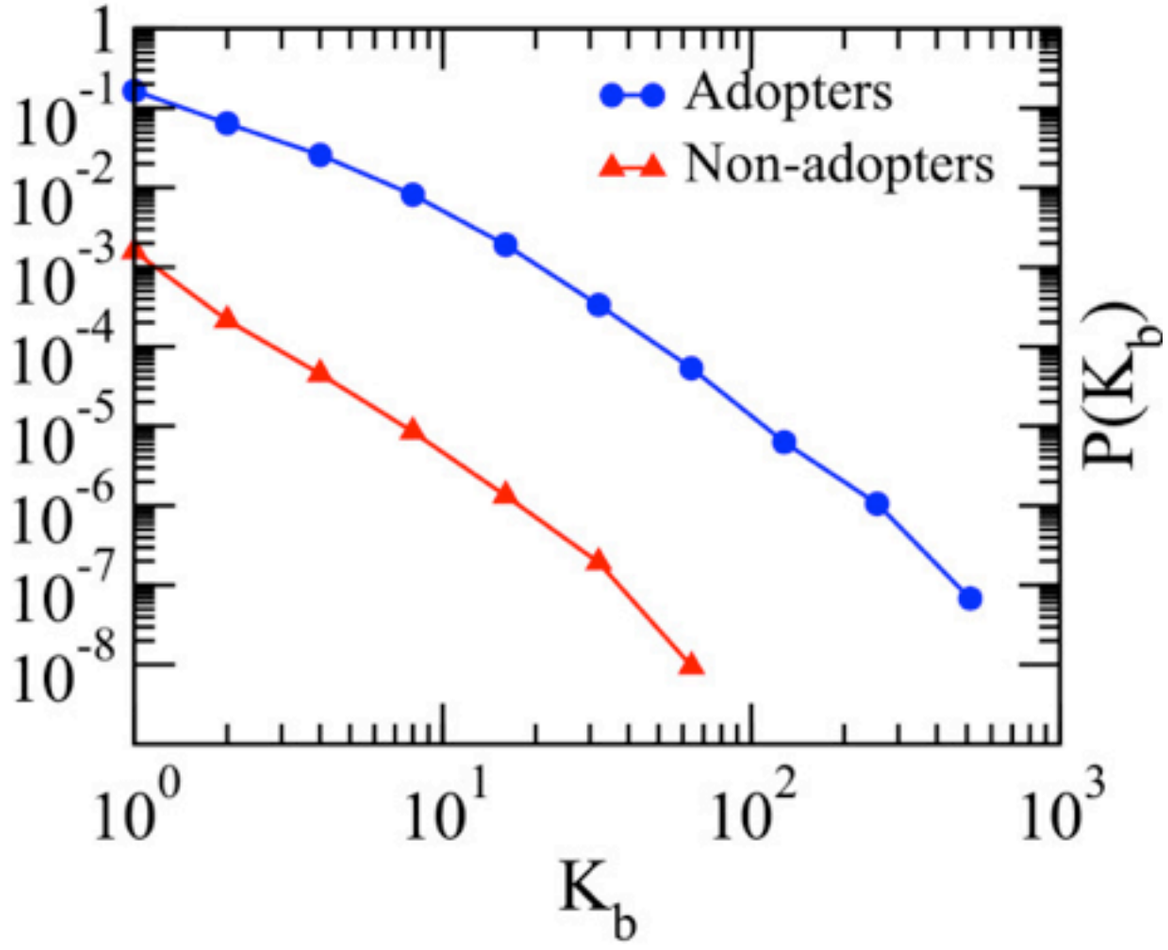
Evolution of similarity ***before and after*** link creation

Selection and influence

Bi-directional causality relation between similarity and link creation

Influence in the adoption of books





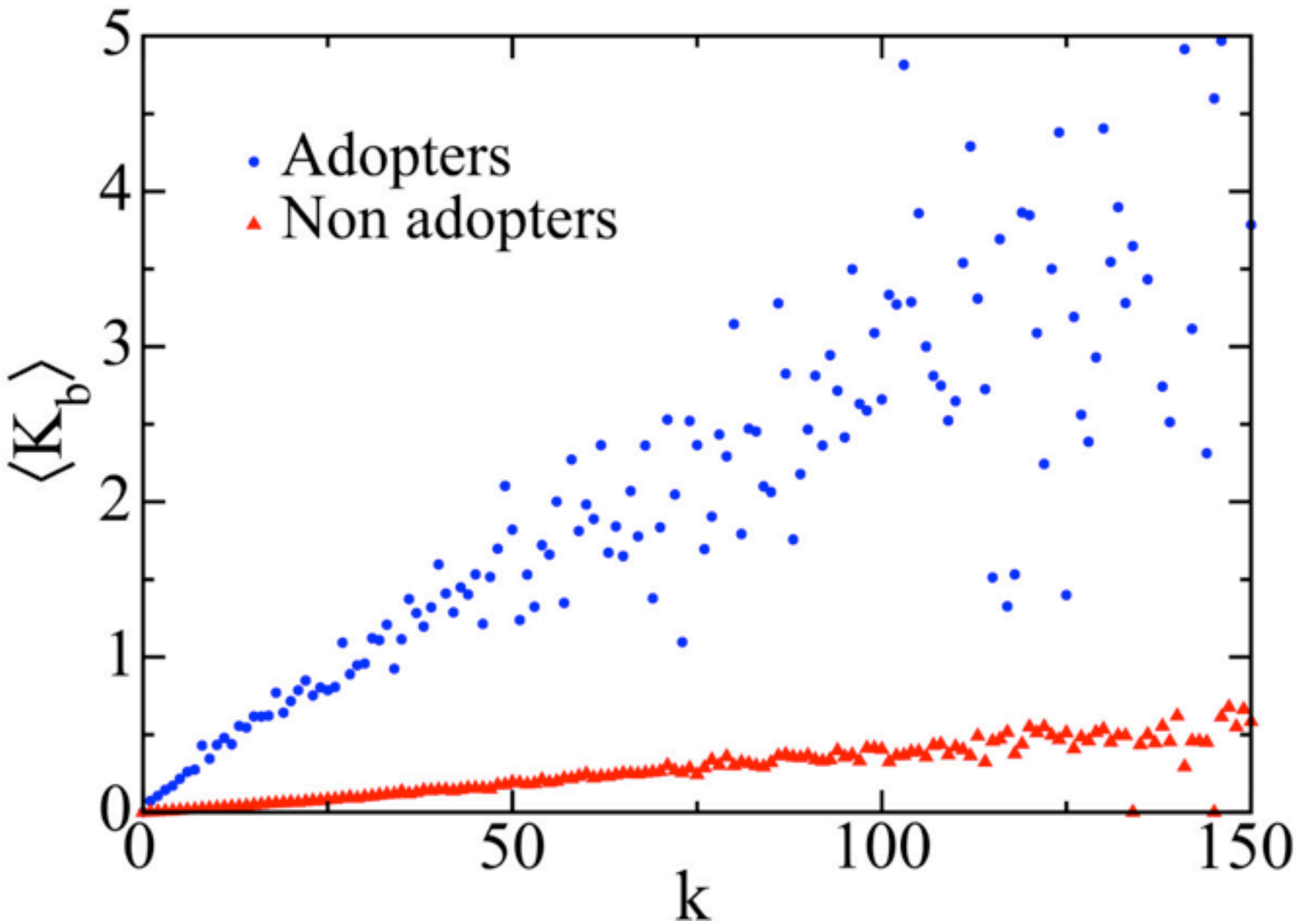
Distribution of the fraction/number of neighbours who have read book b at time t, for '**adopters**' of book b between t and t+1, i.e., users:

-without book b at time t

-with book b at time t+1

and for **non-adopters** (users without the book at both t and t+1)

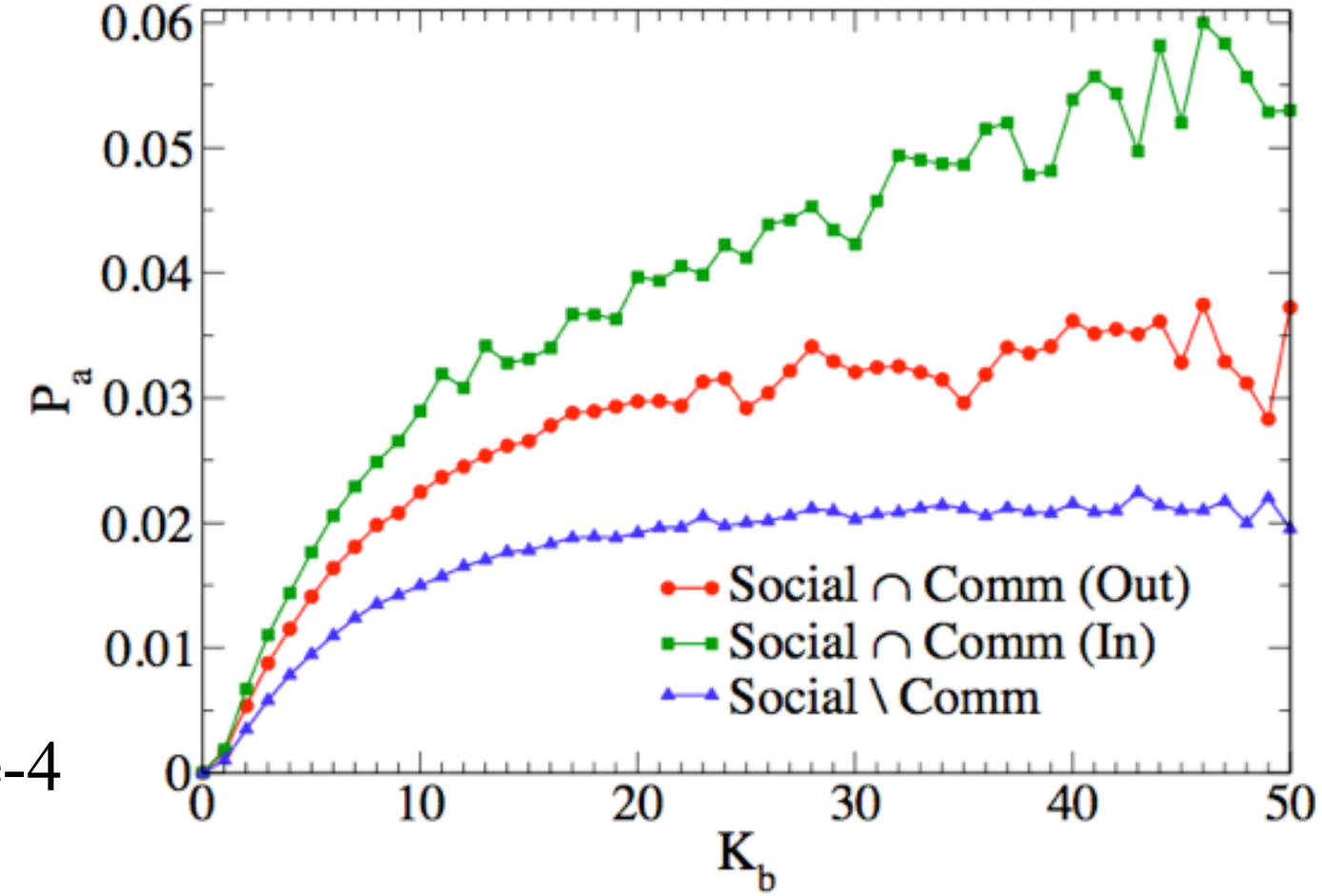
Influence in the adoption of books



Average ***number of neighbours who have read book b at time t,*** vs user's degree, for '**adopters**' and '**non-adopters**' of book b between t and t+1

Adopters have been more 'exposed' to the book than non-adopters

Influence in the adoption of books



P(0)~1e-4

Probability to adopt a book vs number of neighbours having read this book:
-very small P(0), fast increase
-saturation at large K_b (diminishing returns)
-effect of communication
-stronger effect of in-communication, i.e., **potential recommendations**, vs out-communication

What about predicting new links?

