

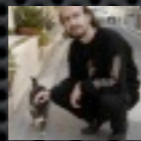
Folks in Folksonomies:

Social Link Prediction from Shared Metadata

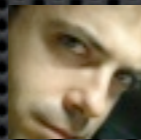
Rossano Schifanella



Alain Barrat



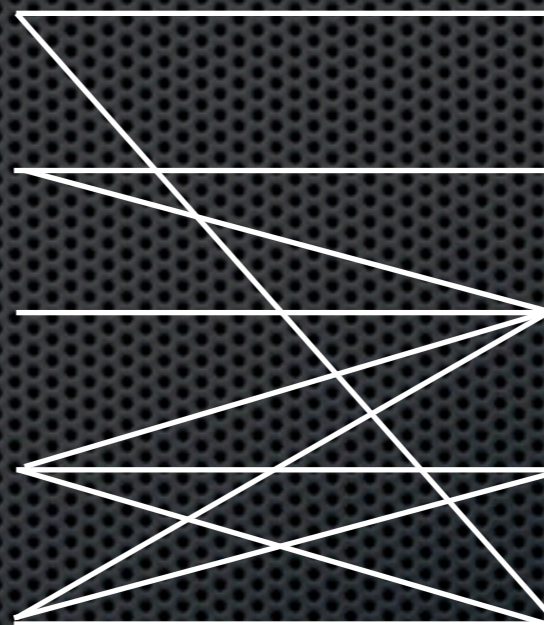
Ciro Cattuto



Benjamin Markines



Filippo Menczer



Department of Computer Science,
University of Torino

Centre de Physique Théorique, Marseille

Complex Networks & Systems
Laboratory, ISI Foundation, Torino

School of Informatics and Computing,
Indiana University

Center for Complex Networks &
Systems Research, Indiana University

Danah Boyd on **homophily**

Danah Boyd on **homophily**

*In a networked world, **people connect to people like themselves. What flows across the network flows through edges of similarity...***

I interviewed gay men who thought Friendster was a gay dating site because all they saw were other gay men. I interviewed teens who believed that everyone on MySpace was Christian because all of the profiles they saw contained biblical quotes...

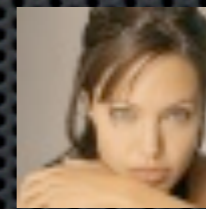
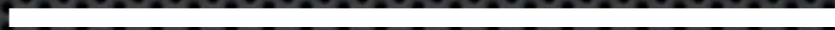
In an era of networked media, we need to recognize that networks are homophilous and operate accordingly. Technology does not inherently disintegrate social divisions. In fact, more often than not, it reinforces them...

*“Streams of Content, Limited Attention: The Flow of Information through Social Media”
Web2.0 Expo. New York, November 2009*

- Given two users, how does the alignment of their tag vocabularies relate to their proximity on the social network?

