Examples of online social network analysis

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
- Real-world interactions
- Online networks/ social web

NEW (large-scale) DATASETS,

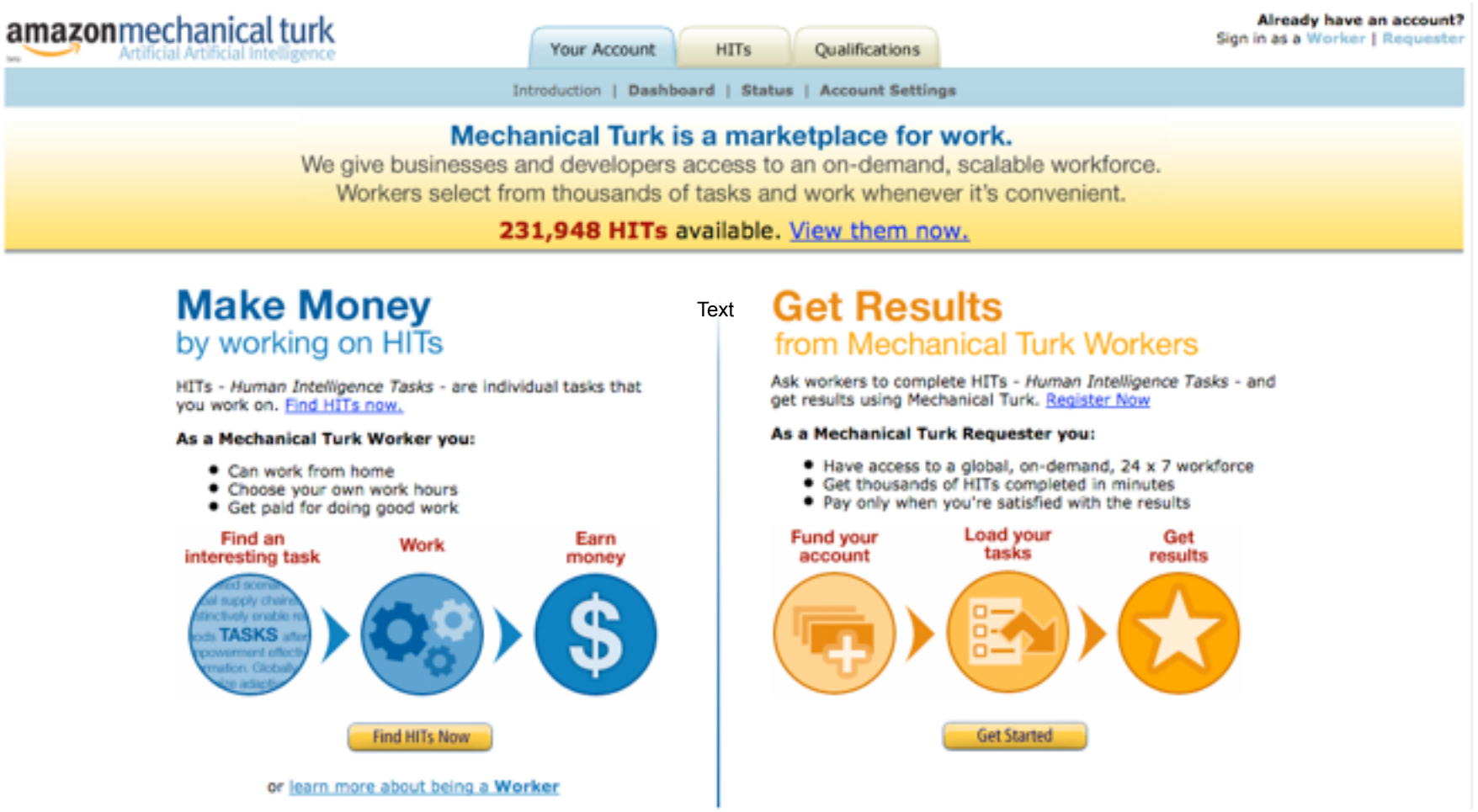
longitudinal data

New laboratories

- Social network properties
  - homophily
  - selection vs influence
- Triadic closure, preferential attachment
- Social balance
- Dunbar number
- Experiments at large scale...

Another social science lab:

crowdsourcing, e.g. Amazon Mechanical Turk



http://experimentalturk.wordpress.com/

Judgment and Decision Making, Vol. 5, No. 5, August 2010, pp. XX–XX

Running experiments on Amazon Mechanical Turk

Gabriele Paolacci*
Advanced School of Economics, Ca' Foscari University of Venice

Jesse Chandler
Woodrow Wilson School of Public and International Affairs, Princeton University

Panagiotis G. Ipeirotis
Leonard N. Stern School of Business, New York University

Abstract

Although Mechanical Turk has recently become popular among social scientists as a source of experimental data, doubts may linger about the quality of data provided by subjects recruited from online labor markets. We address these potential concerns by presenting new demographic data about the Mechanical Turk subject population, reviewing the strengths of Mechanical Turk relative to other online and offline methods of recruiting subjects, and comparing the magnitude of effects obtained using Mechanical Turk and traditional subject pools. We further discuss some additional benefits such as the possibility of longitudinal, cross cultural and prescreening designs, and offer some advice on how to best manage a common subject pool.

Keywords: experimentation, online research

New laboratories

Caveats:

- online links can differ from real social links
- population sampling biases?
- “big” data does not automatically mean “good” data

The social web

- social networking sites
- blogs + comments + aggregators
- community-edited news sites, participatory journalism
- content-sharing sites
- discussion forums, newsgroups
- wikis, Wikipedia
- services that allow sharing of bookmarks/favorites
- ...and mashups of the above services


del.icio.us


Technorati


WIKIPEDIA


flickr
LOVES YOU


YouTube
Broadcast Yourself™


BibSonomy ::


citeulike


reddit


twitter


myspace.com™


Connotea


digg™


LinkedIn

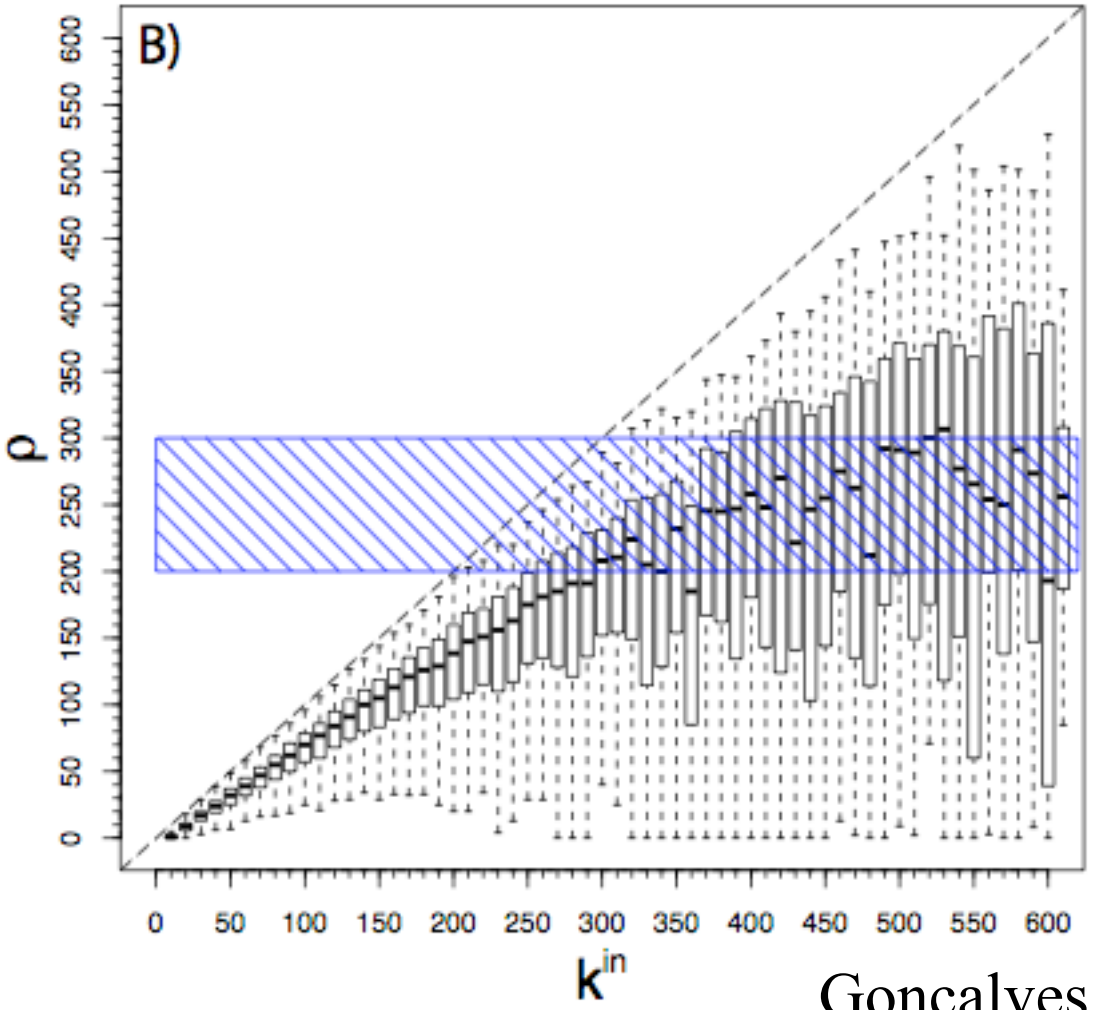

newsvine.com


facebook


THE ICY NORTH (YAHOO, WINDOWS LIVE)
REUNION DOT COM
CLASSMATES DOT COM
FRIENDSTER
FALLPANTY
MOUNTAINS OF WEB 1.0
AOL
CHAT ROOMS
QWOHLM
YAHOO GAMES
MAP OF ONLINE COMMUNITIES
AND RELATED POINTS OF INTEREST
GEOGRAPHIC AREA REPRESENTS
ESTIMATED SIZE OF MEMBERSHIP
MYSPACE
NOOB SEA
THE SHORES OF TUBES
MYSPACE BRAND
BLURTY
OK CUPID
CYWORLD
ORKUT
XANGA
FACEBOOK
LJ
GULF OF YOUTUBE
SEA OF CULTURE
PICZO
BROADCASTER
FLICKR
LAST.FM
DEVIANT ART
STRAITS OF WEB 2.0
P2P SHOALS
OCEAN OF SUBCULTURE
DIGG
FARK
REDDIT
SPAMARETA
VIRAL STRAITS
BAY OF ANGST
PRACTICALS
FOCUS ON REAL LIFE
FOCUS ON WEB
INTELLECTUALS
THE COMPASS-ROSE-SHAPED ISLAND
SECOND LIFE
LINEAGE
WoW
EQ
HERE BE ANTHROPOMORPHIC DRAGONS
THE BLOGIPELAGO
SULAWESI
THE WET SEA
THE WIKIPEDIA PROJECT
WIKI-PEDIA
USENET
SEA OF MEMES
IRC ISLES
SOURCEFORGE
(NOT A COMPLETE SURVEY. SIZES BASED ON BEST FIGURES I COULD FIND BUT INVOLVED SOME GUESSWORK. DO NOT USE FOR NAVIGATION.)
SPRING 2007


EMAIL
SMS
SPOKEN LANGUAGE
FACEBOOK
NORTHERN WASTELAND OF UNREAD UPDATES
FARMVILLE
HAPPY FARM
PLAINS OF ASSORTED PUBLIC EMAIL INTERACTIONS
LAMEBOOK BAY
DOPAMINE SEA
MMO ISLE
WoW
GULF OF LAG
UPDATED
MAP OF ONLINE COMMUNITIES
SIZE ON MAP REPRESENTS VOLUME OF DAILY SOCIAL ACTIVITY (POSTS, CHAT, ETC). BASED ON DATA GATHERED OVER THE SPRING AND SUMMER OF 2010
MYSPACE
ORKUT
LINKEDIN
ODNOKLASSNIKI
FLICKR
BLIZZARD BAY
YOUTUBE
FORUMS
TROLL BAY
SEA OF MEMES
TWITTER
BAY OF DRAMA
SEA OF OPINIONS
SEA OF PROTOCOL CONFUSION
AIM
GREAT FIREWALL
GULF OF CHINA
SKYPE
QQ
BLOGOSPHERE
(CORE REGION)
GOSSIP BLOGS
LIBERAL BLOGS
BAY OF FLAME
ABOUT THIS MAP
COMMUNITIES RISE AND FALL, AND TOTAL MEMBERSHIP NUMBERS ARE NO LONGER A GOOD MEASURE OF A COMMUNITY'S CURRENT SIZE AND HEALTH. THIS UPDATED MAP USES SIZE TO REPRESENT TOTAL SOCIAL ACTIVITY IN A COMMUNITY — THAT IS, HOW MUCH TALKING, PLAYING, SHARING, OR OTHER SOCIALIZING HAPPENS THERE. THIS MEANT SOME COMPARING OF APPLES AND ORANGES, BUT I DID MY BEST AND TRIED TO BE CONSISTENT.
ESTIMATES ARE BASED ON THE BEST NUMBERS I COULD FIND, BUT INVOLVED A GREAT DEAL OF GUESSWORK, SPECTRAL INFERENCE, RANDOM SAMPLING, NONSENSICAL SAMPLING, A 20,000-CELL SPREADSHEET, EYEBALLING, CAROLING, TEA-LEAF READING, GOAT SACRIFICES, AND GUT INSTINCT (I.E. MAKING THINGS UP).
FORUMS
4CHAN
/b/

An example: Dunbar number on twitter



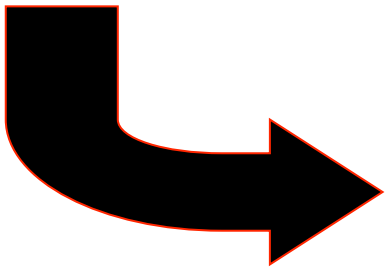
Fraction of reciprocated connections as a function of in-degree

Gonçalves et al, PLoS One 6, e22656 (2011)

Sharing and annotating

Examples:

- Flickr: sharing of photos
- Last.fm: music
- aNobii: books
- Del.icio.us: social bookmarking
- Bibsonomy: publications and bookmarks
- …



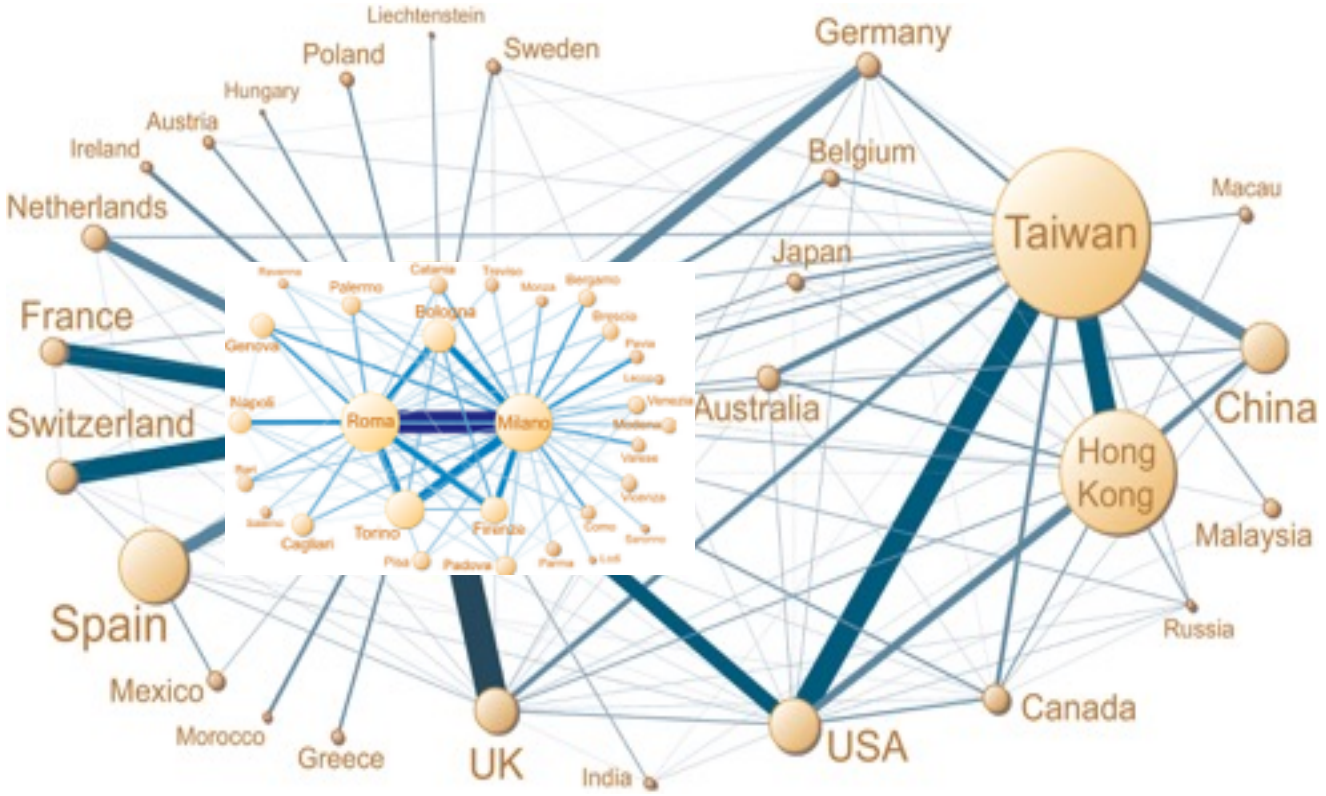
- "Social" networks
- "specialized" content-sharing sites
- Users expose profiles (content) and links