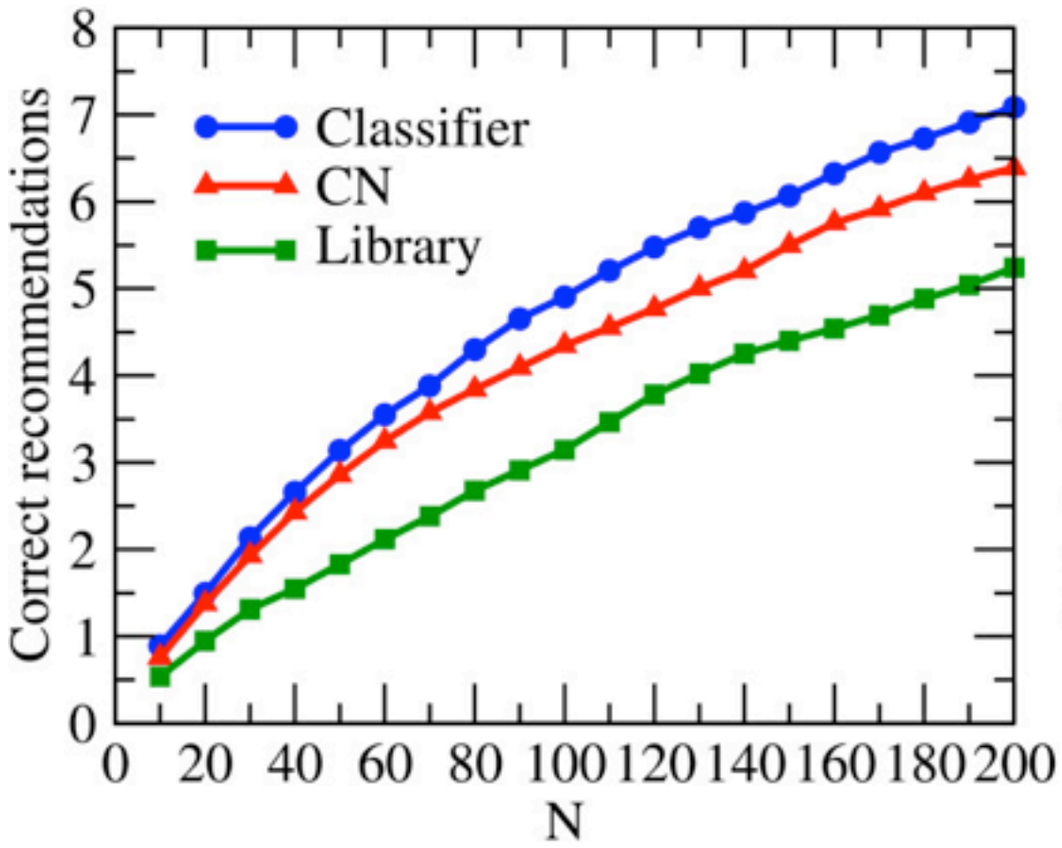
Creation of new links influenced by:

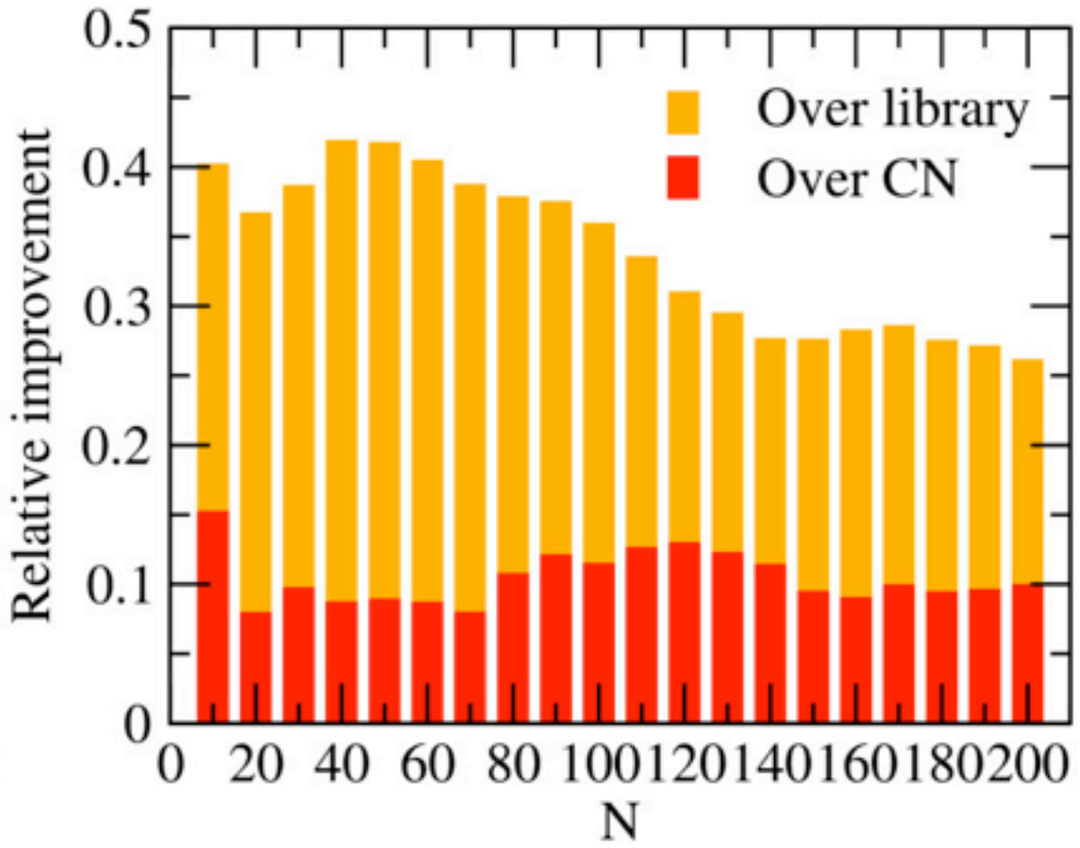
-Proximity on the graph (friends of friends,...)
-Strength of communication
-Selection mechanism (profile similarity)

=> use topological + profile features, train classifier to select most predictive features, create recommendation system

What about predicting new links?

Test set of users making at least 20 new ties at distance 2





Precision at N for the recommendation made with the classifier combining all the relevant features and for two unsupervised baselines (common neighbors and library similarity). Right: Relative improvement on the classifier-based approach over the baselines.

=> not so good results, due to low activity, small fraction of new links at d=2, directionality of links, topicality of social network?

Dynamical real-world social/behavioral networks

Mining social interaction networks

– Bluetooth, wifi (O' Neill et al 2006; Scherrer et al 2008; Eagle, Pentland 2009)
– MIT Reality mining project (sociometric badges)
– MOSAR european project (hospitals)

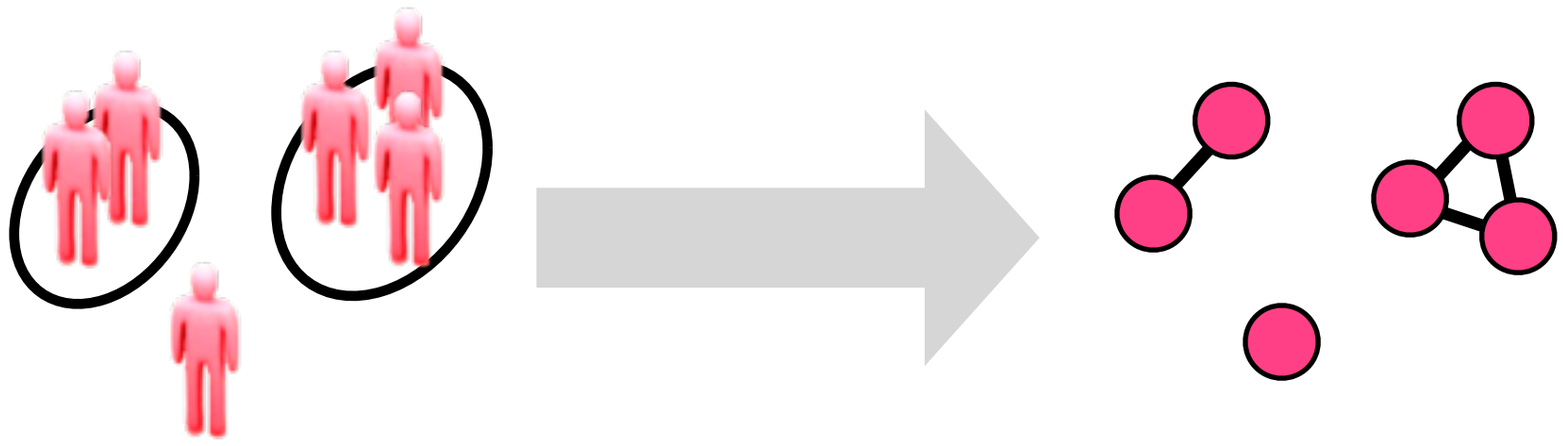
SocioPatterns: large-scale time-resolved data on face-to-face proximity across a variety of contexts

(Ciro Cattuto's talk on monday)