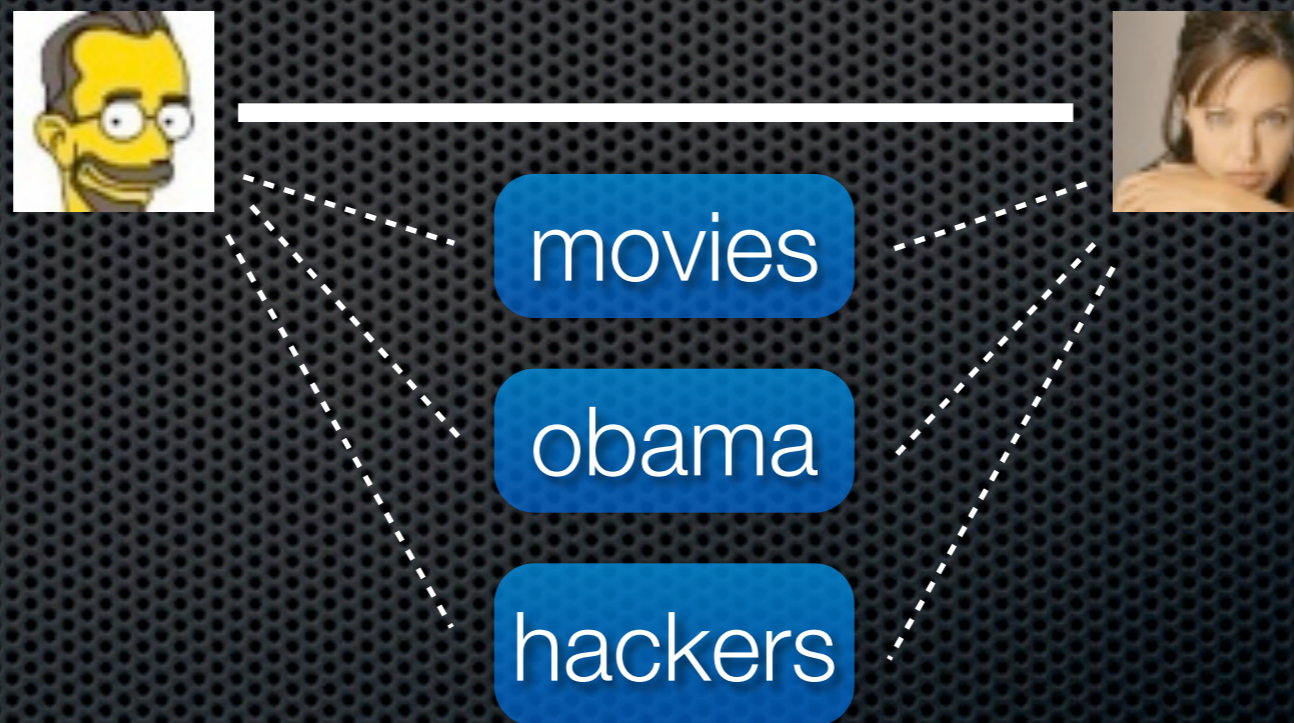
- Given two users, how does the alignment of their tag vocabularies relate to their proximity on the social network?

1. Lexical alignment between social neighbors?



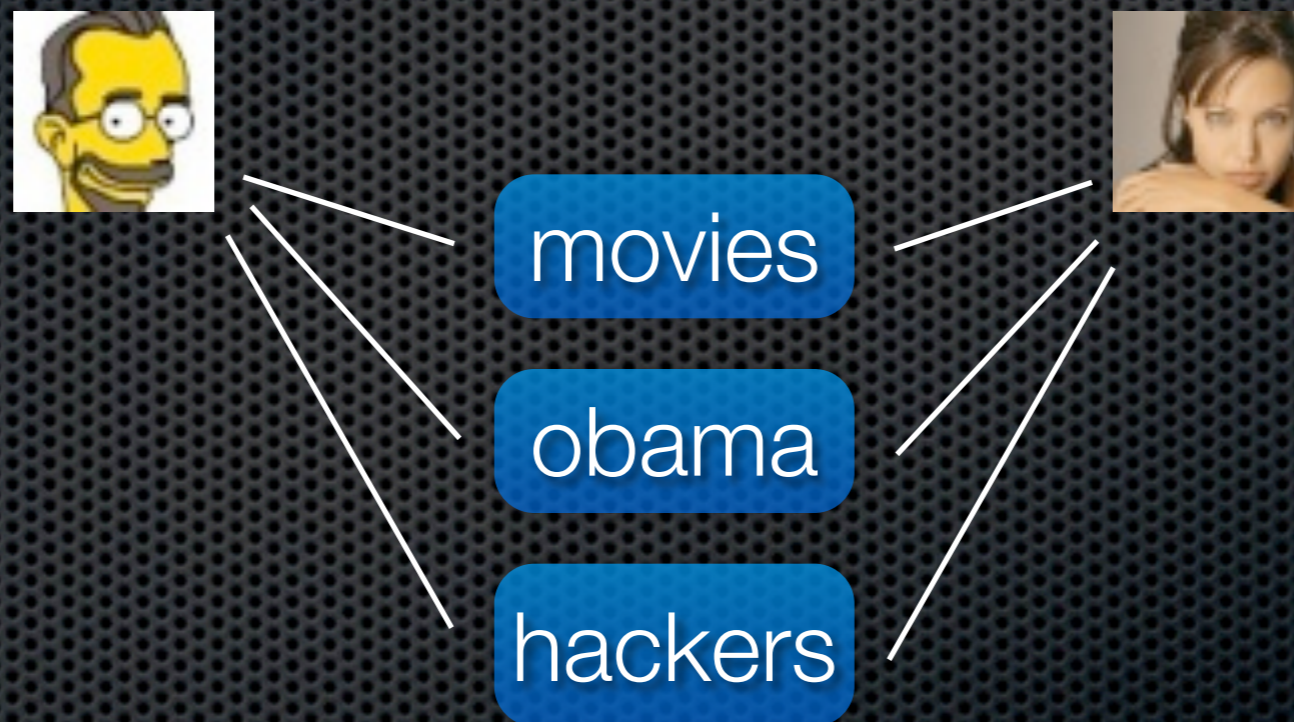
- Given two users, how does the alignment of their tag vocabularies relate to their proximity on the social network?