Case study: aNobii

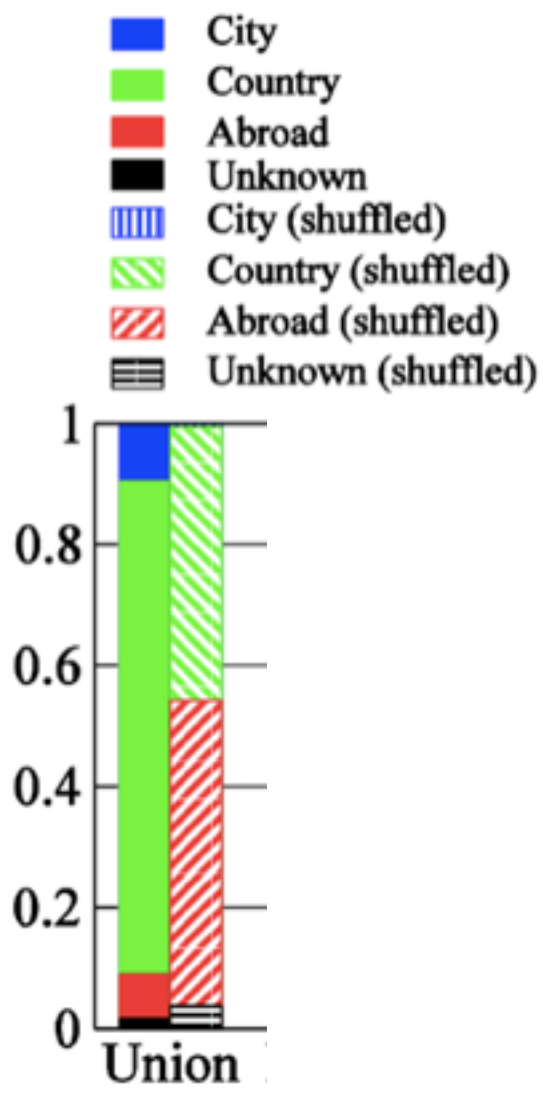
(similar analysis done also for last.fm and flickr)

- User's profile:
  - Books read by user
  - Wishlist of books
  - Tags describing the books
  - Groups of discussion
  - Geographical information
- Social network (directed)
- ~100 000 users

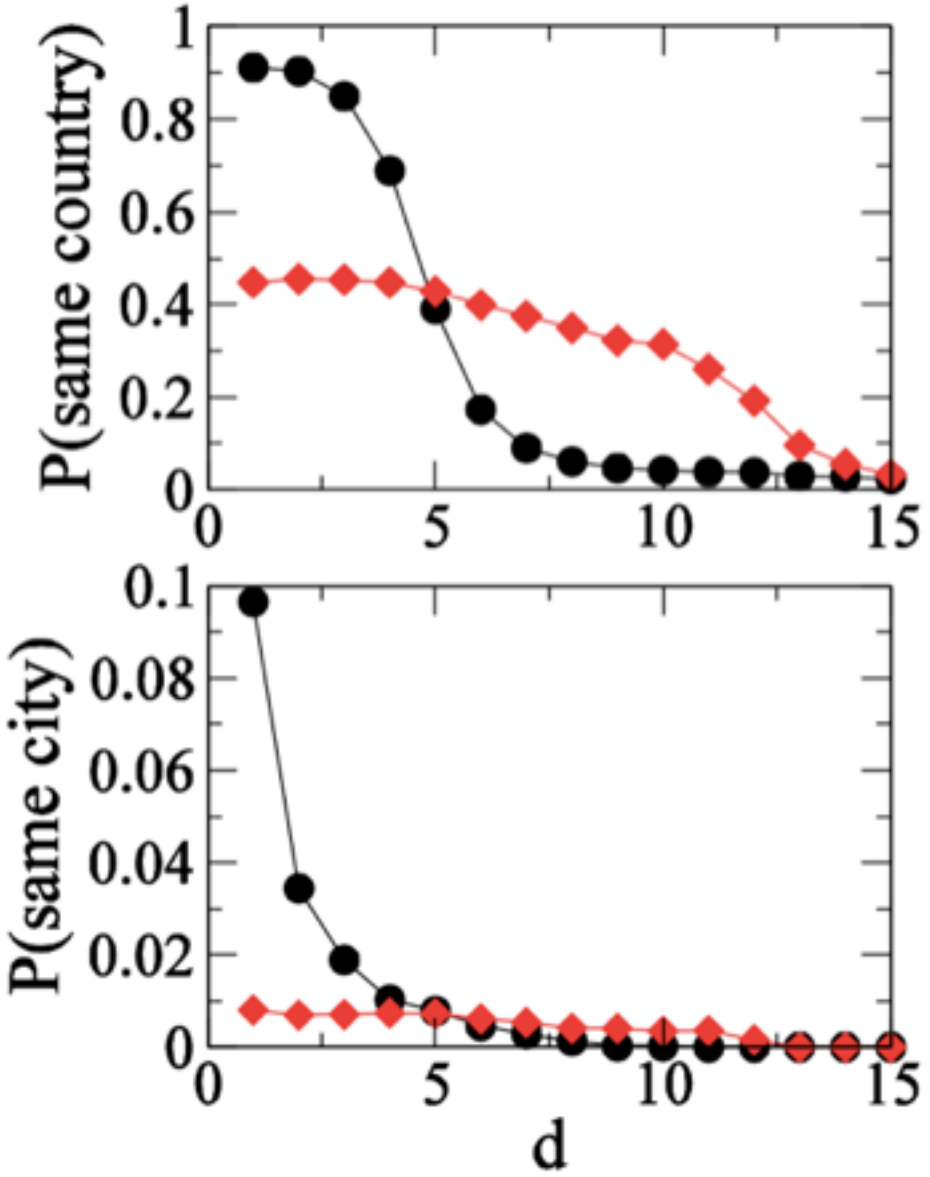
Geography


Liechtenstein
Poland
Sweden
Germany
Hungary
Austria
Ireland
Belgium
Netherlands
Macau
Taiwan
Japan
France
Switzerland
Australia
China
Hong Kong
Malaysia
Russia
Spain
Mexico
Morocco
Greece
UK
India
USA
Canada
Roma
Milano
Torino
Firenze
Bologna
Genova
Napoli
Palermo
Cagliari
Pisa
Padova

Geography



Fraction of links



Distance on network

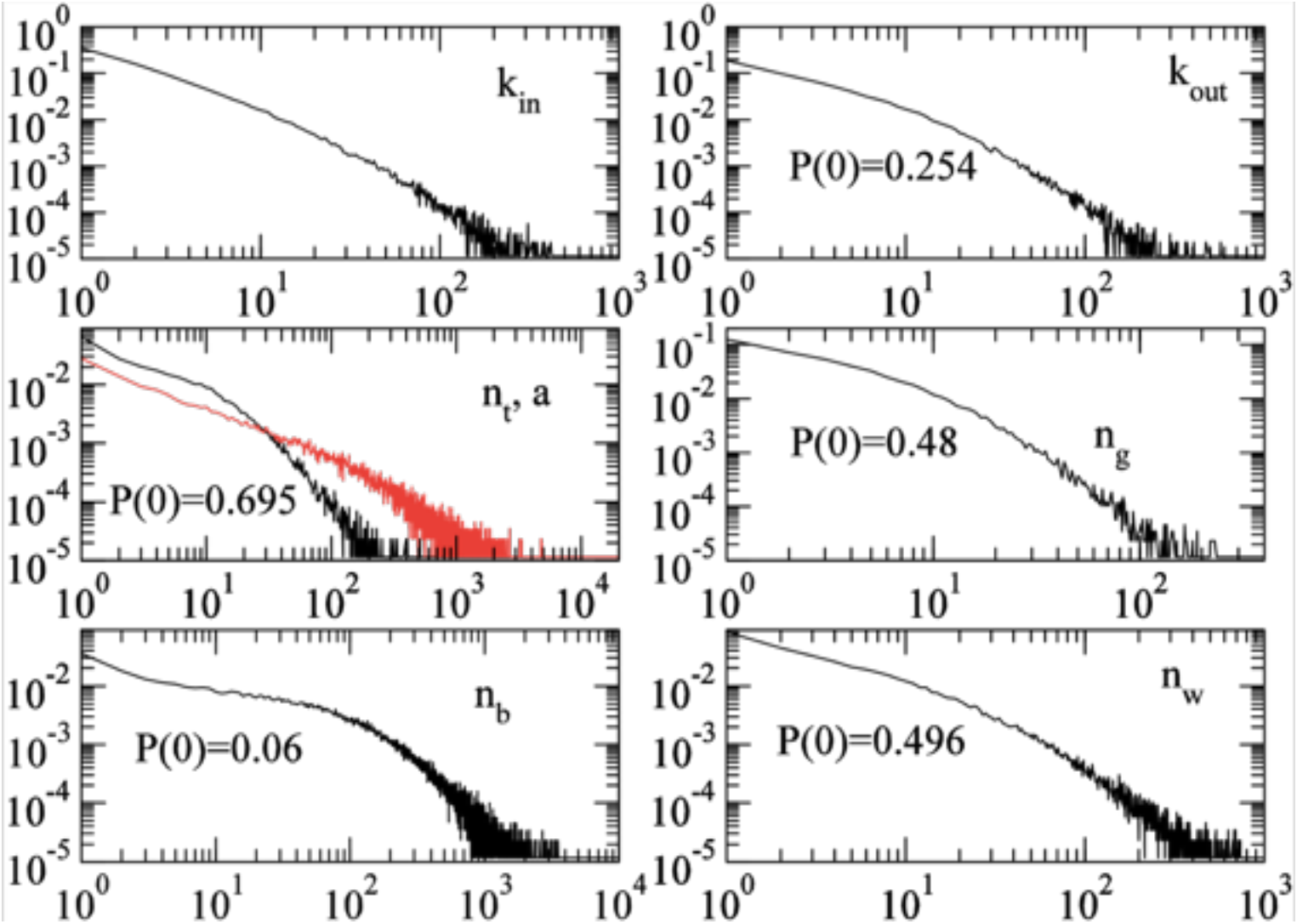
Activity measures

Heterogeneity of all users' activity amounts

Networking

Tagging/Groups

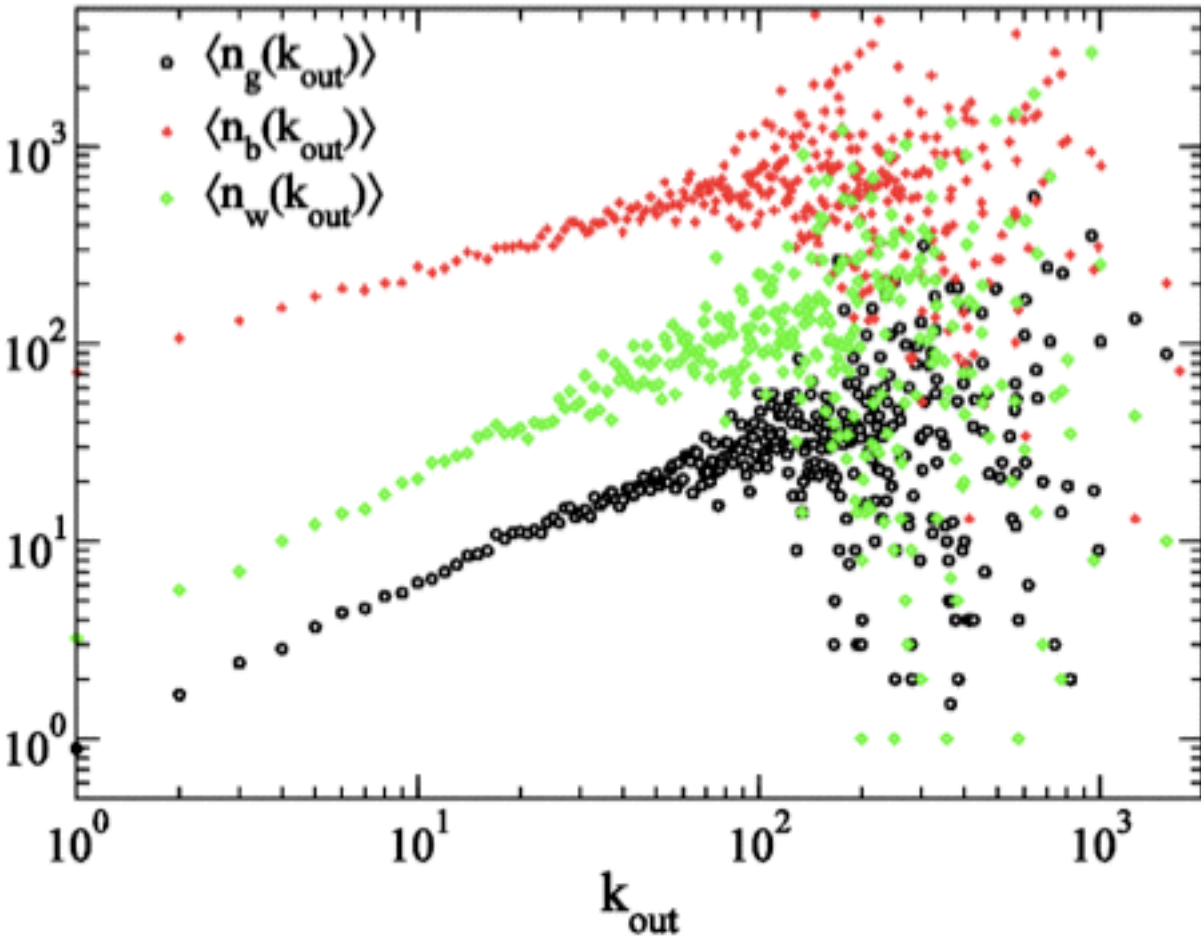
Books



Correlations

Correlation between user's activity types:

Sharing and annotating activities

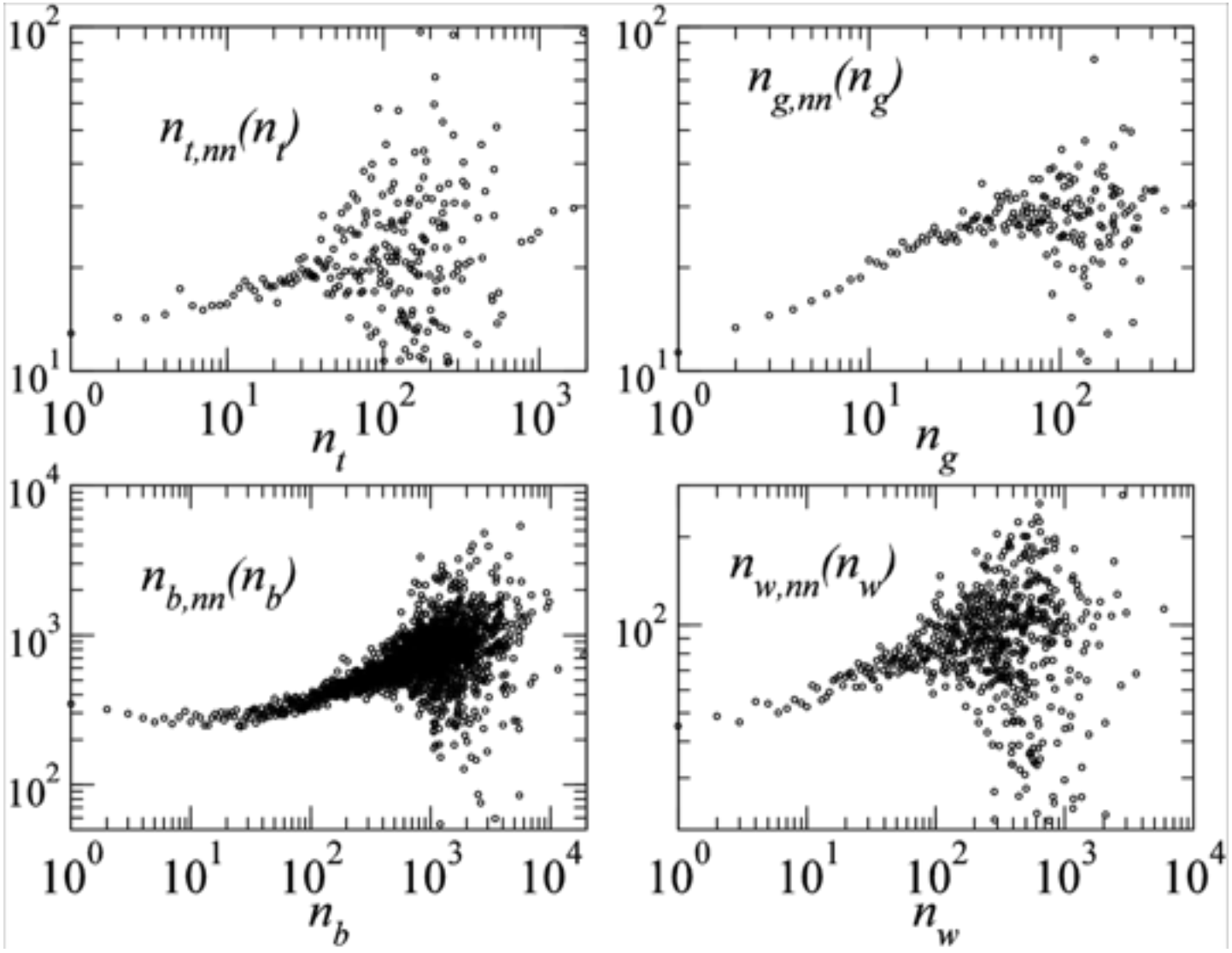


Social networking

Mixing patterns

average activity of nearest neighbors as a function of own activity

The more a user is active, the more its neighbours are active



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?