The SocioPatterns collaboration

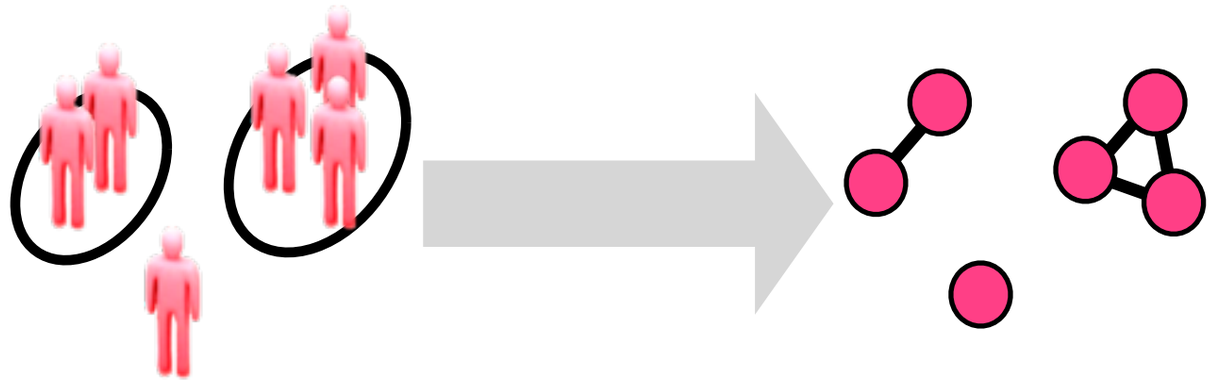
SocioPatterns.org

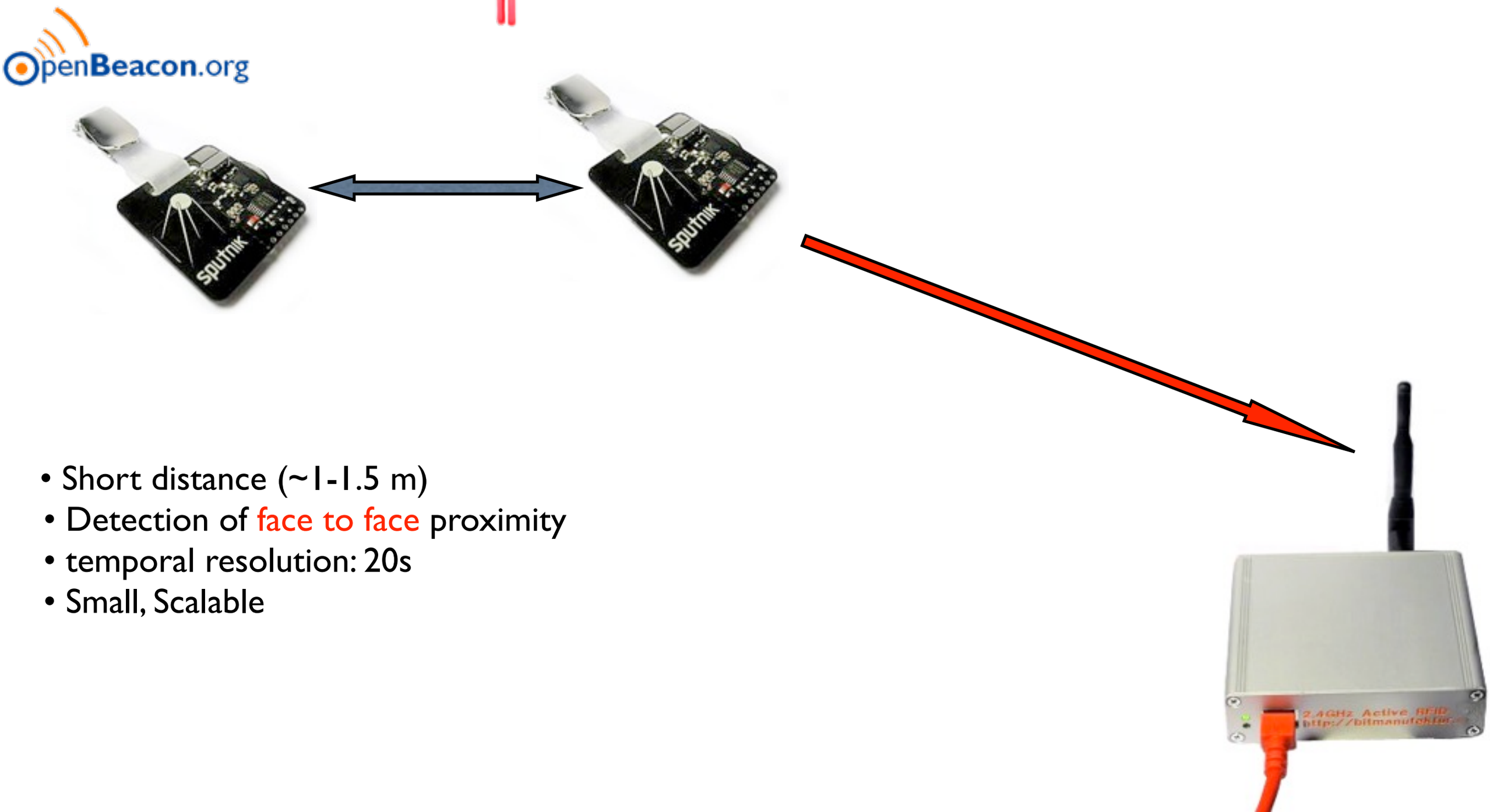
what are the statistical and dynamical properties of the networks of contact and co-presence of people in social interaction?



fine-grained **spatial** (~ m) and **temporal** (<min) resolution

Infrastructure: active RFID badges

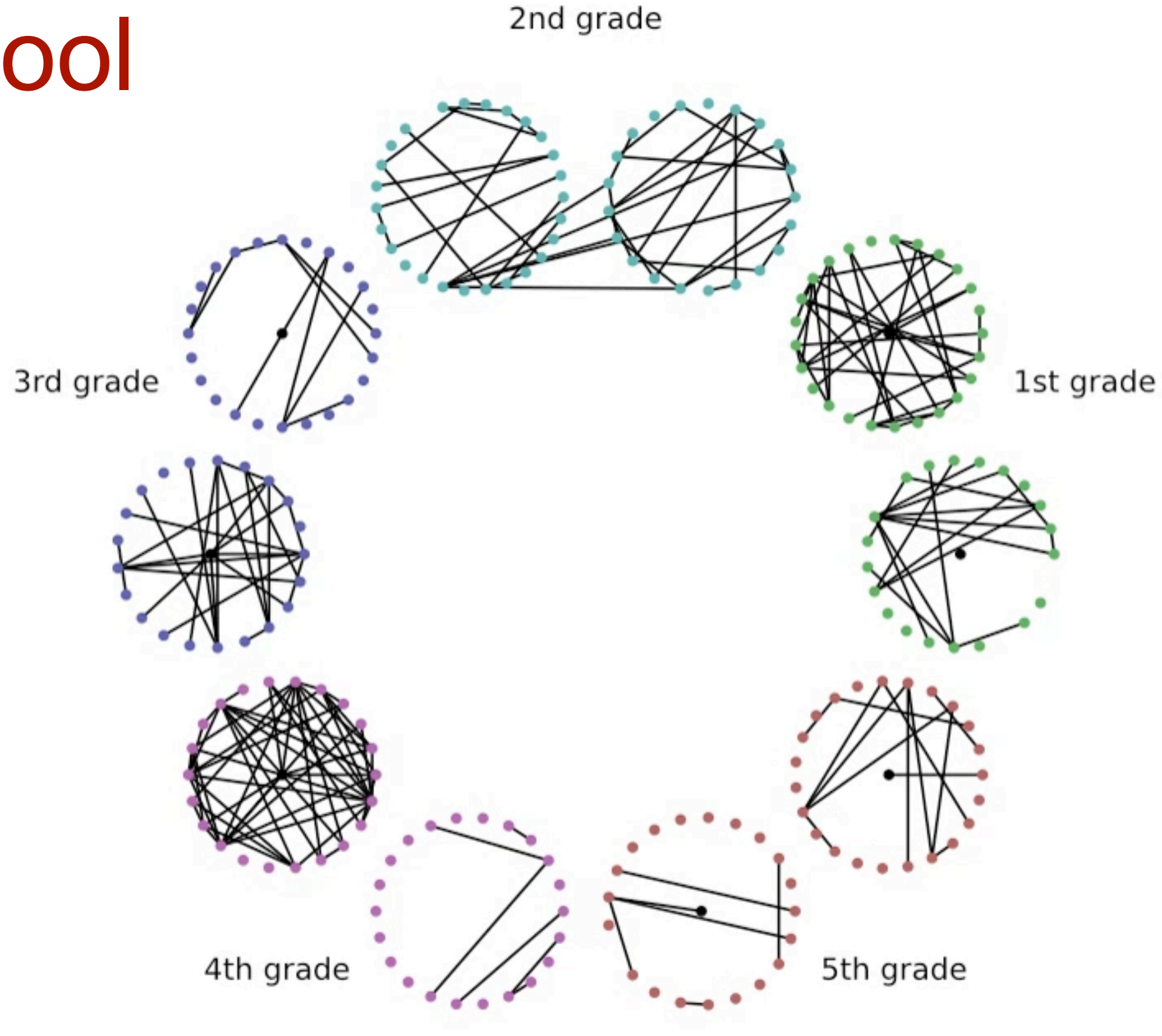




- Short distance (~1-1.5 m)
- Detection of face to face proximity
- temporal resolution: 20s
- Small, Scalable

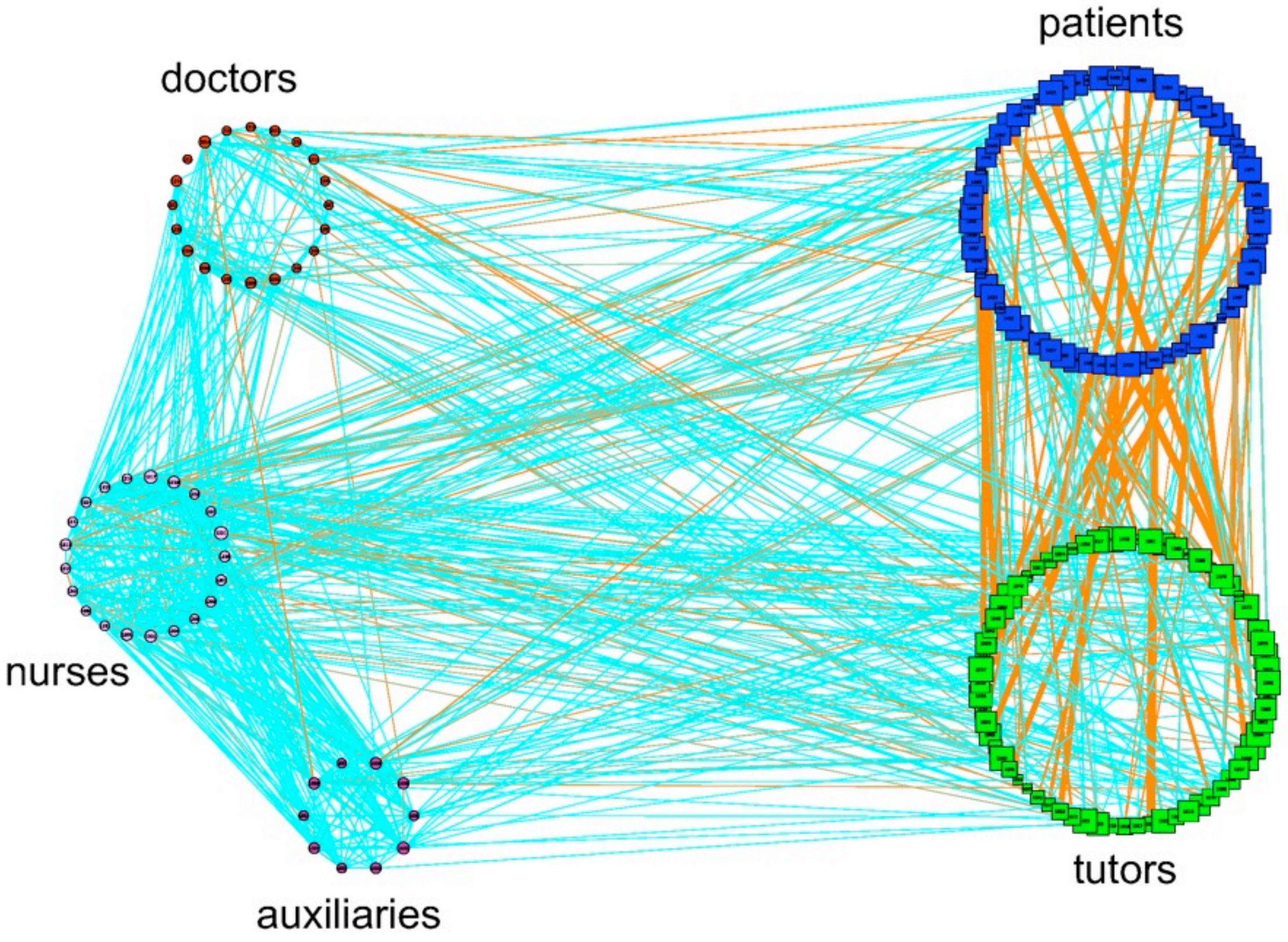
DATE	EVENT	SIZE	DURATION
May 2008	Workshop, Torino, IT	~65	3 days
Jun 2008	ISI offices, Torino, IT	~25	3 weeks
Oct 2008	ISI workshop, Torino, IT	~75	3 days
Dec 2008	Chaos Comm. Congress, Berlin, DE	~600	4 days
Apr–Jul 2009	Science Gallery, Dublin, IE	~30,000	3 months
Jun 2009	ESWC09, Crete, GR	~180	4 days
Jun 2009	SFHH, Nice, FR	~360	2 days
Jul 2009	ACM **and more....**	~120	3 days
Oct 2009	Primary school, Lyon, FR	~250	2 days
Nov 2009	Bambino Gesù Hospital, Rome, IT	~250	10 days
Jun 2010	ESWC10, Crete, GR	~200	4 days
Apr 2010	Practice Mapping, Gijon, ES	~100	10 days
Jul 2010	H-Farm, Treviso, IT	~200	6 weeks
2010, 2012	Hospital, Lyon, FR	~100	10 days
Nov 2011, 2012	High school, Marseilles, FR	~150	1 week