1. Lexical alignment between social neighbors?

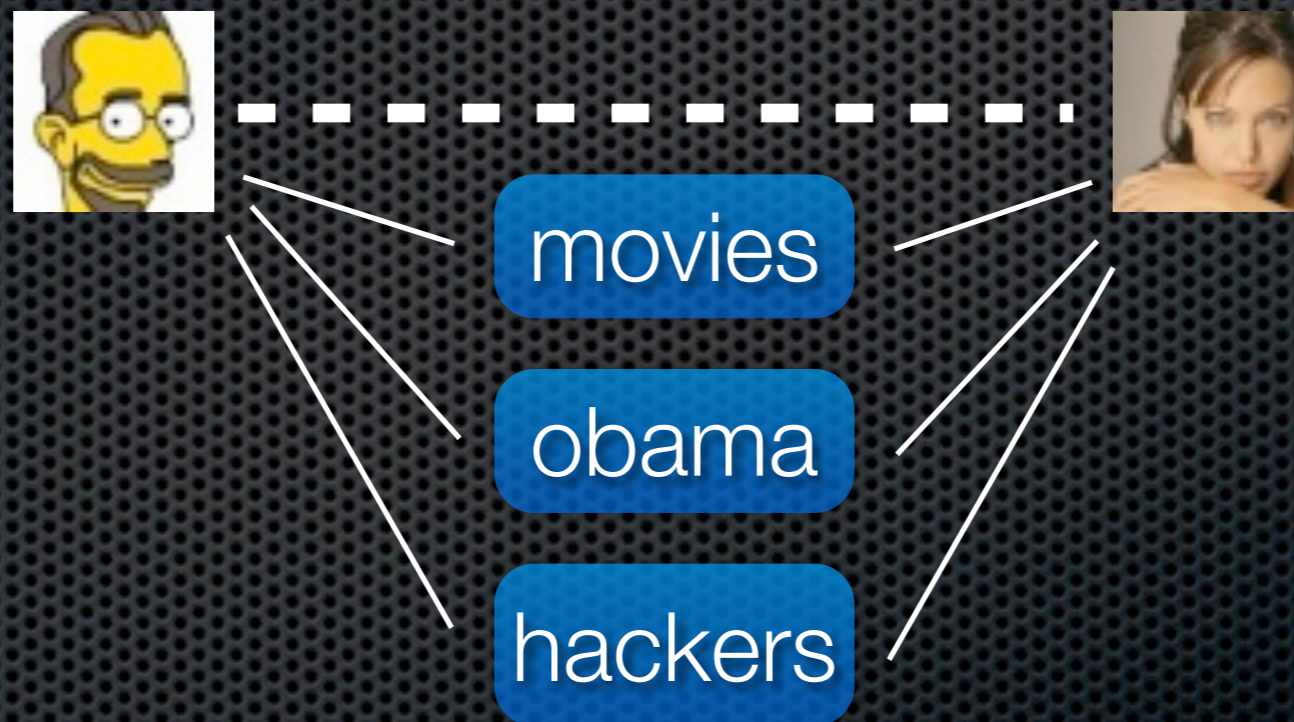


2. Predict social links from analysis of similarity, extracted from annotations?

2. Predict social links from analysis of similarity, extracted from annotations?



2. Predict social links from analysis of similarity, extracted from annotations?