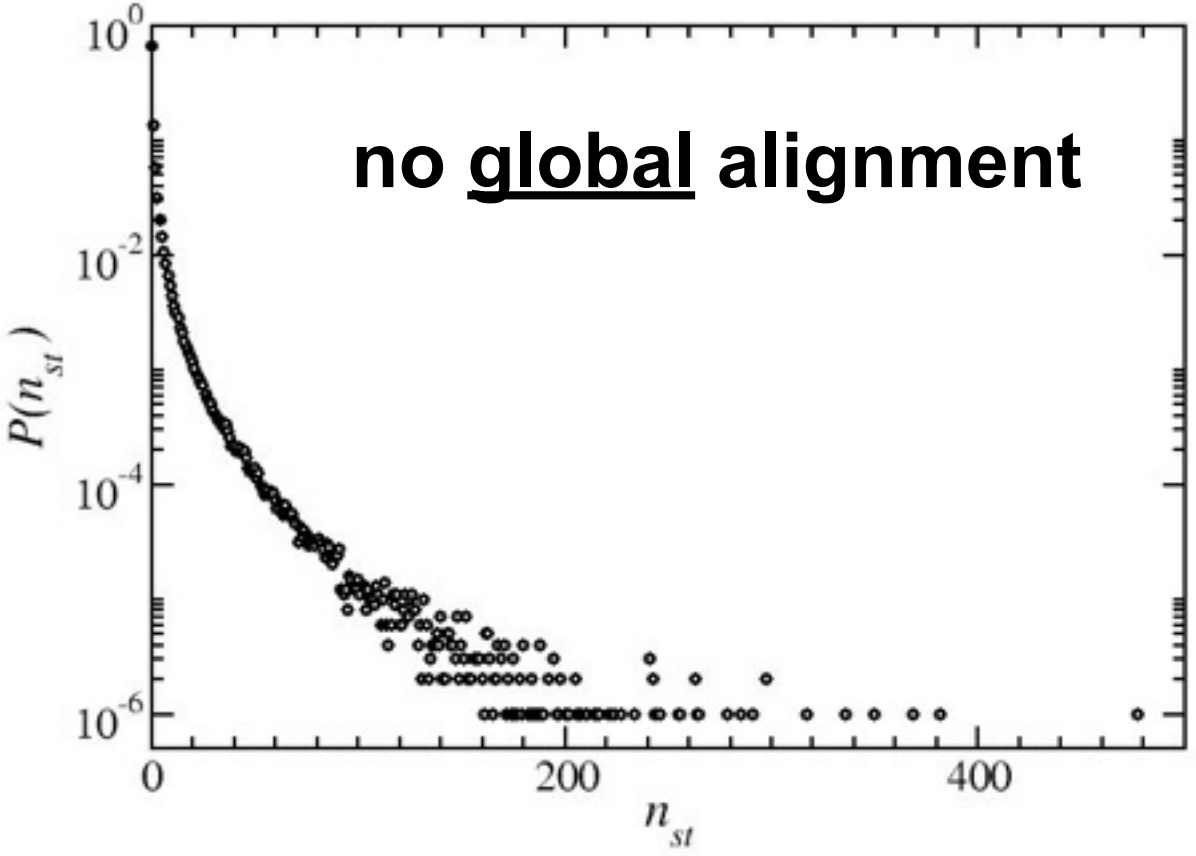
measures of alignment:

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

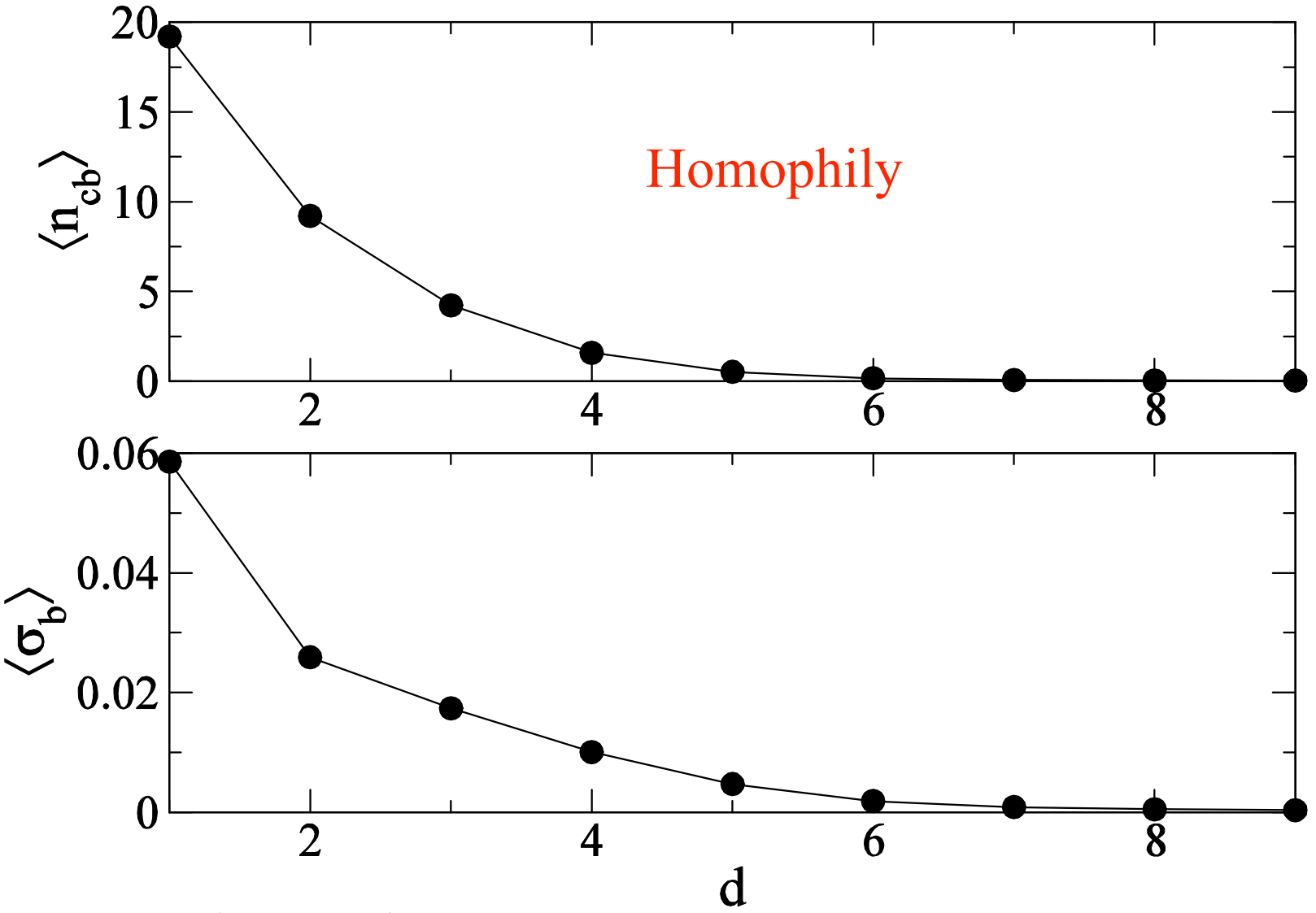
Alignment of users' profiles

random pairs of users:

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



Alignment along the network



Average number of common books of two users

Average normalized similarity measure between two users

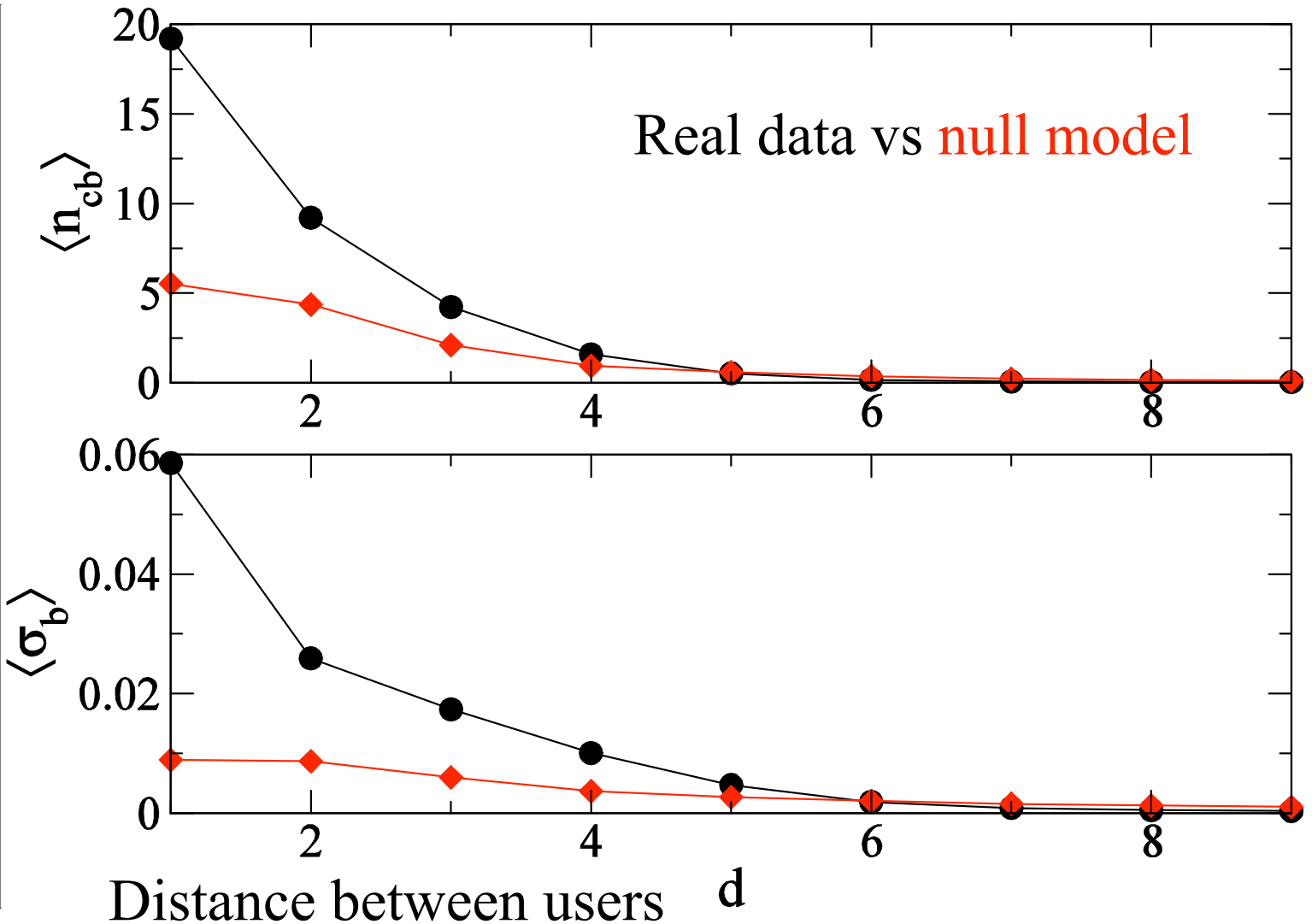
Distance between users on social network

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

Origin of homophily?

Suppose that there are two friends named Ian and Joey, and Ian's parents ask him the classic hypothetical of social influence: “If your friend Joey jumped off a bridge, would you jump too?" Why might Ian answer “yes”?

- because Joey’s example inspired Ian (**social contagion/influence**)
- because Joey infected Ian with a parasite which suppresses fear of falling (biological contagion)
- because Joey and Ian are friends on account of their shared fondness for jumping off bridges (**manifest homophily**, on the characteristic of interest)
- because Joey and Ian became friends through a thrill-seeking club, whose membership rolls are publicly available (**secondary homophily**, on a different yet observed characteristic)
- because Joey and Ian became friends through their shared fondness for roller-coasters, which was caused by their common thrill-seeking propensity, which also leads them to jump off bridges (**latent homophily**, on an unobserved characteristic)
- because Joey and Ian both happen to be on the Tacoma Narrows Bridge in November, 1940, and jumping is safer than staying on a bridge that is tearing itself apart (**common external causation**)

http://arxiv.org/abs/1004.4704

is obesity contagious on Facebook ?

fact: obese individuals are clustered

1. because of selection effects, in which people are choosing to form friendships with others of similar obesity status?

2. because of the confounding effects of homophily according to other characteristics, in which the network structure indicates existing patterns of similarity in other dimensions that correlate with obesity status?

3. because changes in the obesity status of a person's friends was exerting a (presumably behavioral) influence that affected his or her future obesity status?

N. A. Christakis et al., N. Engl. J. Med. 2007; 357:370-37

Origin of homophily?

selection vs influence

Need to observe temporal evolution

aNobii, dynamics

Successive snapshots at intervals of 15 days

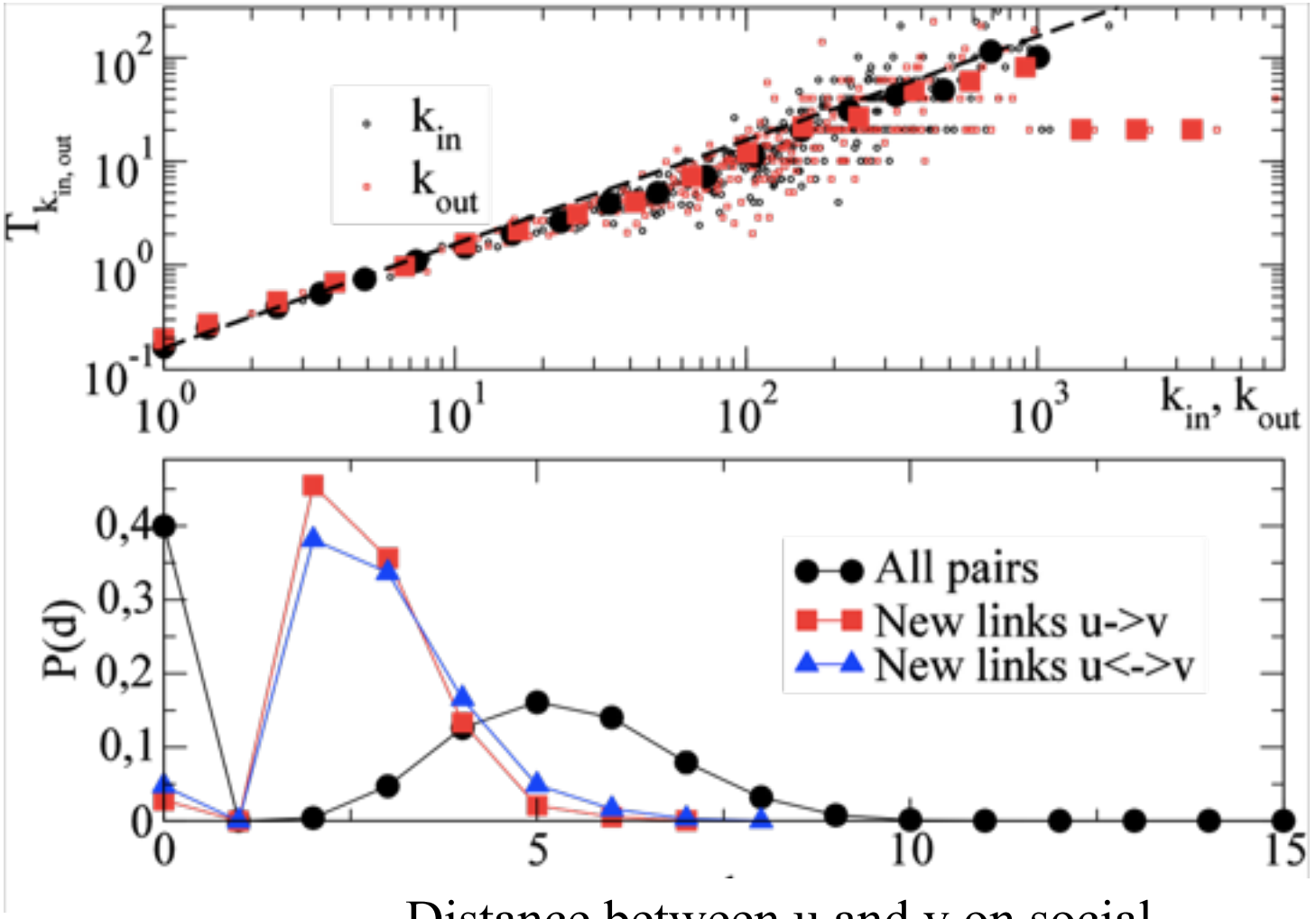
- New nodes
- New links from new to old nodes

  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?