School

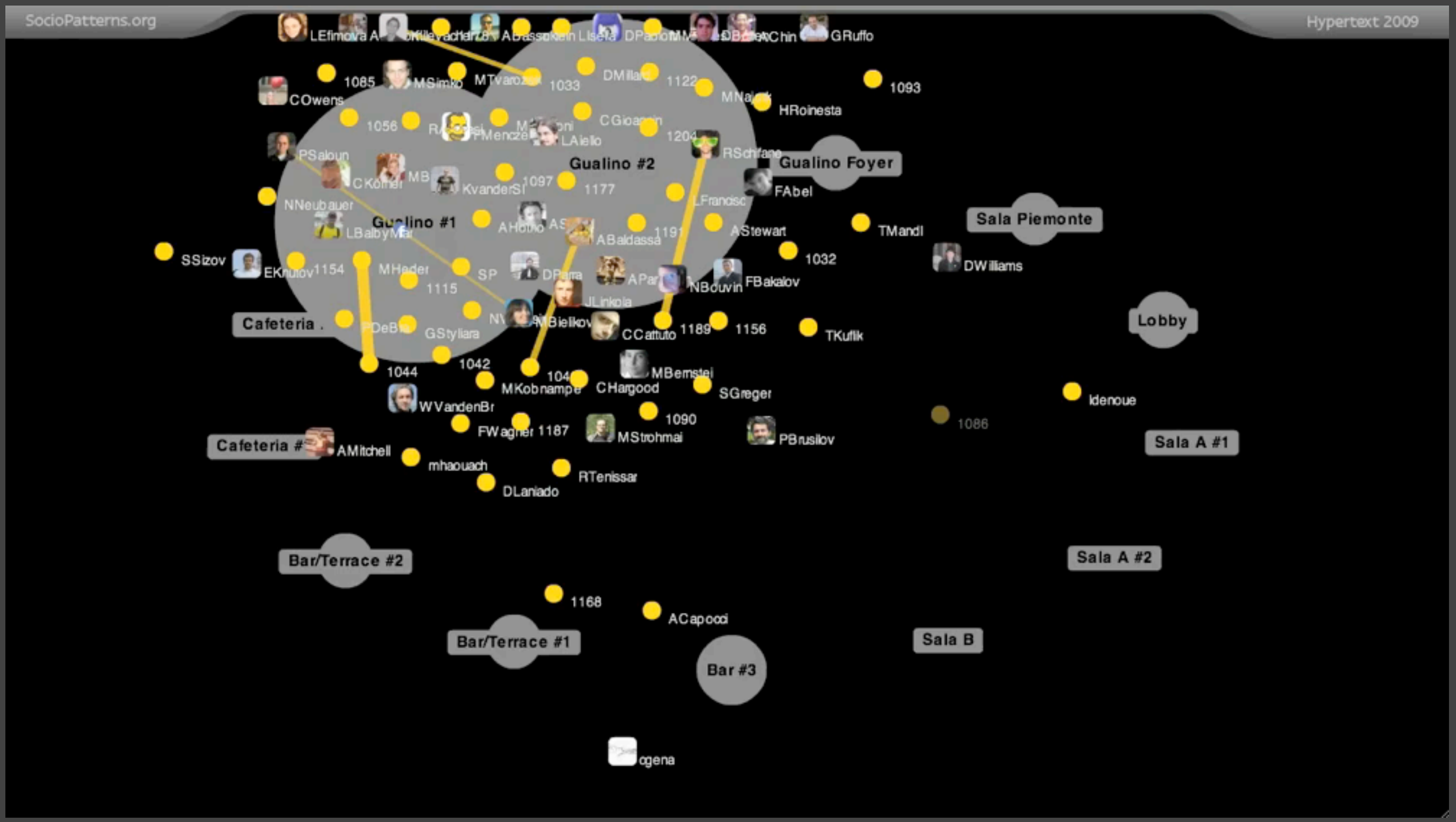


Thu, 11:20- 12:00

Hospital

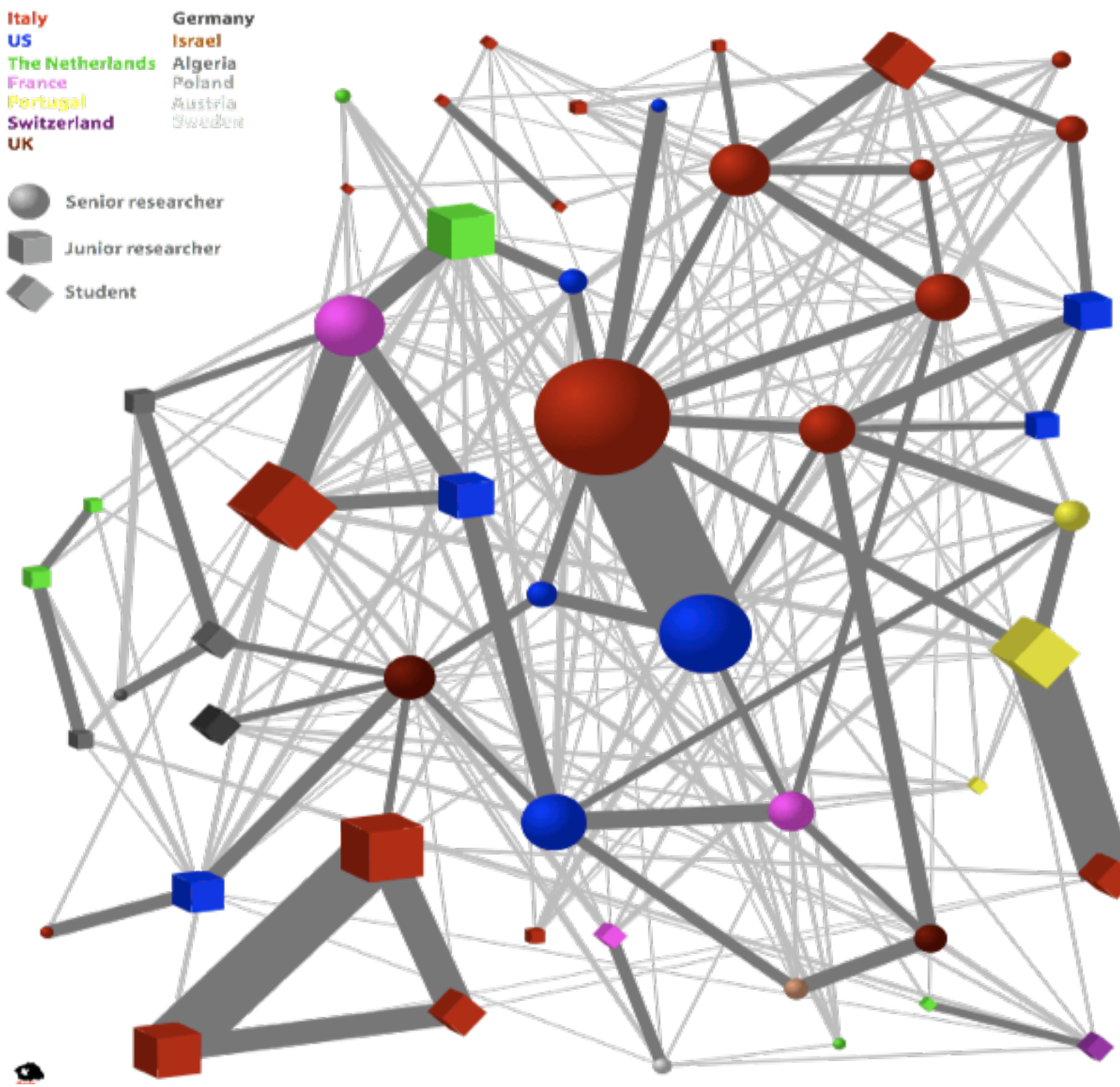


dynamical network of f2f proximity



http://www.vimeo.com/6590604

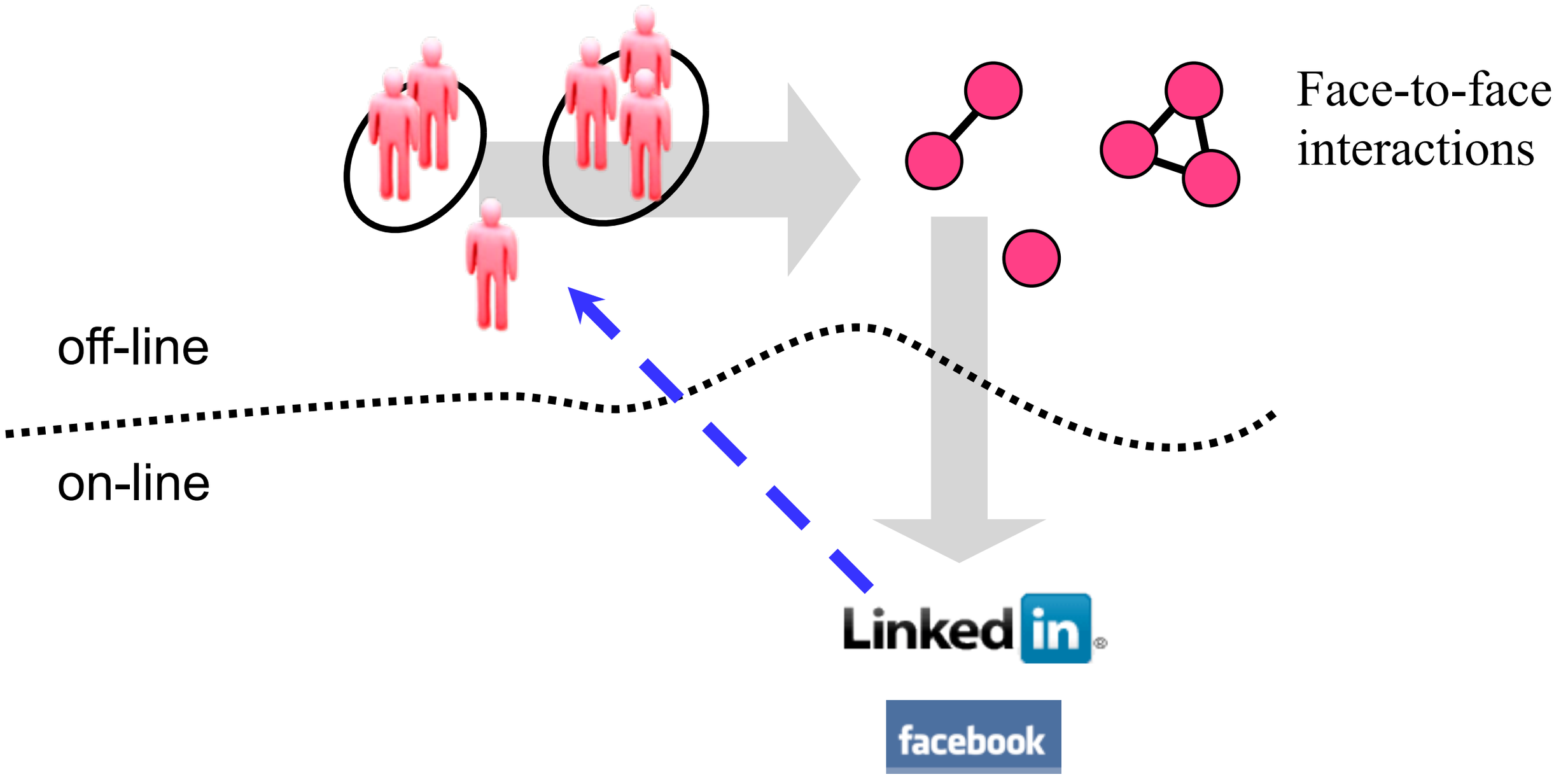
cumulative contact networks



- heterogeneity
- homophily
- clustering
- small diameter

...and many research directions...