Data sets: tags + social links

✦ **Flickr (“narrow” folksonomy)**

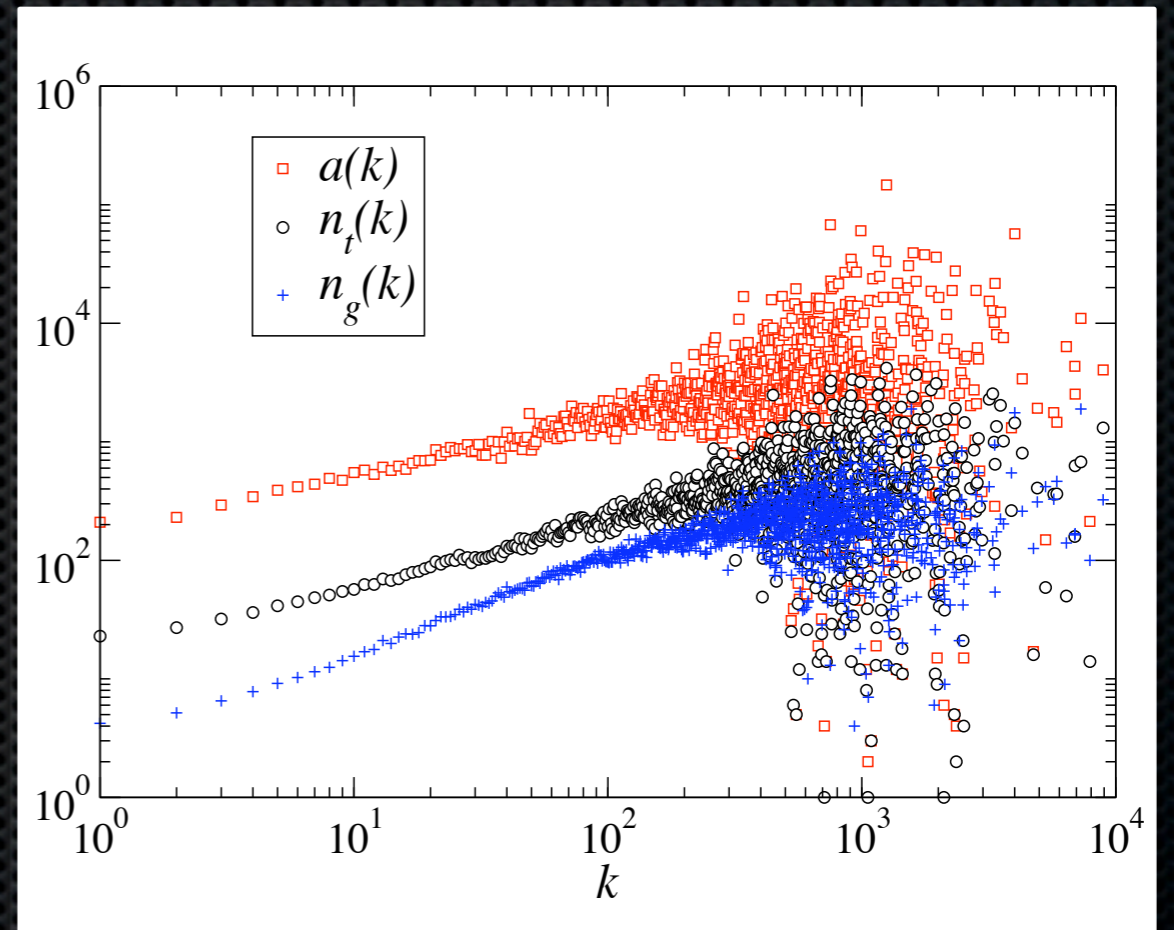
- ✦ Content from Jan 2004 – Jan 2006
- ✦ API crawl (2007) based on photos and tags
- ✦ Separate crawler for “contacts” and groups
- ✦ G0: 118K users, 2.2M edges; complete tag/grp/contact info
- ✦ G1: 984K users, 16.7M edges; neighbors added