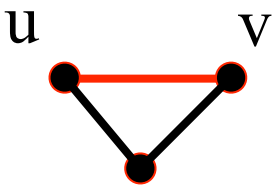
Dynamics: new nodes, new links



Distance between u and v on social network before creation of link (u,v)

Preferential attachment dynamics of new nodes

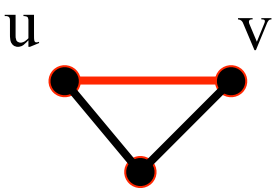
Triangle closure (many new links between users who were at distance 2)



Dynamics: selection or influence?

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

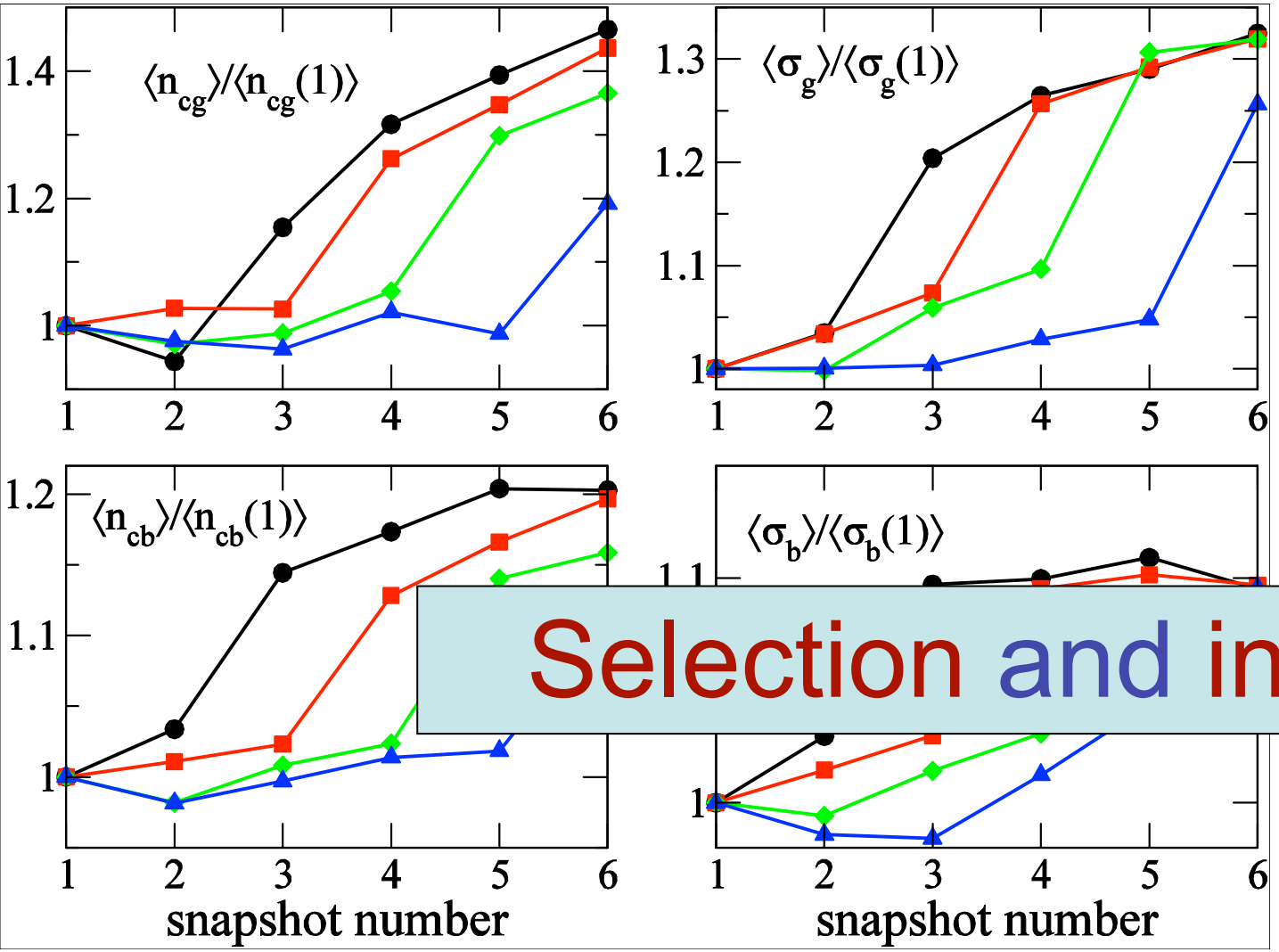
New links between already present users



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?

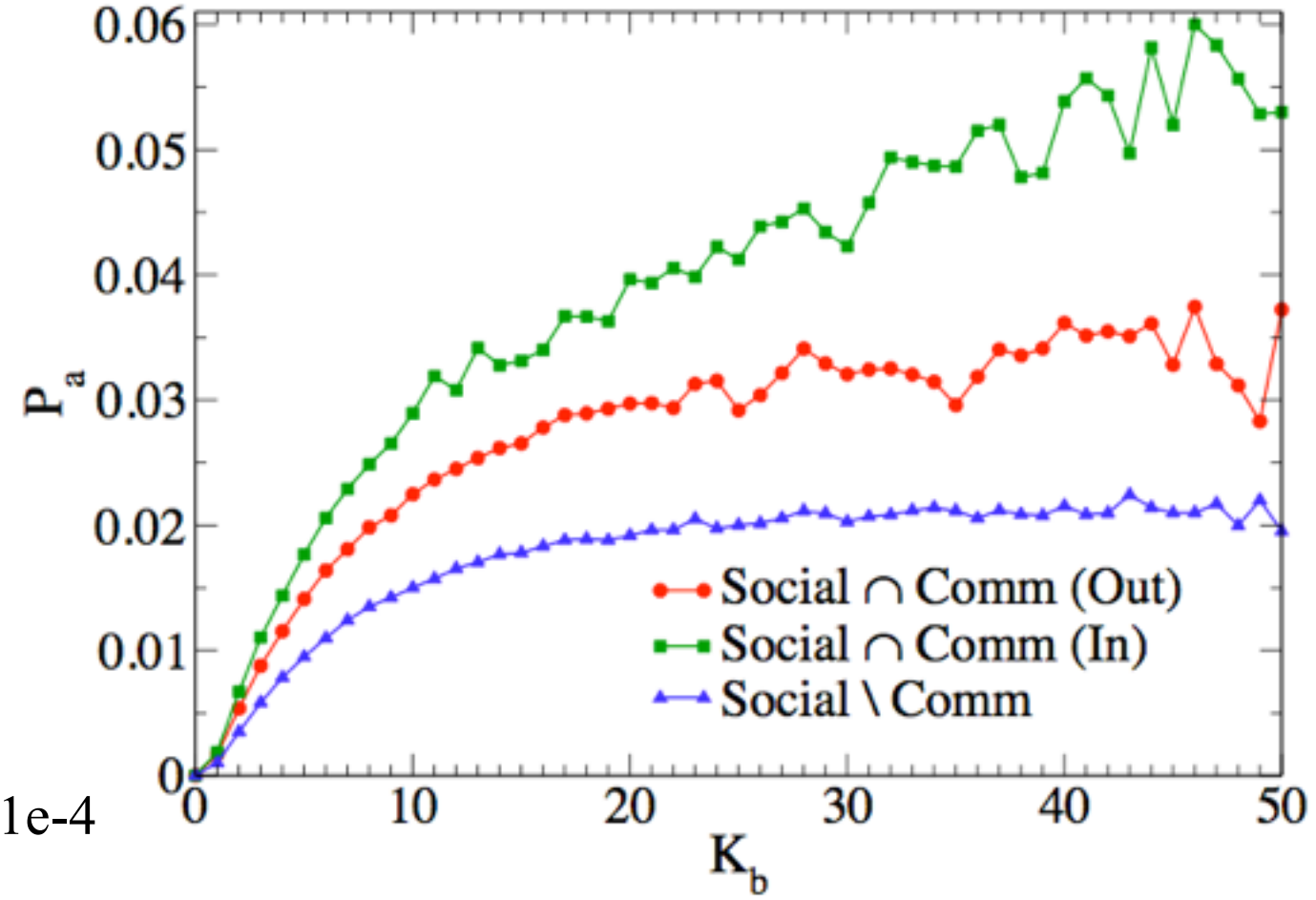


Evolution of similarity ***before and after*** link creation

Selection and influence

Bi-directional causality relation between similarity and link creation

Influence



P(0)~1e-4

Probability to adopt a book between t and t+1 vs number of neighbours having read this book at t

Summary and related work

- Similar results for other networks: Last.fm, flickr
- Possibility to predict *existence* of links
- "Laboratories" for social network analysis and testing of sociological theories, see also e.g.
  - Crandall et al., Proc of Knowledge discovery and Data Mining 2008
  - Leskovec, Huttenlocher, Kleinberg, arxiv:1003.2424, 1003.2429
  - Szell, Lambiotte, Thurner, arxiv:1003.5137 (PNAS 2010)
  - Gonçalves, Perra, Vespignani, arxiv:1105.5170
  - …

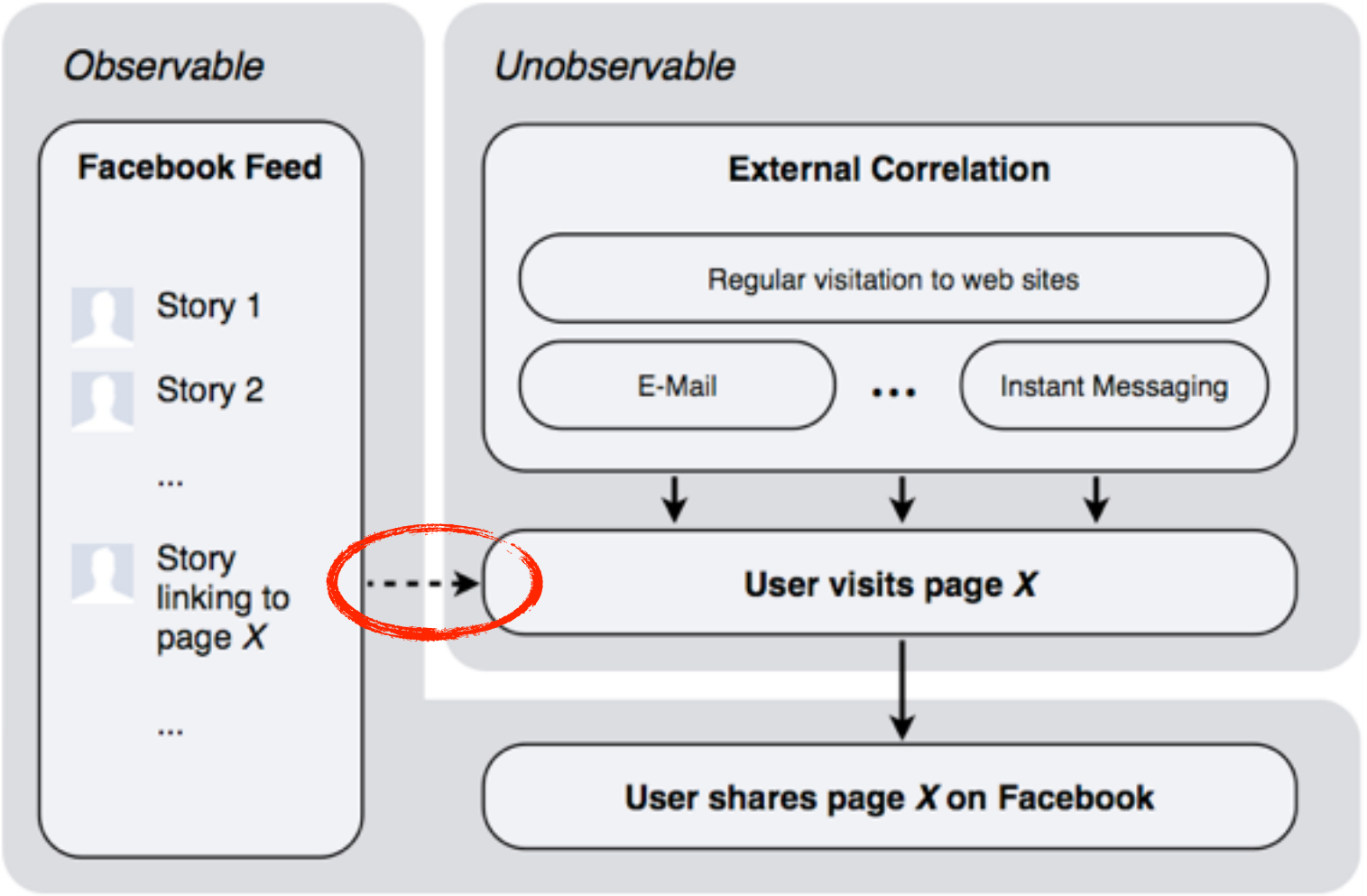
- Prediction of *creation* of links
- Recommendations
- Study of adoption mechanisms (book, author)

R. Schifanella et al., Proc. of Web Search and Data Mining (WSDM) 2010 , arxiv:1003.2281
L. Aiello et al., Proc. of Socialcom 2010, arxiv:1006.4966

a controlled experiment

E. Bakshy *et al.*, *The Role of Social Networks in Information Diffusion*, WWW2012

sharing links on Facebook



experimental design

feed

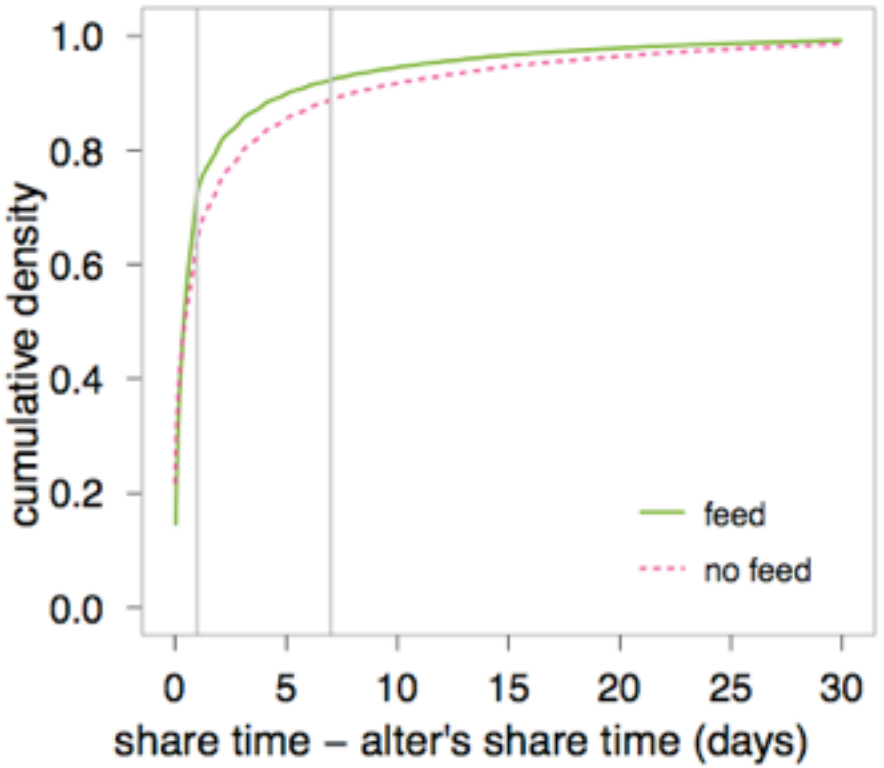
no-feed

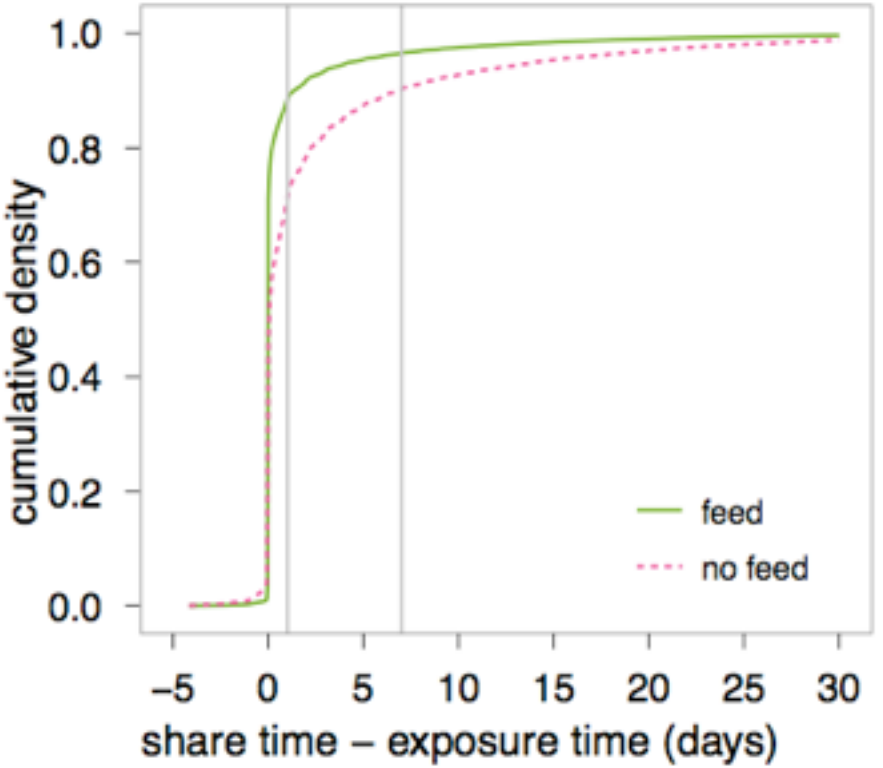
balancing the demographics

Demographic Feature (% of subjects)	**feed**	**no feed**
Gender		
FEMALE	51.6%	51.4%
MALE	46.7%	47.0%
UNSPECIFIED	1.5%	1.5%
Age		
17 OR YOUNGER	12.8%	13.1%
18-25	36.4%	36.1%
26-35	27.2%	26.9%
36-45	13.0%	12.9%
46 OR OLDER	10.6%	10.9%
Country (top 10 & other)		
UNITED STATES	28.9%	29.1%
TURKEY	6.1%	5.8%
GREAT BRITAIN	5.1%	5.2%
ITALY	4.2%	4.1%
FRANCE	3.8%	3.9%
CANADA	3.7%	3.8%
INDONESIA	3.7%	3.5%
PHILIPPINES	2.1%	2.3%
GERMANY	2.3%	2.3%
MEXICO	2.0%	2.1%
226 OTHERS	37.5%	37.7%

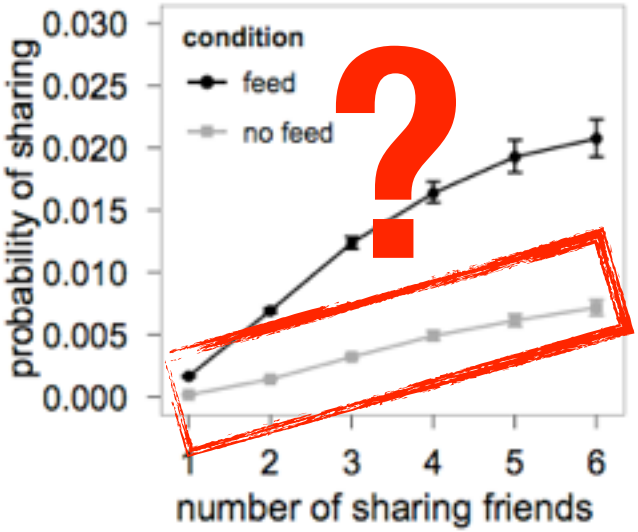
Table 1: Summary of demographic features of subjects assigned to the *feed* ($N = 160,688,092$) and *no feed* ($N = 218,743,932$) condition. Some subjects may appear in both columns.