On- and off-line social networking



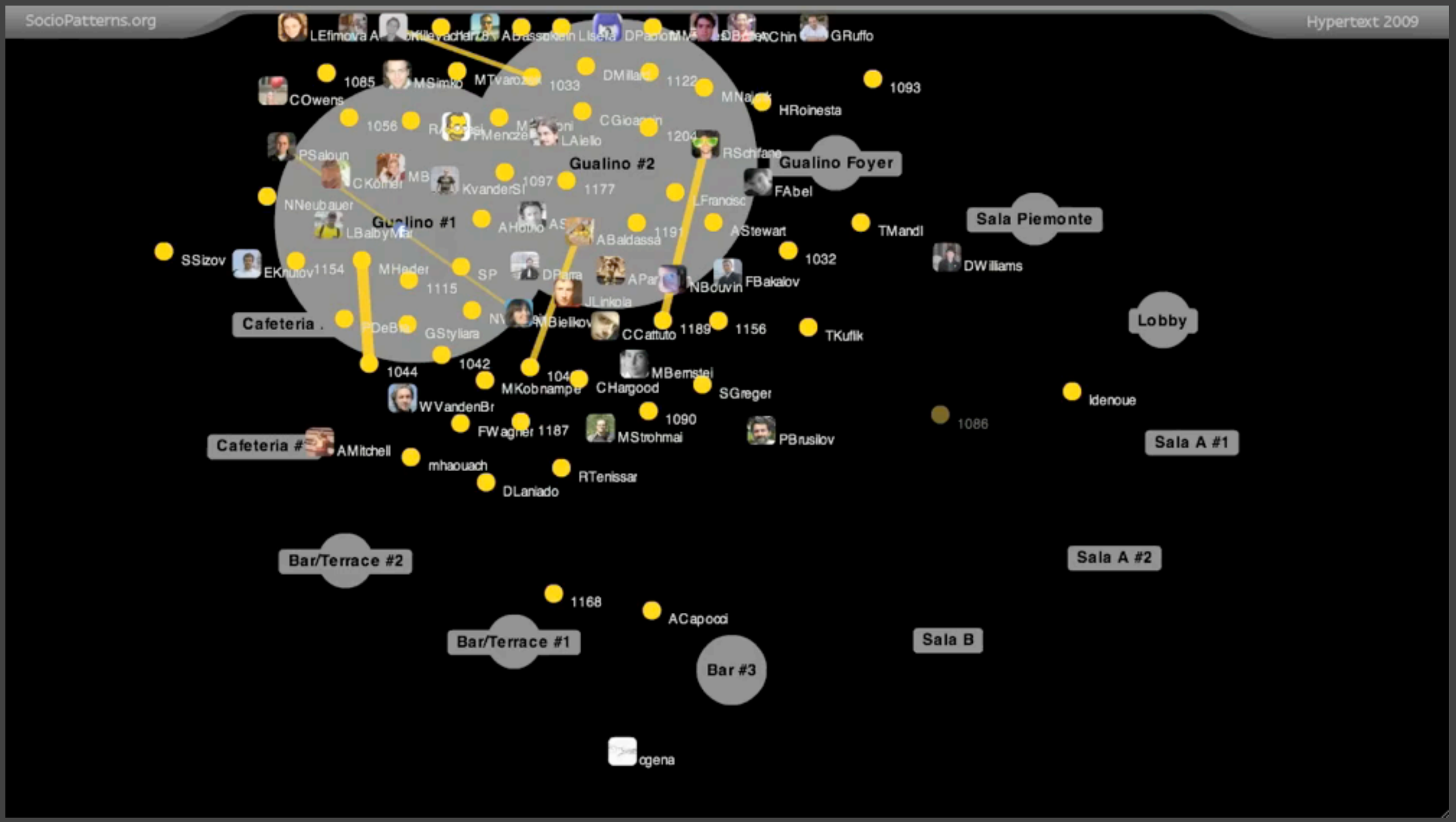
On- and off-line social networking

Live Social Semantics

- Based on the SocioPatterns platform
- Integrates
  - Face-to-face real-time proximity data (RFID badges)
  - Semantic web data
  - Data on online social networks

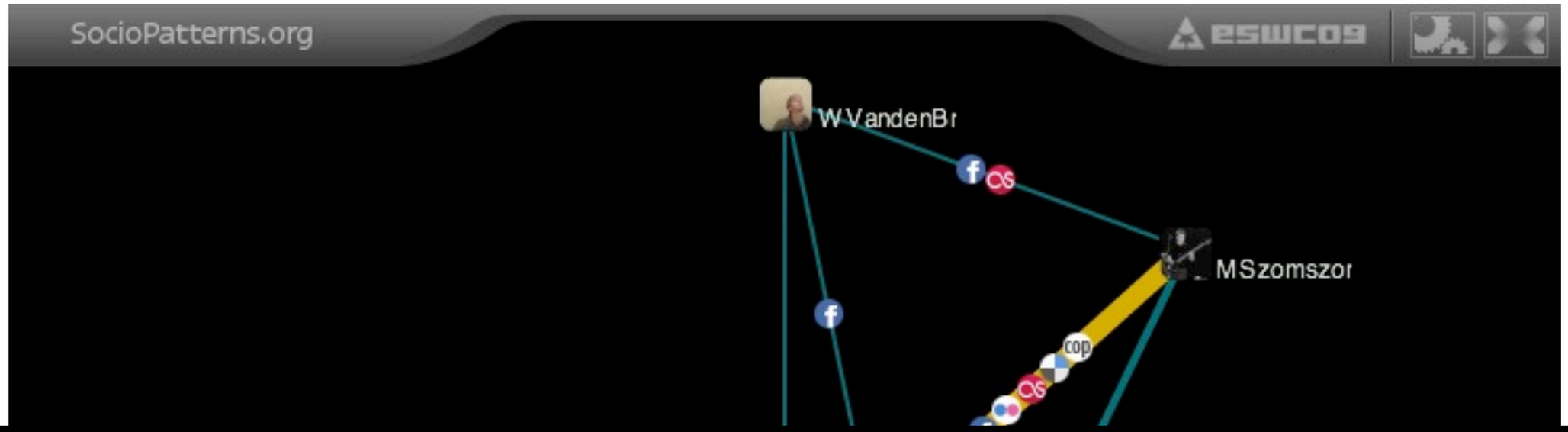
- Deployed at:
  - European Semantic Web Conference 2009, Heraklion, May 31-June 6
  - ACM Hypertext 2009, Torino, June 29-July 2
  - Extended Semantic Web Conference 2010, Heraklion, June 1-4

On- and off-line social networking

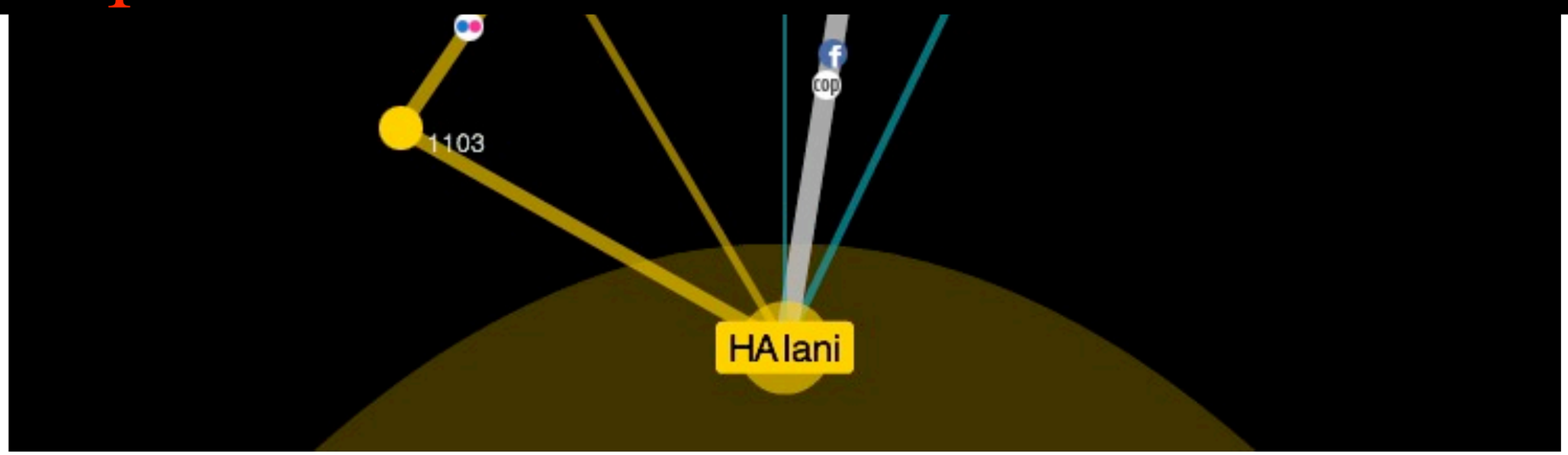


http://www.vimeo.com/6590604

On- and off-line social networking

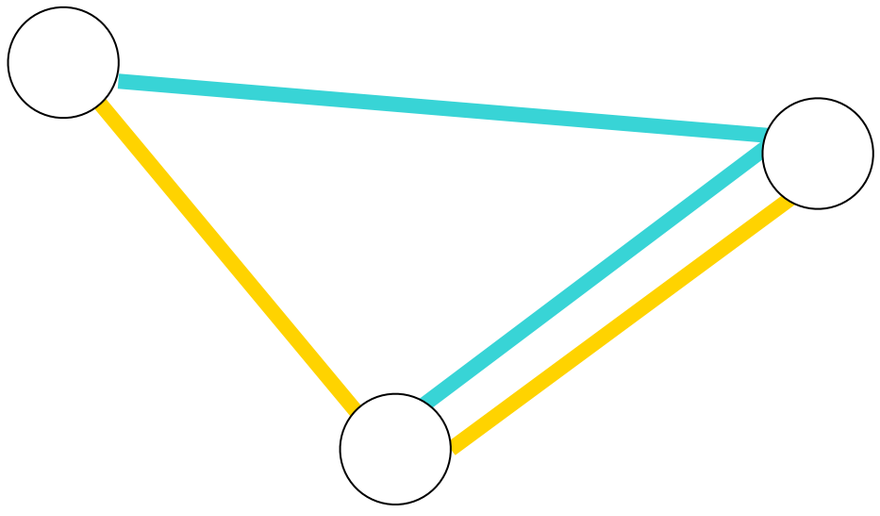


- Direct feedback of information to participants
- Suggestions for new online links
- Comparison of on- and off-line networks



Network comparison

- Set of N nodes (here individuals)
- Two (or more) types of relationships/links between nodes



- Interplay?
- How to measure it?

Network comparison

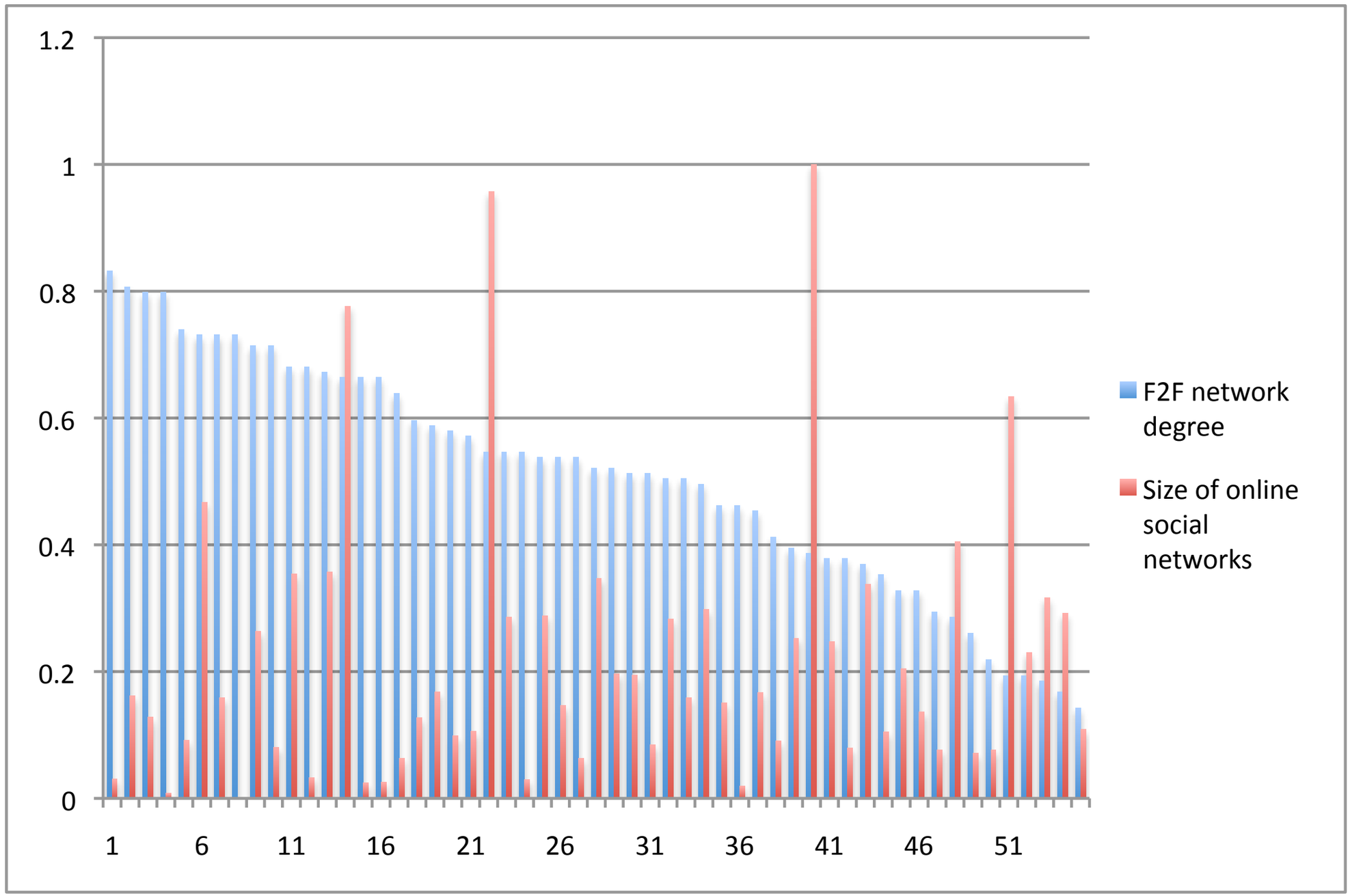
- *Weighted* conference social network (CSN): information on face-to-face contacts

  (i,j): $[\Delta t_{ij}^{(1)}, \Delta t_{ij}^{(2)}, \ldots, \Delta t_{ij}^{(n_{ij})}]$

  $\langle \Delta t_{ij}^{(k)} \rangle$, $\max(\Delta t_{ij}^{(k)})$, $w_{ij}=\Sigma \Delta t_{ij}^{(k)}$, n_{ij} …

- Online social network (OSN)

Online and off-line degrees



Network comparison

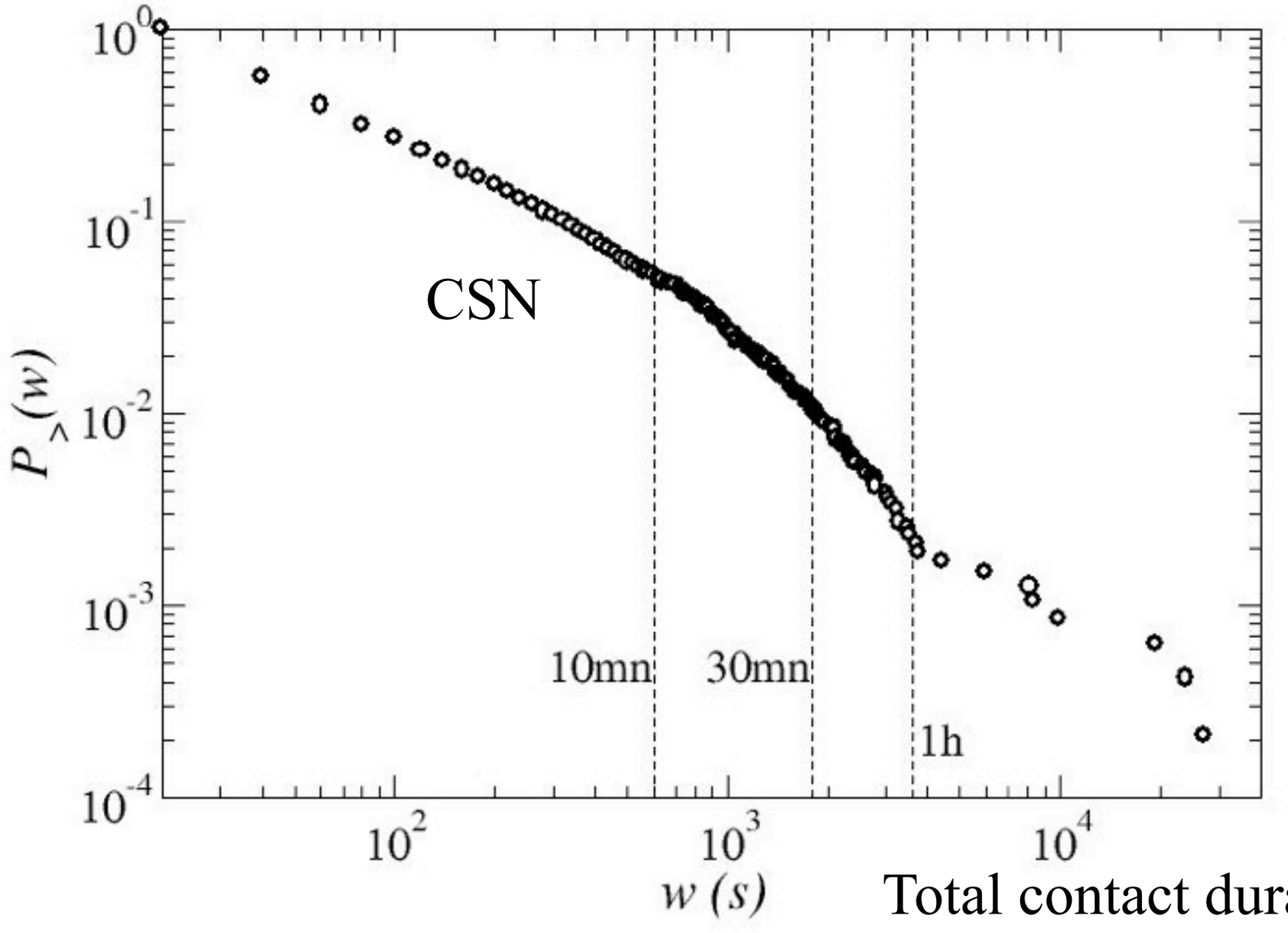
Total contact duration $w_{ij}=\Sigma_k \Delta t_{ij}^{(k)}$

Ex: ESWC

- Averaged over all links in the CSN: 160s
- Averaged over links belonging to both networks: 900s