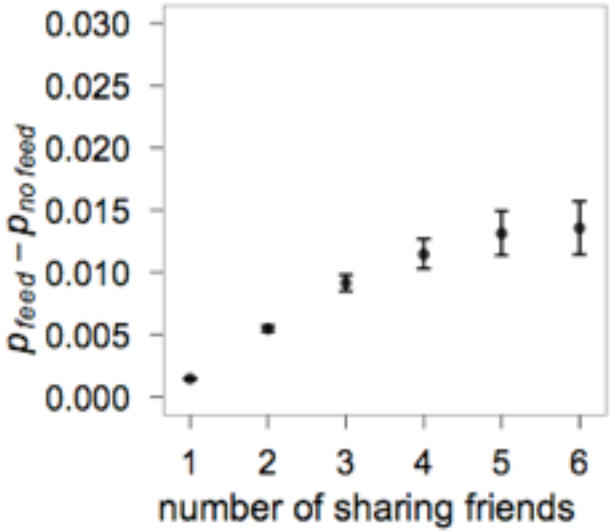
timing of shares

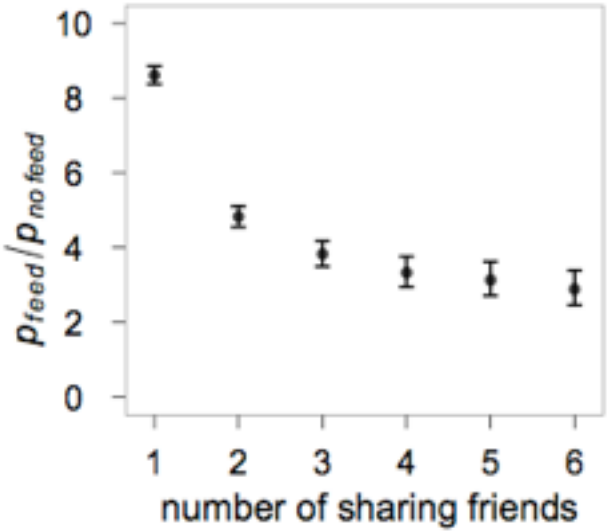




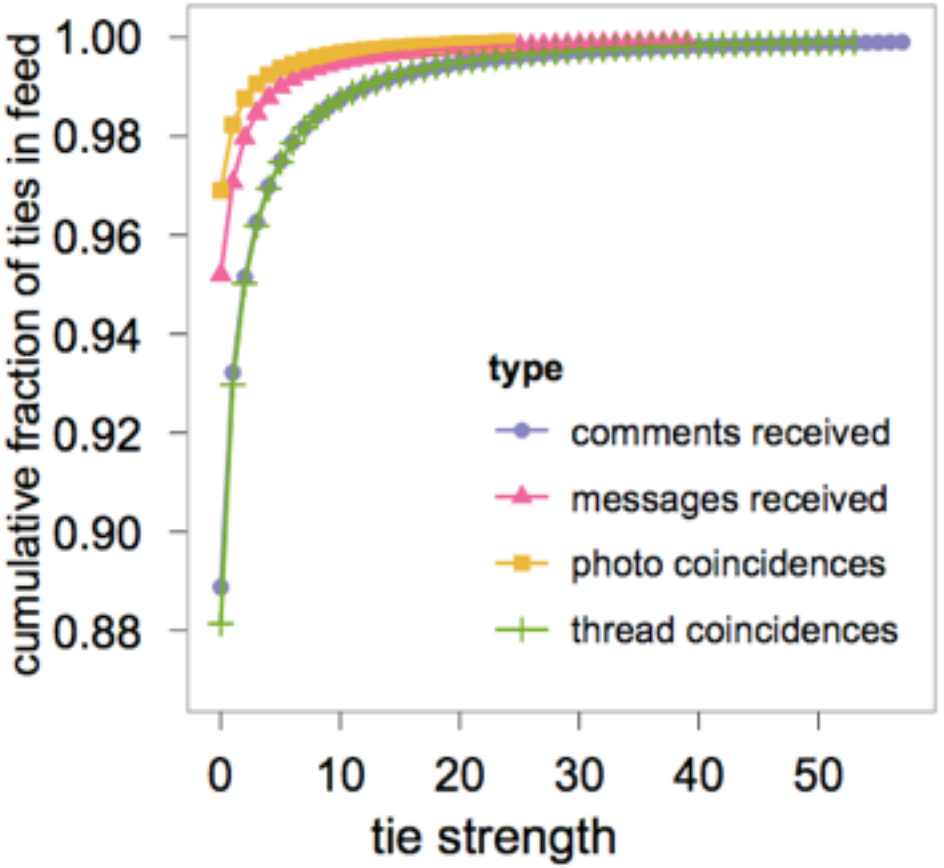
effect of multiple sharing friends



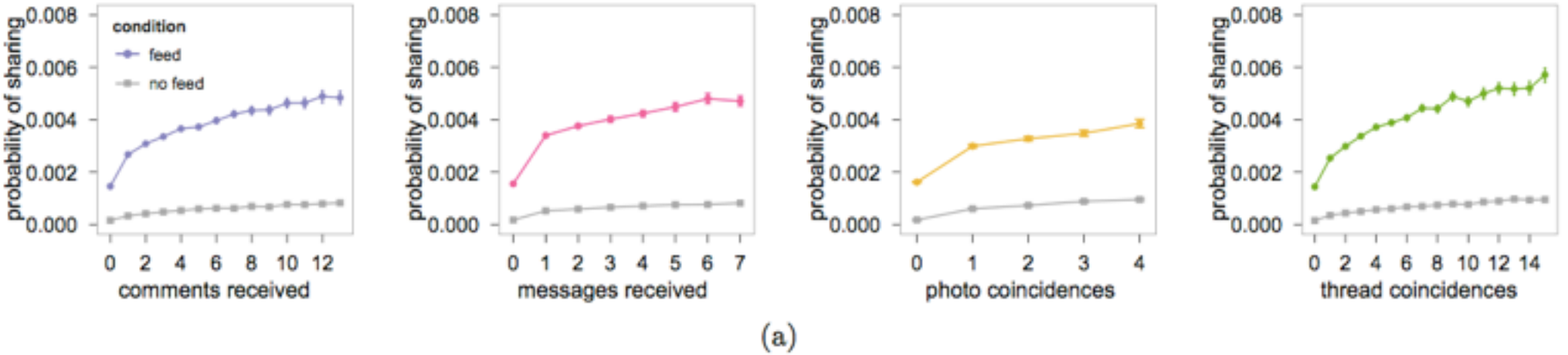




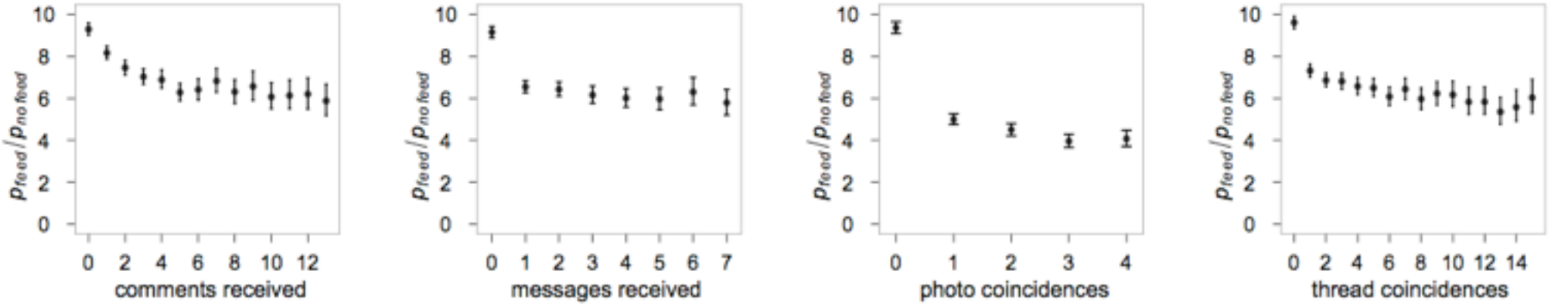
the impact of tie strength



the impact of tie strength



(a)

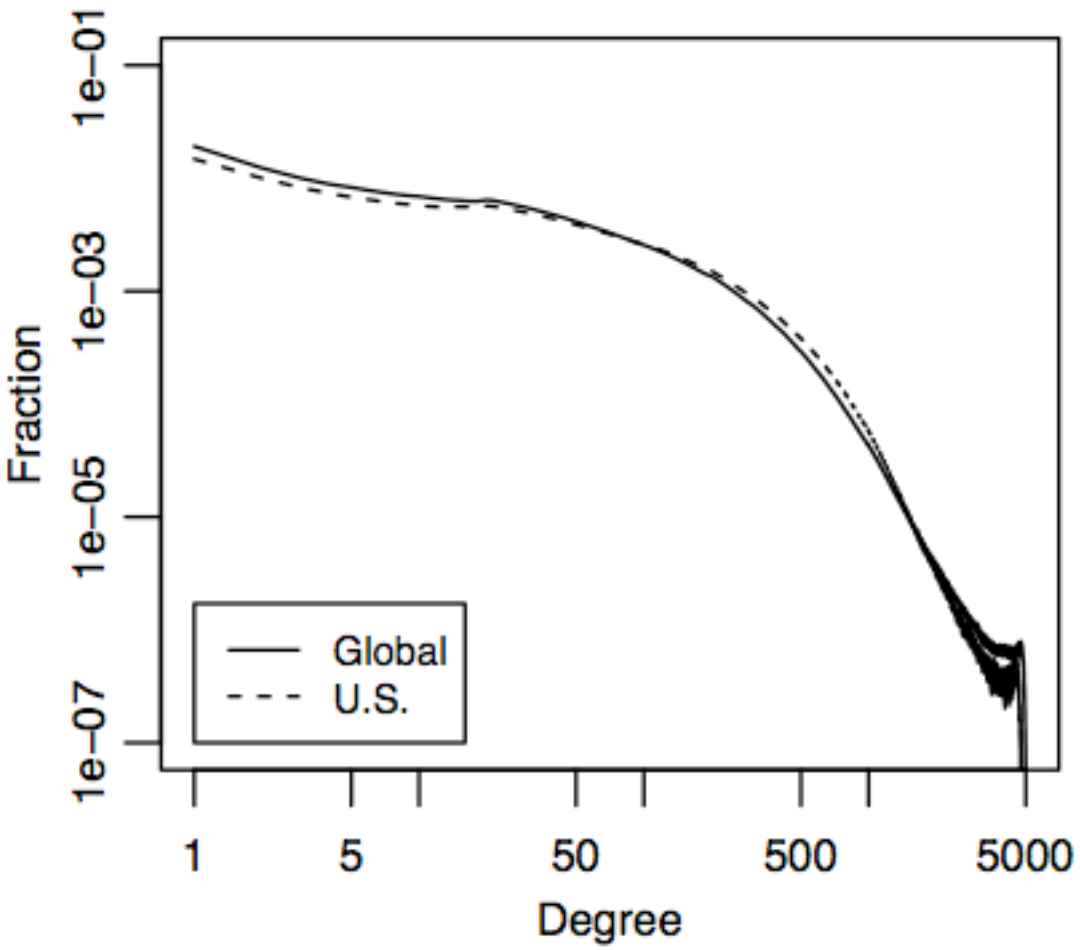


http://arxiv.org/abs/1201.4145

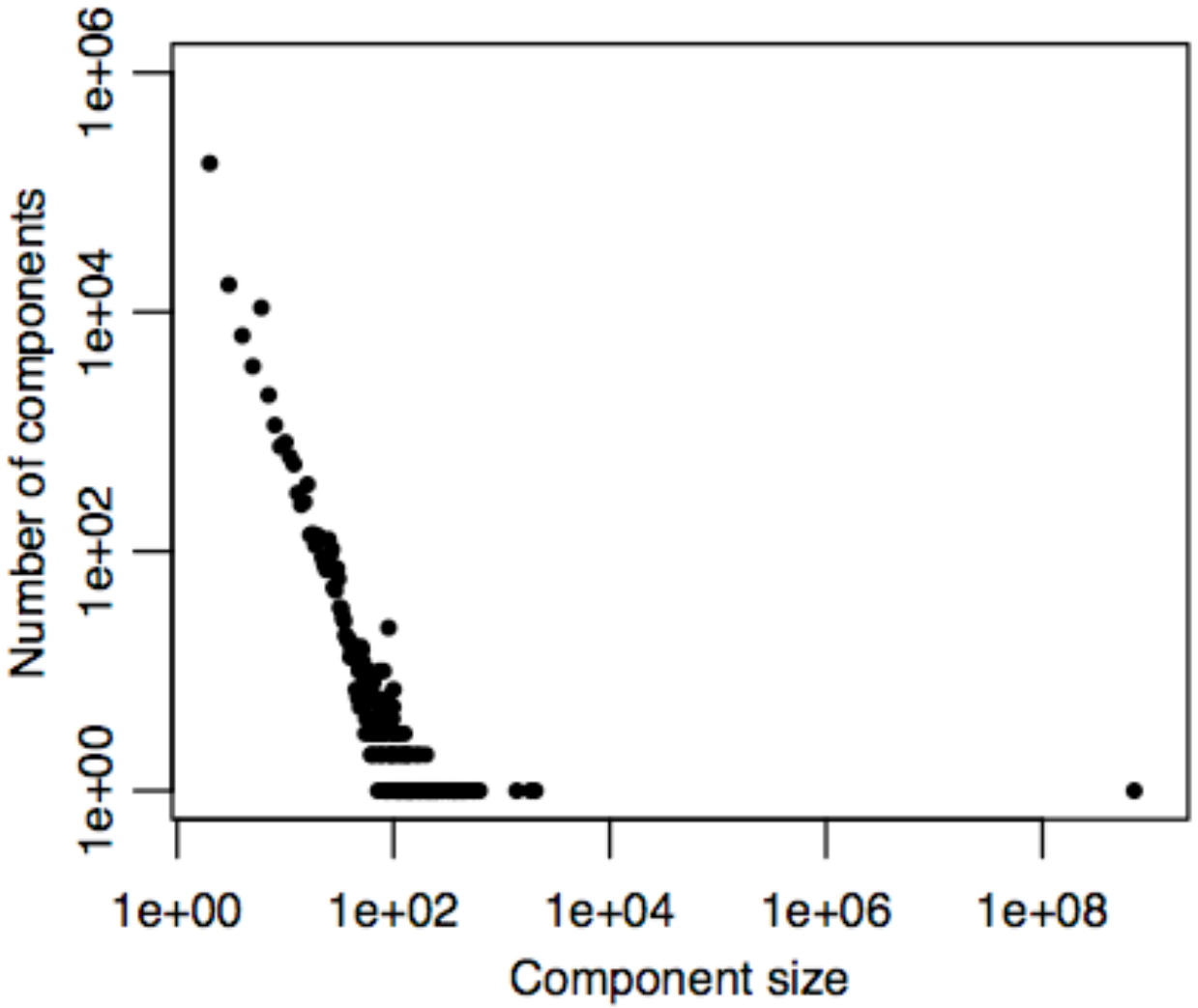
The case of facebook

The Anatomy of the Facebook Social Graph, arXiv:1111.4503
Four Degrees of Separation, arxiv:11.4570
The Role of Social Networks in Information Diffusion, arxiv:1201.4145