Network comparison

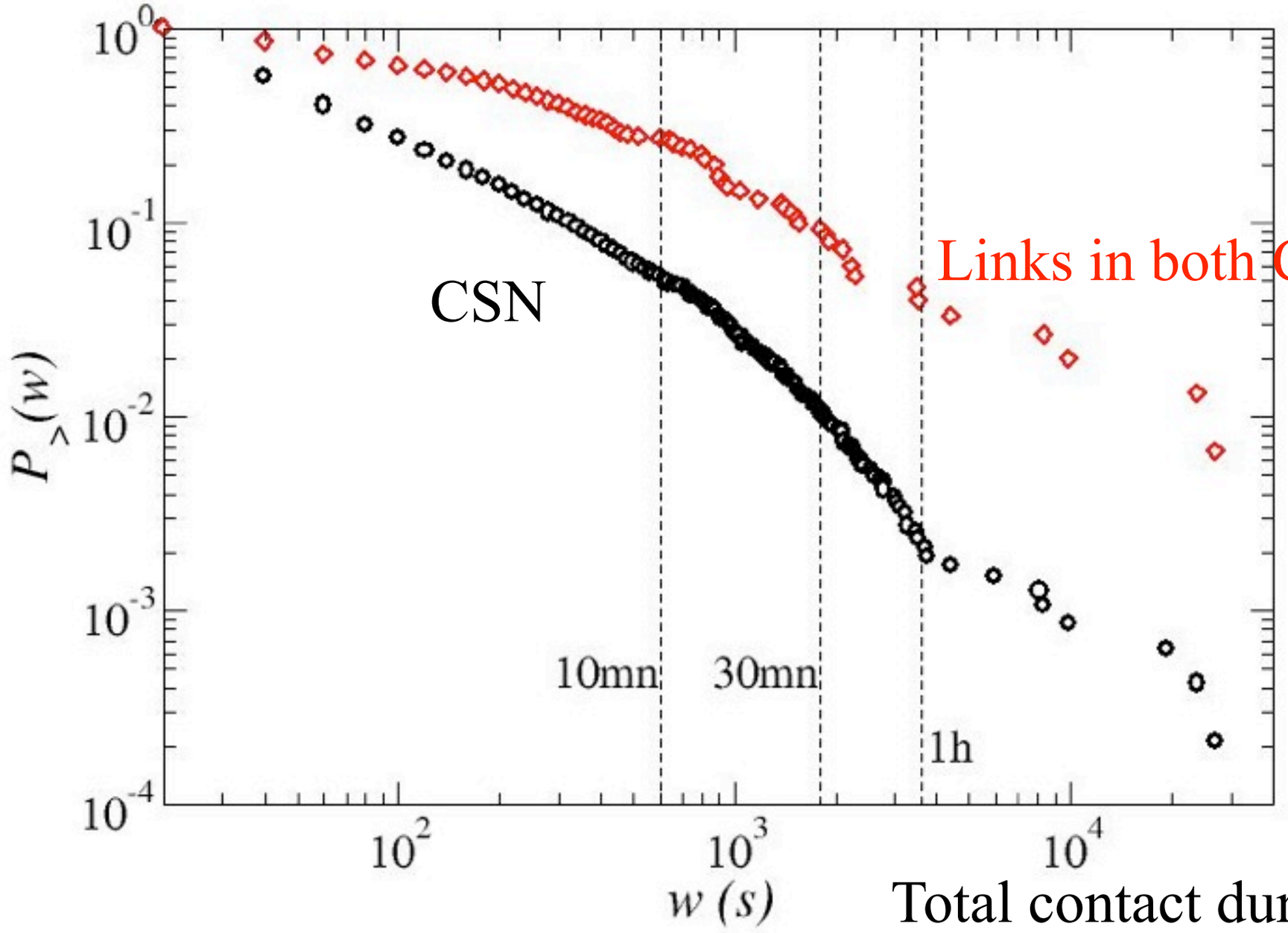
Distributions



Total contact duration

Network comparison

Distributions



Behavioral similarity I

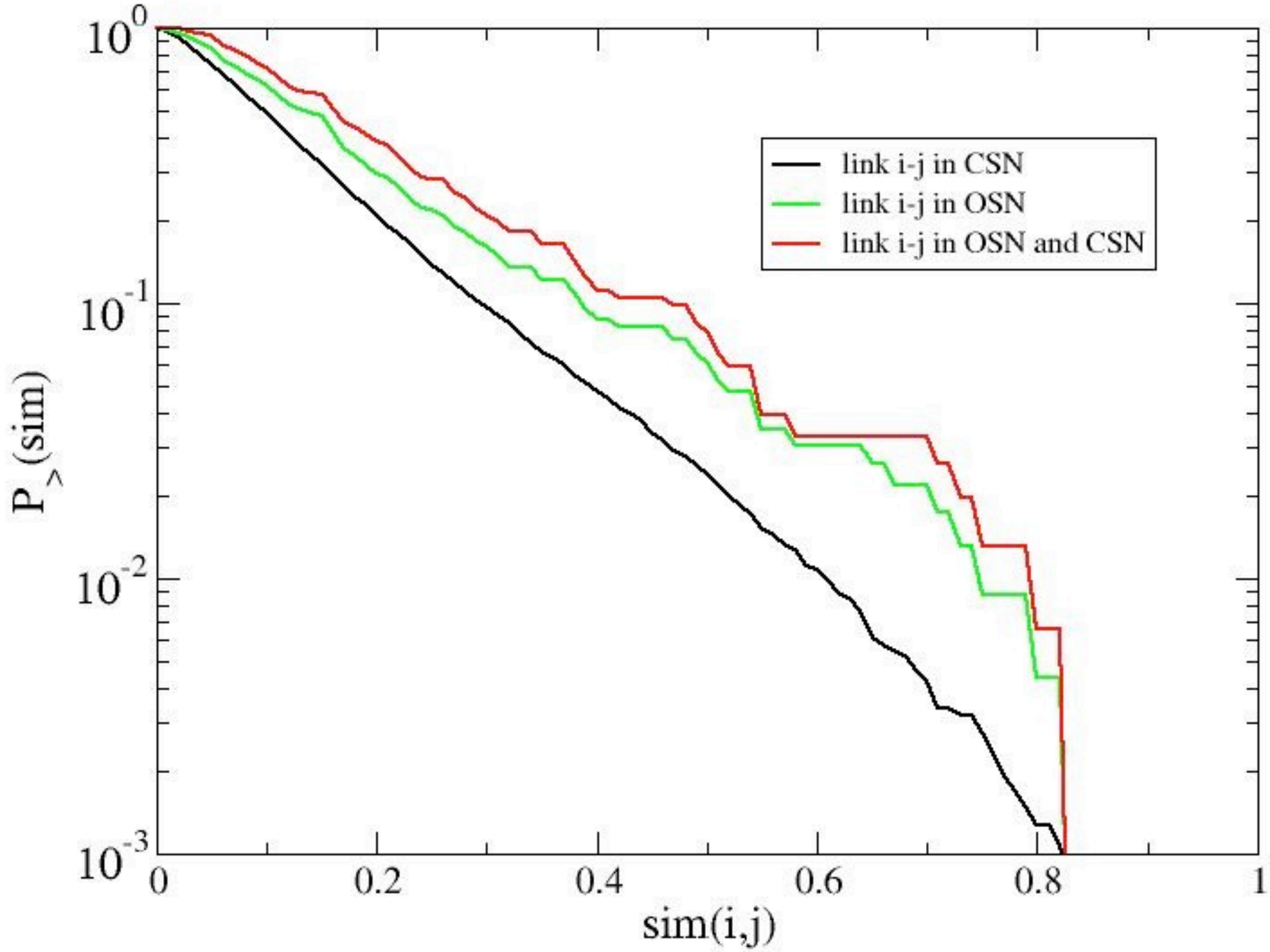
Similarity of neighborhoods of i and j in a weighted network:

- For each node: vector of normalized weights
- Similarity (i,j)= scalar product of vectors of i and j

$$sim(i,j) = \sum_k \frac{w_{ik} w_{jk}}{\sqrt{\sum_l w_{il}^2 \sum_l w_{jl}^2}}$$

sim(i,j)=0 if no common contacts
sim(i,j)=1 if same contacts and same relative amounts of time spent with the various contacts

Behavioral similarity I



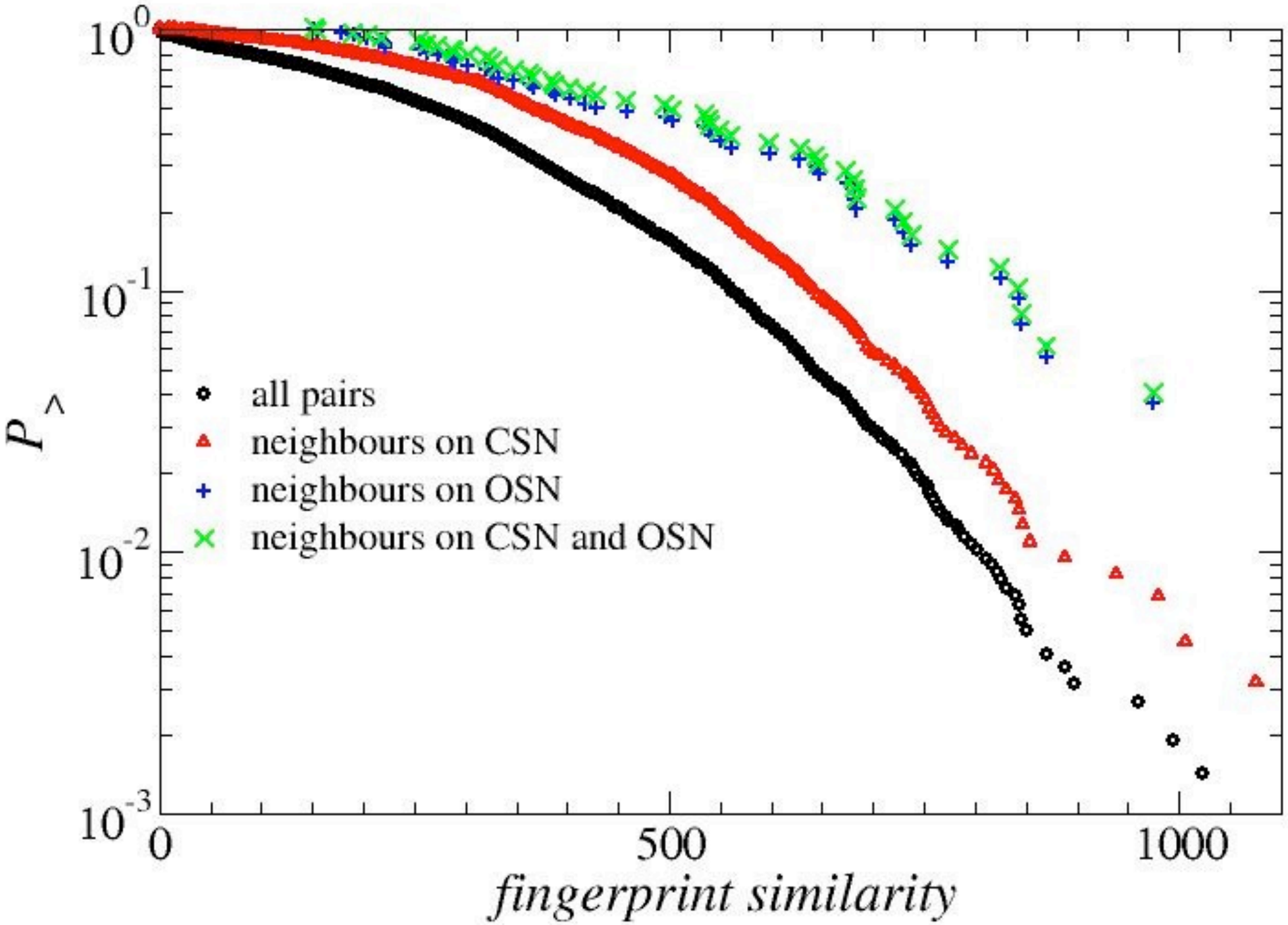
Behavioral similarity II

RFID badges: send packets to readers

For each badge, "fingerprint"= at each time t,
number of packets received by each antenna during [t-Dt,t]

Similarity between badges' fingerprints (averaged over time) gives a proxy for the similarity of trajectories of participants in physical space

Behavioral similarity II



On- vs off-line

Between online-linked attendees:

- More frequent, longer real-world contacts
- Stronger behavioral similarity
- Stronger (spatial) trajectory similarity

⇒Link prediction ? i.e., given the conference social network links, predict online social links?

using

- Total time spent in face-to-face interaction (w)
- Number of face-to-face interactions (n)
- Largest contact duration
- Trajectory similarity
- …

Link prediction

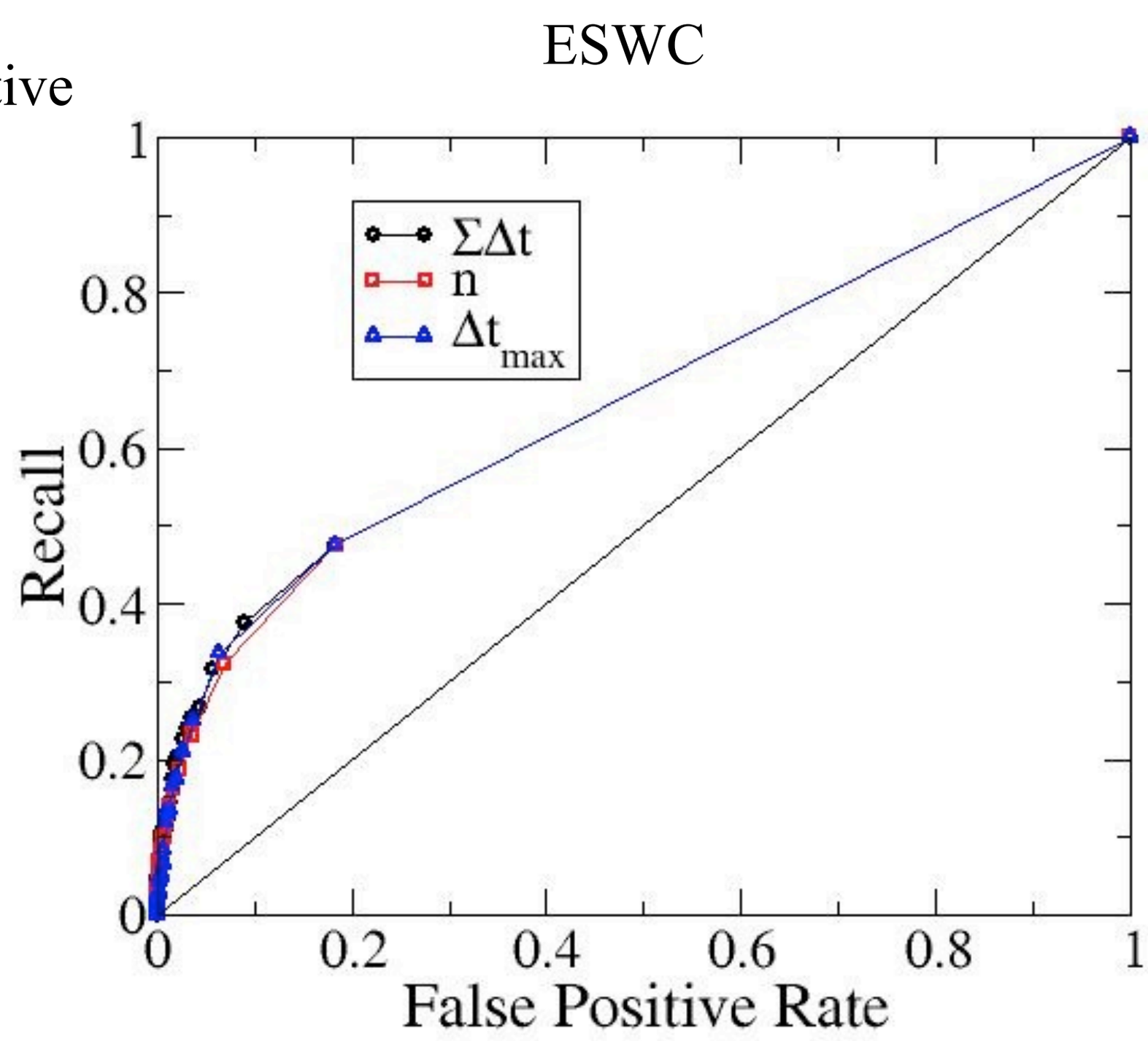
Quantifying the goodness of the prediction

	Predicted	Real
True positive	x	x
False positive	x	0
True negative	0	0
False negative	0	x