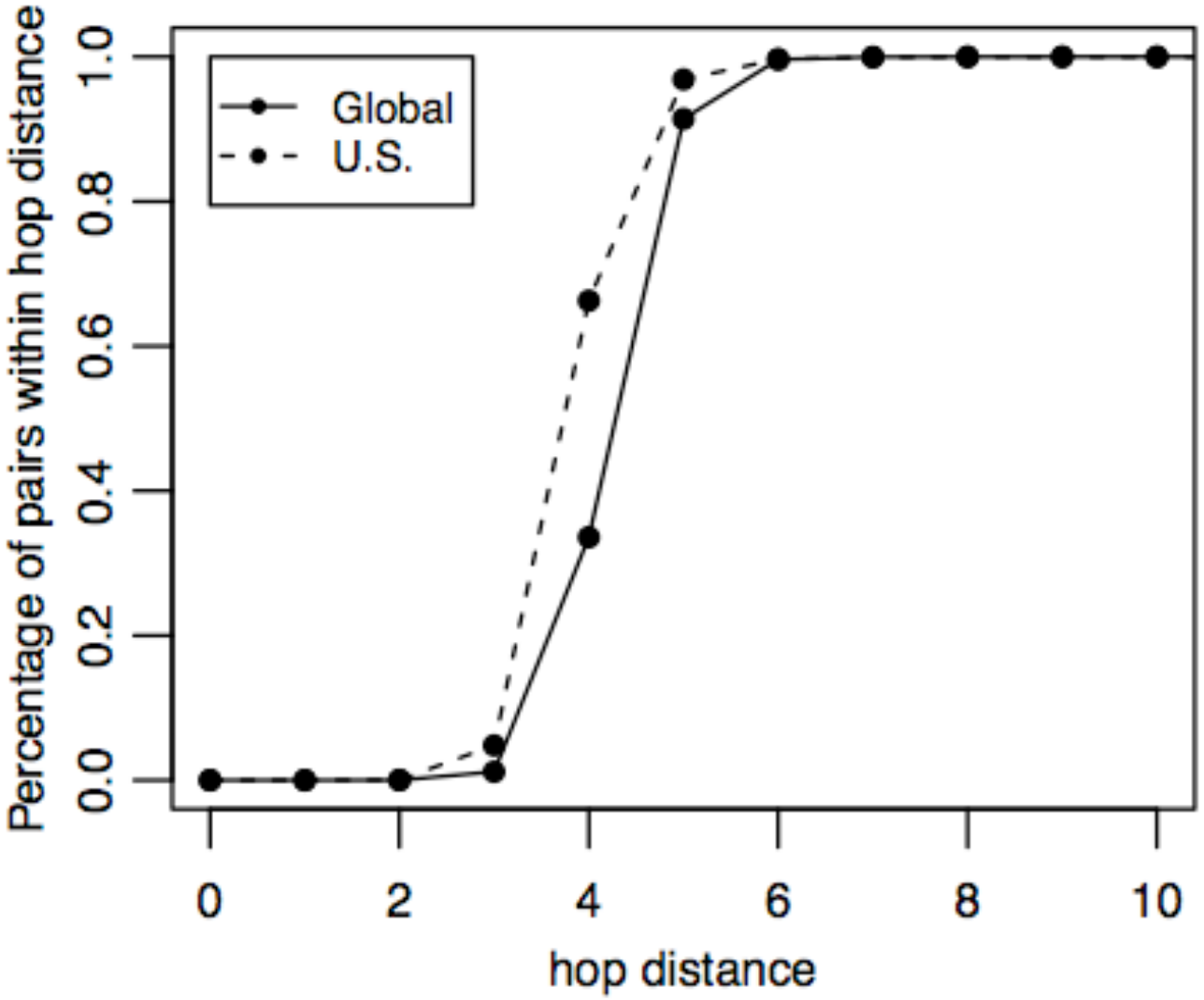
Degree distribution of the facebook network



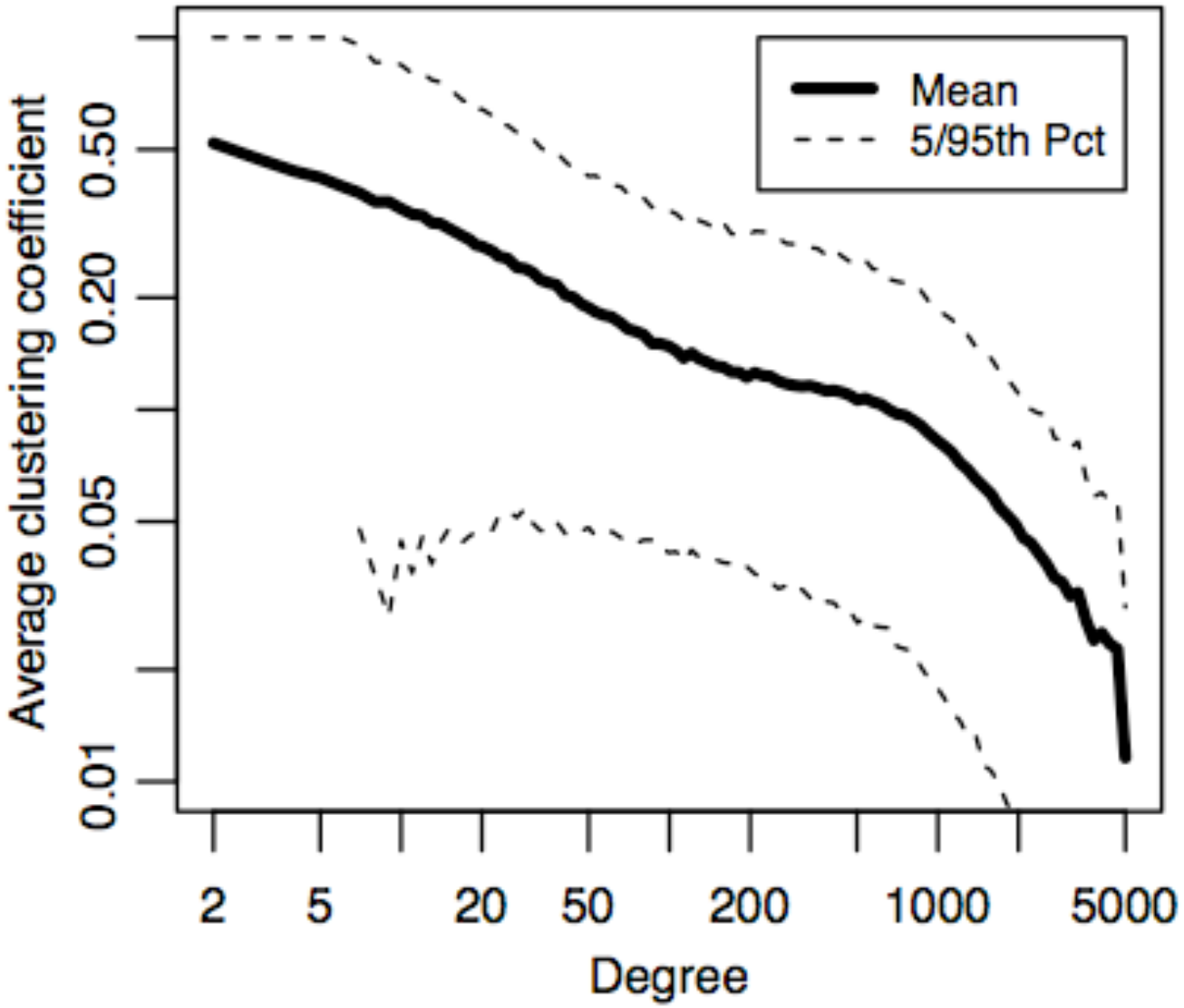
Components


Number of components
1e+00
1e+02
1e+04
1e+06
1e+00
1e+02
1e+04
1e+06
1e+08
Component size