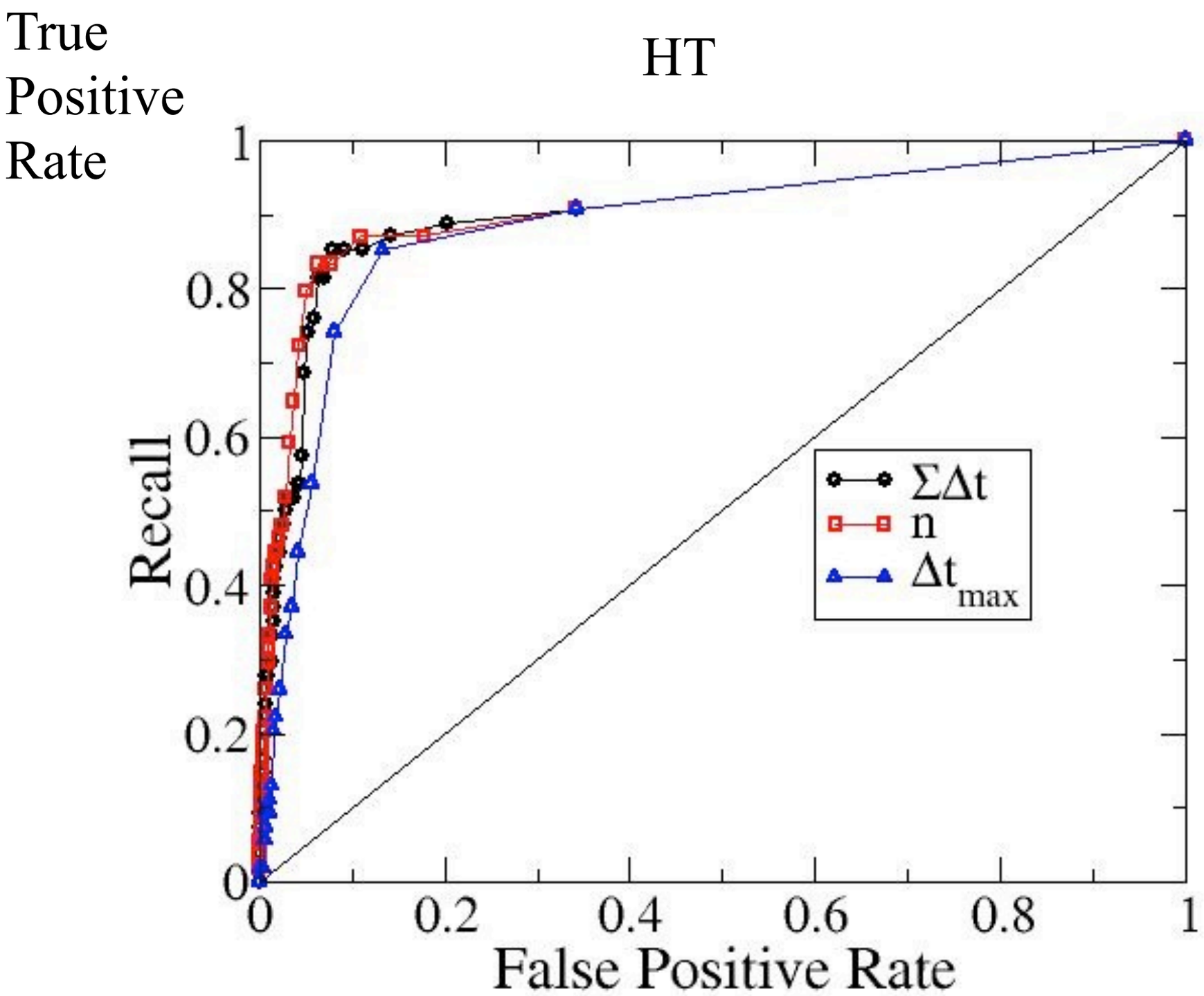
true positive rate (TPR)	**false positive rate (FPR)**
$TPR = TP / P = TP / (TP + FN)$	$FPR = FP / N = FP / (FP + TN)$

Link prediction

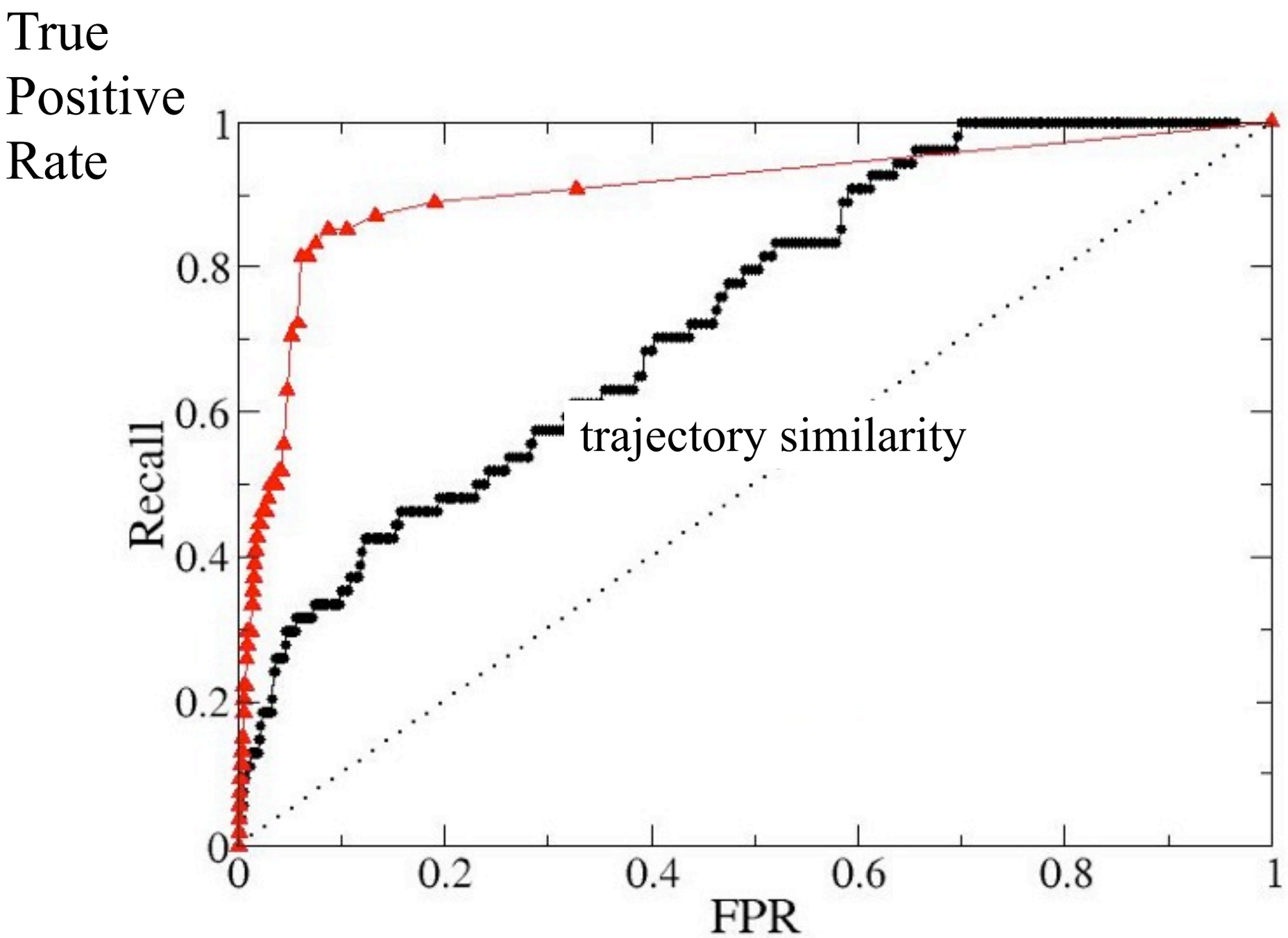
True Positive Rate



Link prediction



Link prediction



Contacts between co-authors

Characteristics	all participants	coauthors	FB	Twitter
average contact duration (s)	42	75	63	72
average edge weight (s)	141	4470	830	1010
average number of events per edge	3.37	60	13	14

Table 5. F2F contact characteristics between (i) all LSS users, (ii) LSS users who are coauthors, (iii) LSS users who are friends on Facebook, and (iv) pairs of users who are linked on Twitter.

Session chairs

Characteristics	all participants, 2009	chairs 2009	all participants, 2010	chairs, 2010
average degree	55	77.7	54	77.6
average strength	8590	19590	7807	22520
average weight	159	500	141	674
average number of events per edge	3.44	8	3.37	12

Table 4. Some characteristics of the ESWC 2010 chairs, and of the links between chairs, compared with the overall averages.

2 different years

Characteristics	all participants, 2009	all participants, 2010	common participants, 2009	common participants, 2010
Average degree	55	54	73	62
Average strength	8590	7807	16426	13216
Average weight	159	141	416	404
Average contact duration in seconds	46	42	52	57
Average number of contact events per edge	3.44	3.37	8	7

Table 3. Average characteristics in each year of the participants to both ESWC 2009 and ESWC 2010, and of the contact patterns between these returning participants, as compared to the average over all participants.

In summary

- Between online-linked attendees:
  - More frequent, longer real-world contacts
  - Stronger behavioral similarity
  - Stronger trajectory similarity
- Link prediction
  - Very good results
  - Better results when including face-to-face contact information
- Also:
  - Strong effect for coauthorship and chairs
  - Comparison of two successive years

Some perspectives & work in progress

★ Real-world contact patterns
"Atlas" of human interaction patterns,
Similarities and differences across contexts

★ Time-varying networks: fundamental and applied issues:

- characterization, modeling, visualization
- role of causality constraints
- characterization of "central", "important" nodes
- representations, "summary" of dynamical networks?
- applications to
  - spreading of information/diseases
  - social network analysis from
    (i) behavior
    (ii) online data
    (iii) combine with survey data

MULTIPLEXITY OF (TEMPORAL) NETWORKS

Collaborators: Online social networks

Luca Aiello (Turin university)
Ciro Cattuto (ISI Foundation, Turin)
Ben Markines (Indiana University)
Filippo Menczer (Indiana University)
Giancarlo Ruffo (Turin university)
Rossano Schifanella (Turin university)

SocioPatterns team and collaborators

Ciro Cattuto (ISI Foundation, Turin)
Wouter Van Den Broeck (ISI Foundation, Turin)

Vittoria Colizza (INSERM, Paris & ISI Foundation, Turin)
Lorenzo Isella (ISI Foundation, Turin)
Anna Machens (CPT Marseille)
André Panisson (ISI & University of Turin)
Jean-François Pinton (ENS Lyon)
Marco Quaggiotto (ISI & Politecnico di Milano)
Juliette Stehlé (CPT Marseille)
Alessandro Vespignani (Northeastern University & ISI)

Milosch Meriac and Brita Meriac (Bitmanufaktur)

Harith Alani (Open University, UK)
Gianluca Correndo, Martin Szomszor (Southampton, UK)

references

★ C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, A. Vespignani, *Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks*, PLoS ONE 5(7), e11596 (2010)

★ R. Schifanella, A. Barrat, C. Cattuto, B. Markines, F. Menczer, *Folks in folksonomies: Social link prediction from shared metadata,* Proc. of WSDM 2010, arxiv:1003.2281

★ L. Aiello, A. Barrat, C.Cattuto, G. Ruffo, R. Schifanella, *Link creation and profile alignment in the aNobii social network.,* Proc. of Socialcom 2010, arxiv:1006.4966

★ A. Barrat, C. Cattuto, M. Szomszor, W. Van den Broeck, H. Alani, *Social dynamics in conferences: analyses of data from the Live Social Semantics application*,: Proceedings of the 9th International Semantic Web Conference ISWC2010.

★ L. Aiello, A. Barrat, C. Cattuto, R. Schifanella, G. Ruffo, *Link creation and information spreading over social and communication ties in an interest-based online social network*. EPJ Data Science 1:12 (2012)