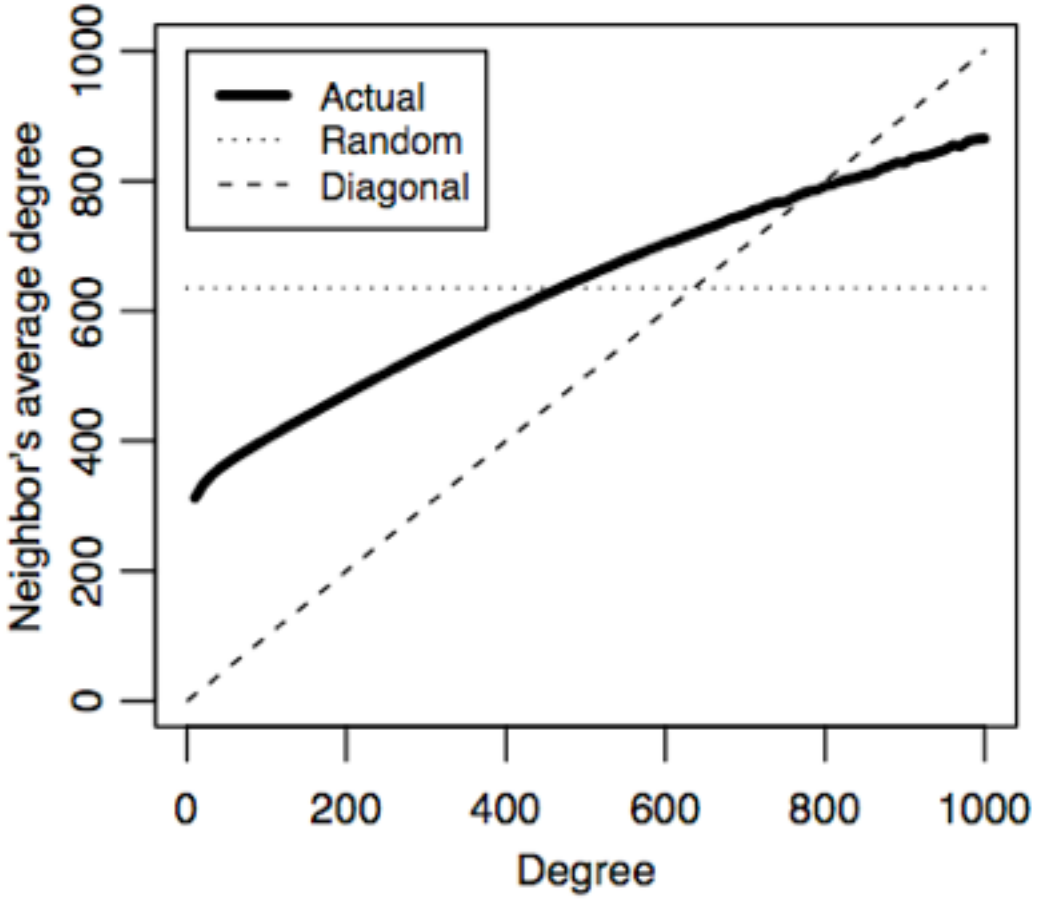
A small-world network

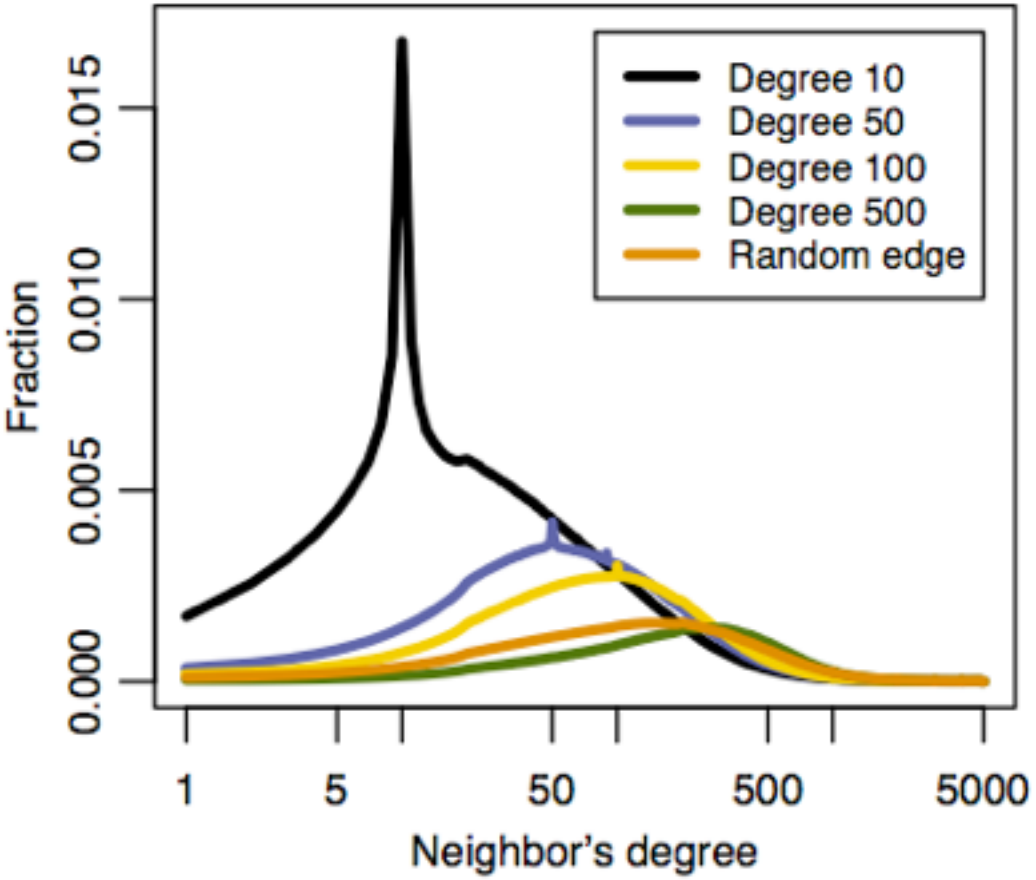


Clustering spectrum

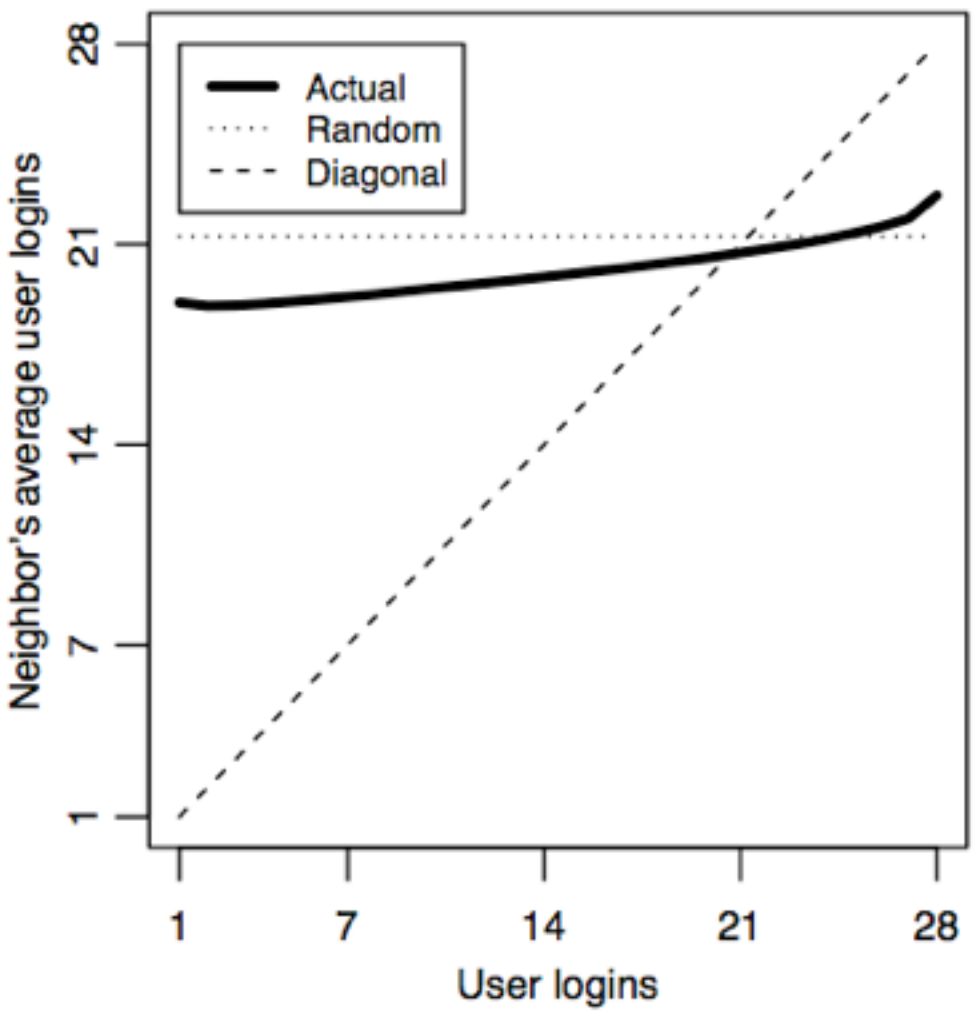


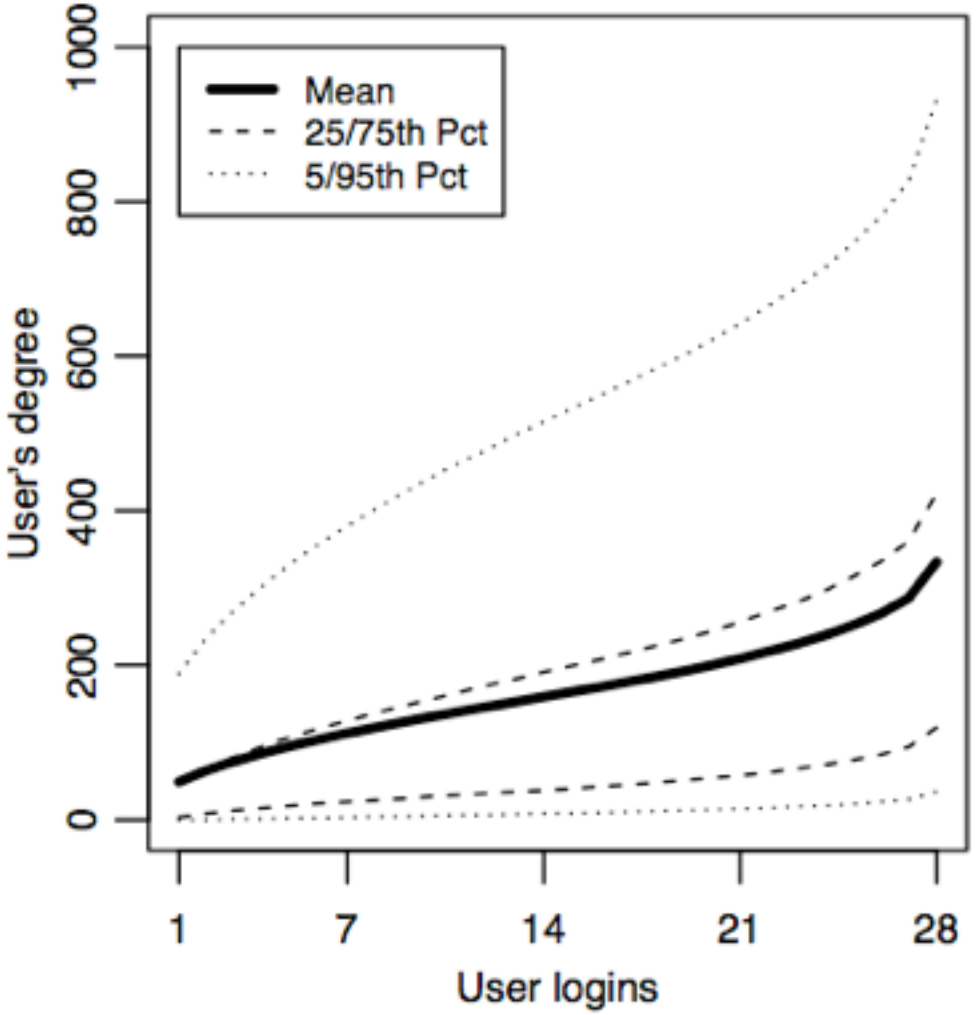
Degree correlations





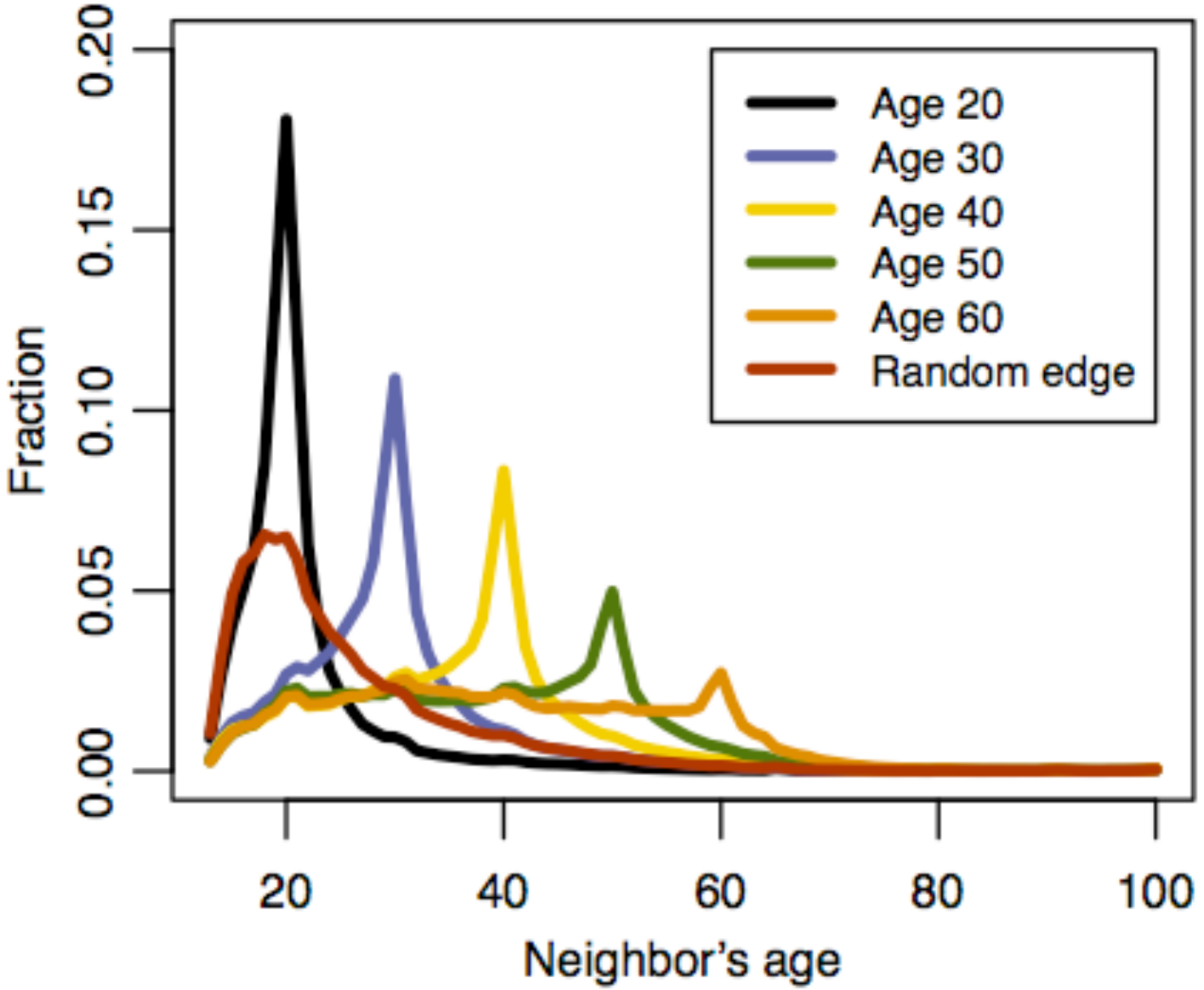
Activity-degree correlations



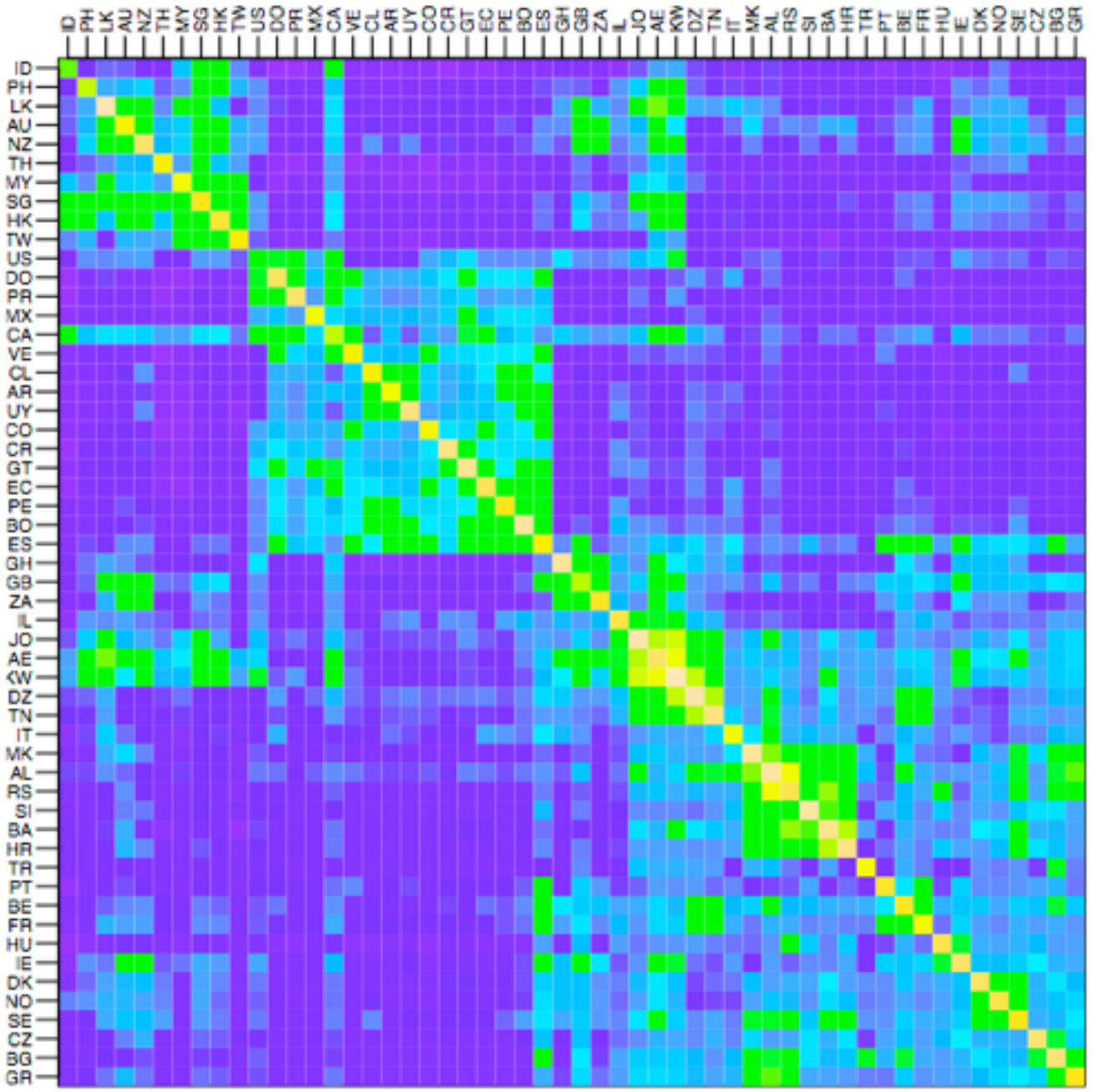


(logins during 28 days)

Age homophily



Geographic homophily



-84% of edges within country

-Modularity=0.75 when clustering by country

Influence in facebook

The Role of Social Networks in Information Diffusion, arxiv:1201.4145

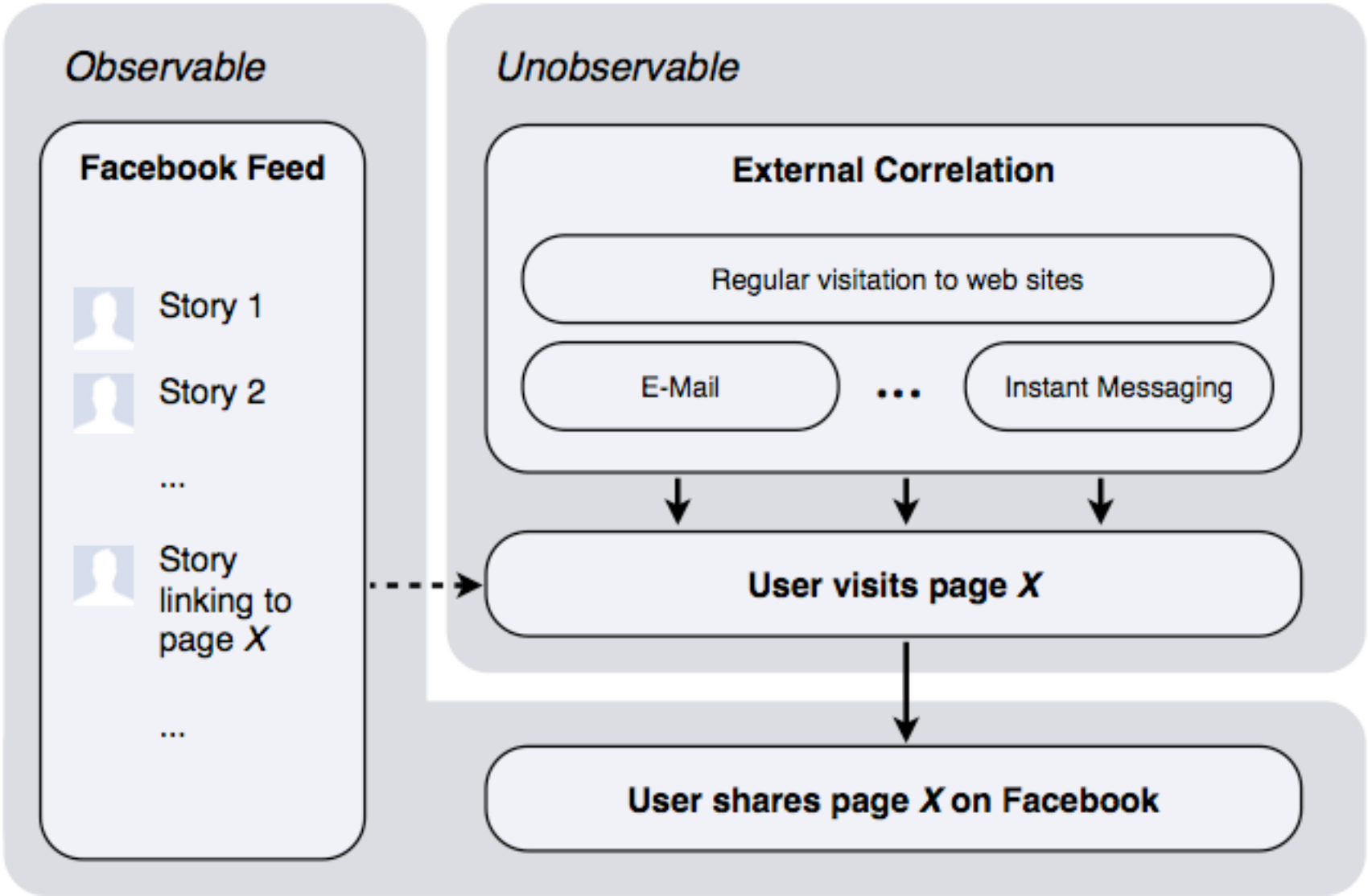
Assume the following scenario:

1. user U exposes a web page X on facebook
2. user V, *friend of U*, exposes *at a later time* X on facebook

Question: was V influenced by U?

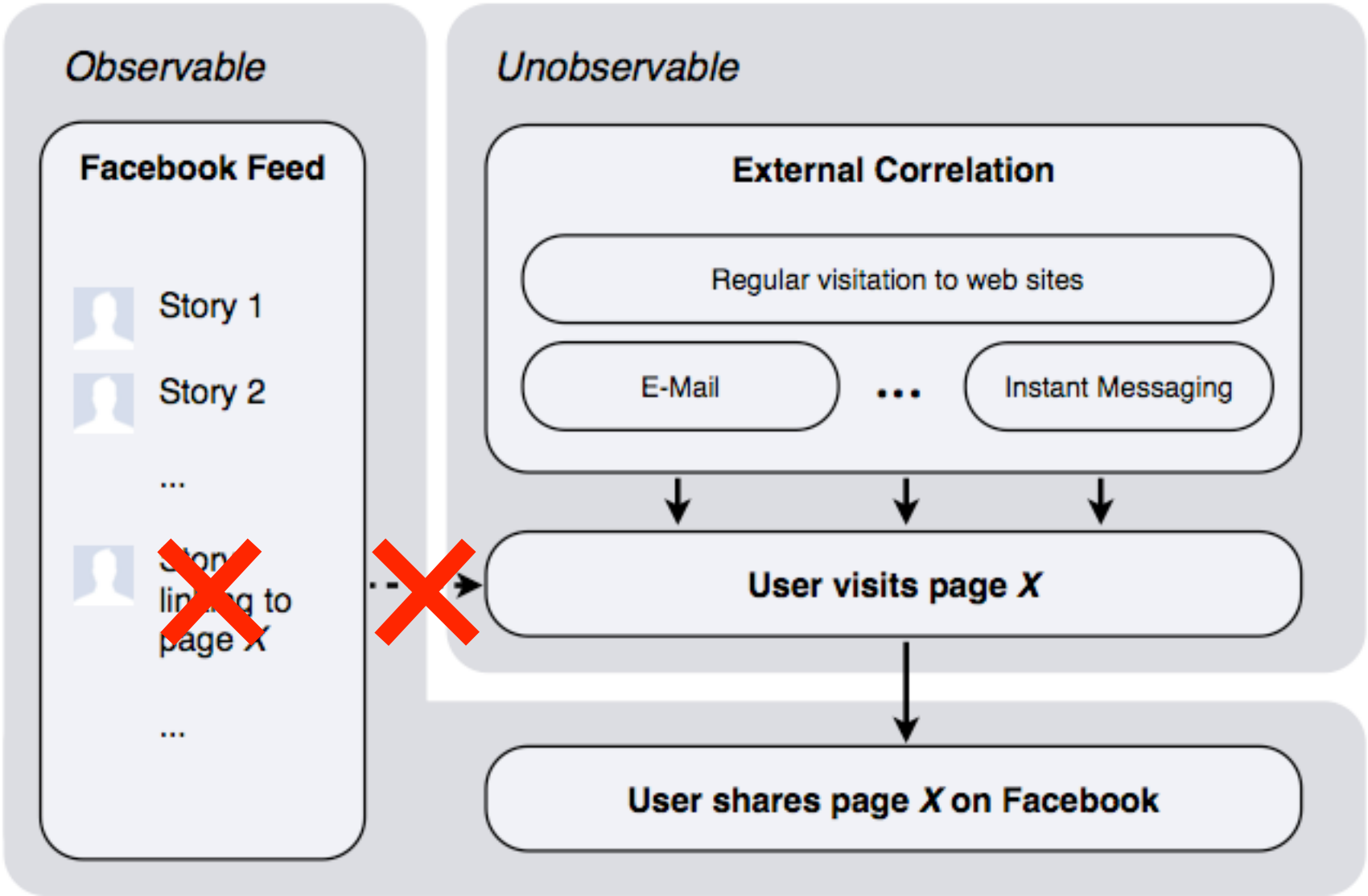
Why is that not obvious?

confounding factors



Controlled experiment:

- *suppress the exposure to X on facebook at random*
- *compare probability for V to share X*
  - *when exposed on facebook*
  - *when not exposed on facebook*

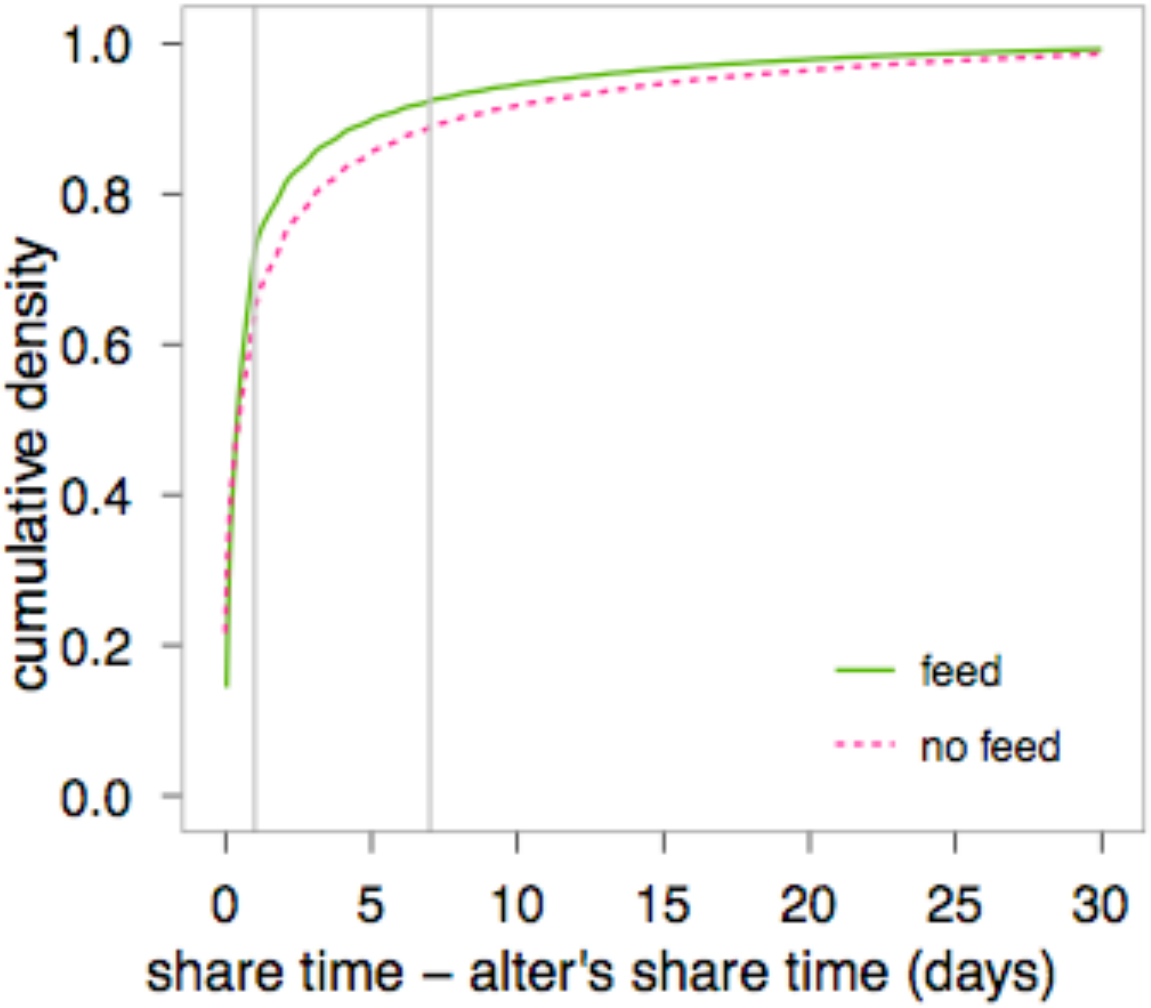


experimental design

feed

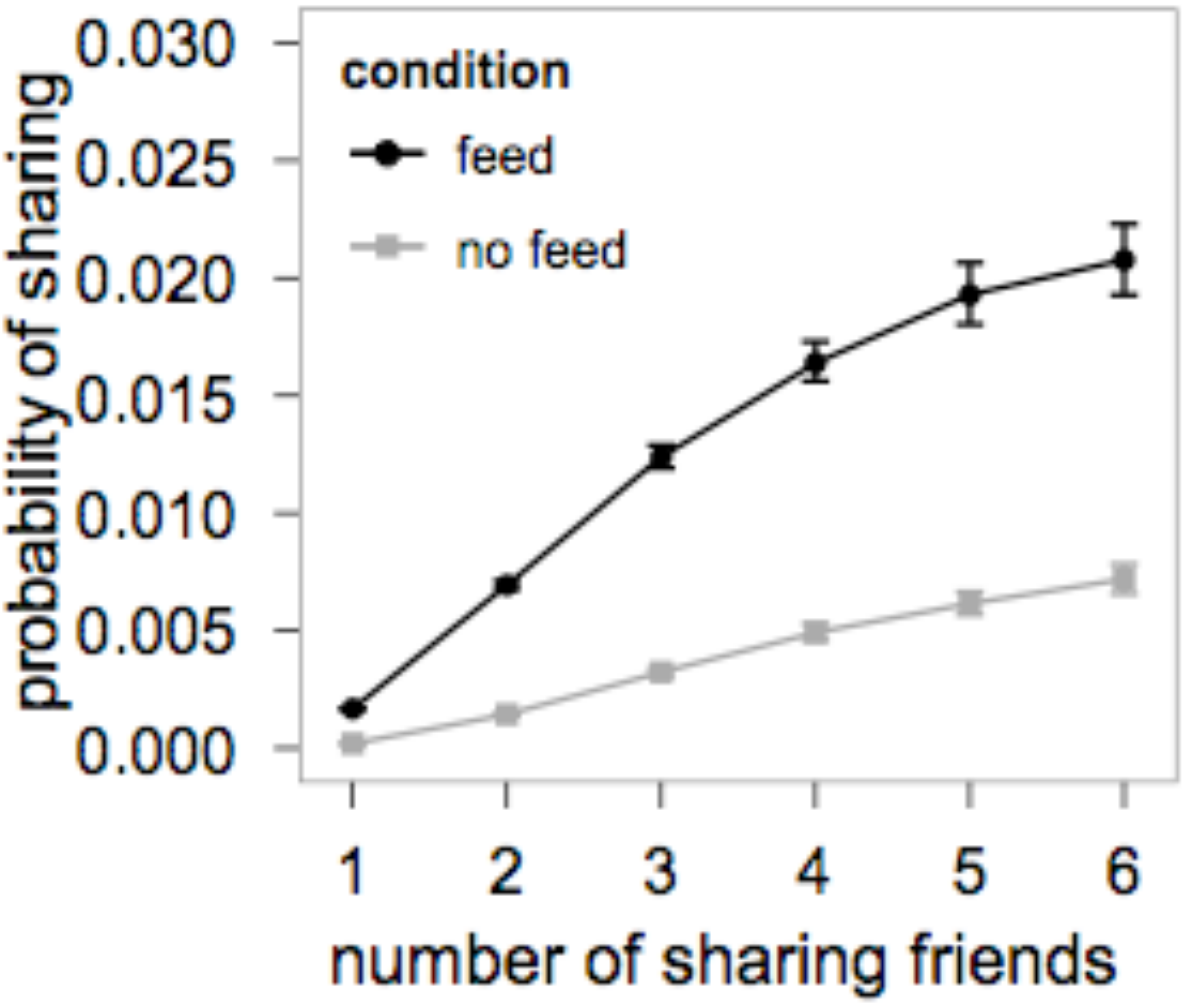
no-feed

Results

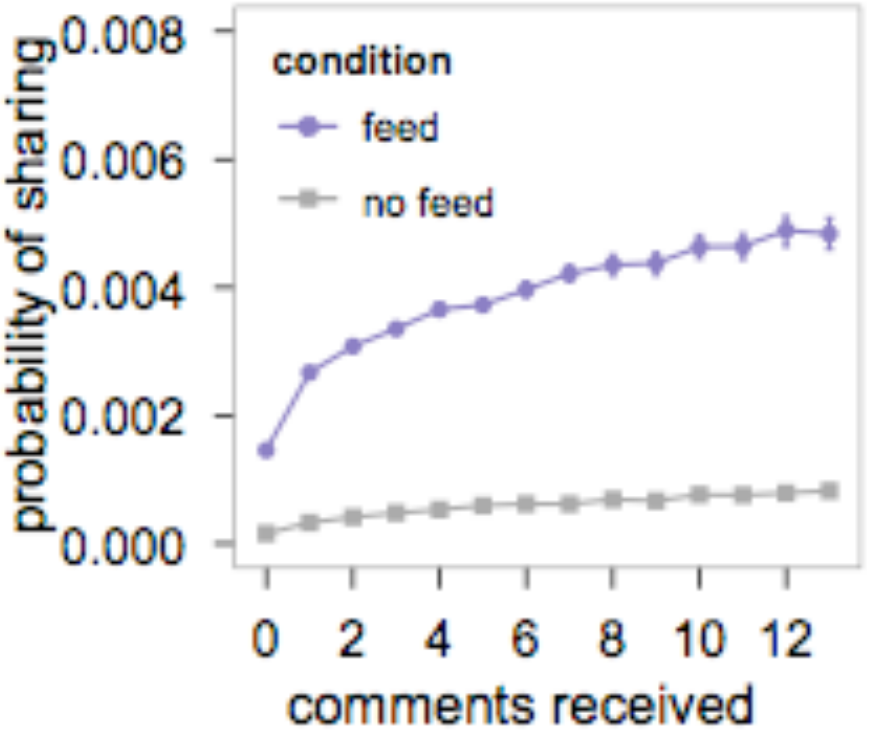


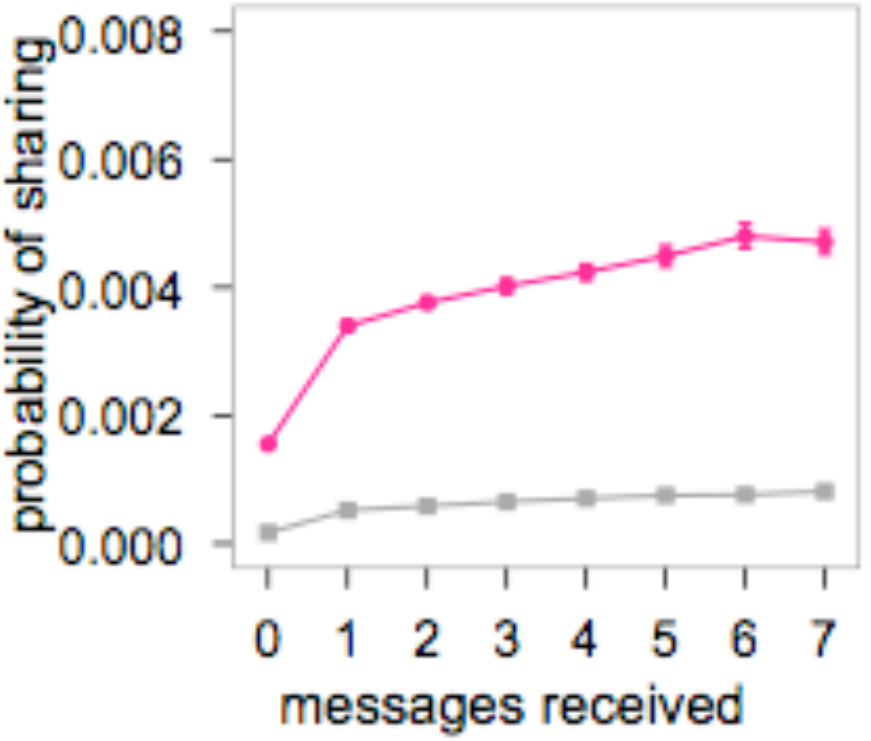
Time difference between time at which a user shares and the time of the first sharing friend

Results


probability of sharing
0.030
0.025
0.020
0.015
0.010
0.005
0.000
condition
feed
no feed
1
2
3
4
5
6
number of sharing friends

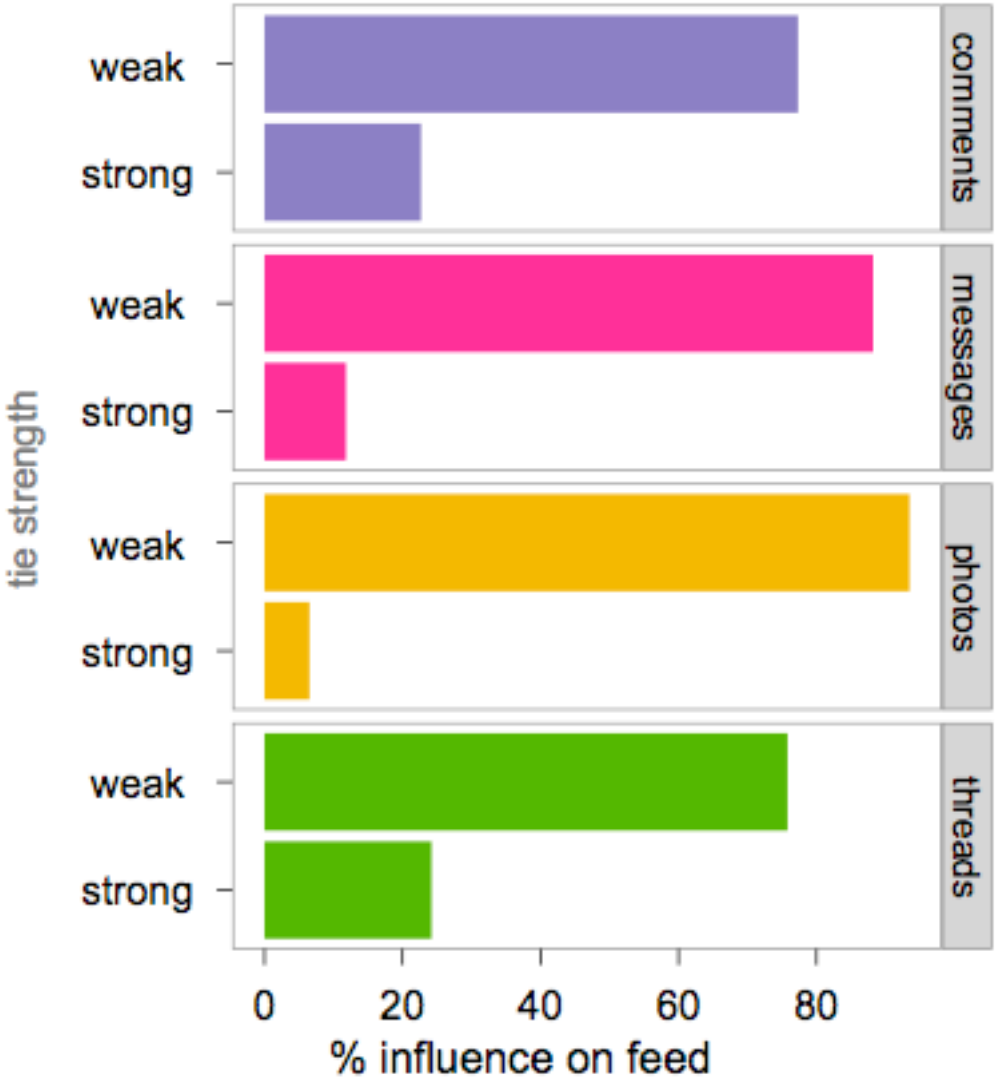
Results





Stronger ties carry more influence

Results



weak ties are collectively more influential

it’s complicated (but interesting!)