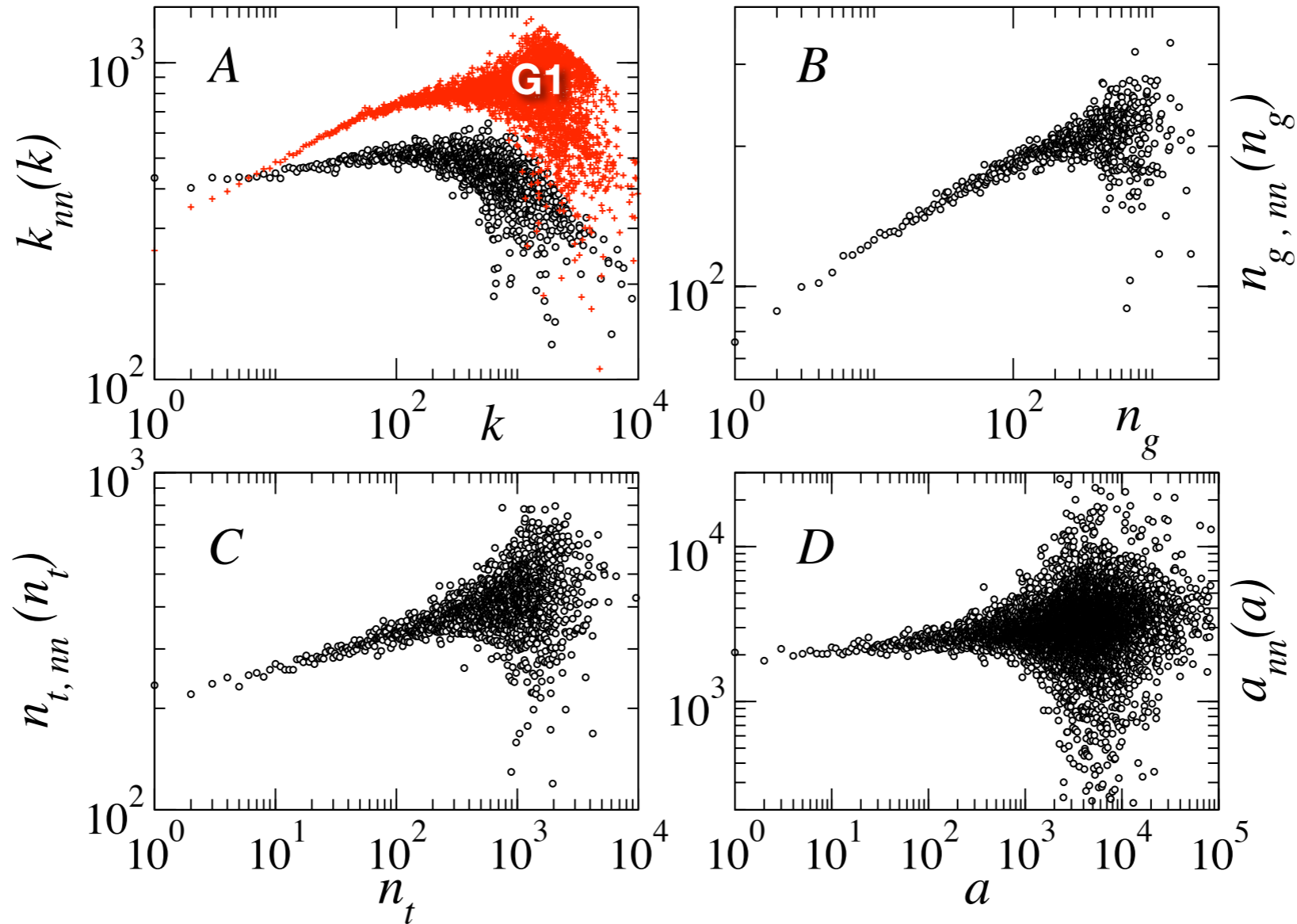
✦ **Last.fm (“broad” folksonomy)**

- ✦ API crawl for “neighbors” and “friends” (2009)
- ✦ Separate crawler for triples and groups
- ✦ 100K users (52K active), 11M triples, 1.4M items, 282K tags, 66K groups
- ✦ smithers.cs.indiana.edu/data/last.fm

Correlations

	n_t	n_g	a
k	0.349	0.482	0.268
n_t		0.429	0.753
n_g			0.304





Mixing patterns
assortative activity trends

Lexical & topical similarity

- Focus on **local alignment** (among social neighbors)

Lexical & topical similarity

- Focus on **local alignment** (among social neighbors)



kitchen kushishabu la library littletokyo lomo love
li meant mom museum nature newborn not1000 OC ocean
id polaroid portrait puppy reading red sad santabarbara
sleep summer sun sunset sunshine sx70 tattoo the
tiger toes tree trip usa vacation wild woman wool wow yard

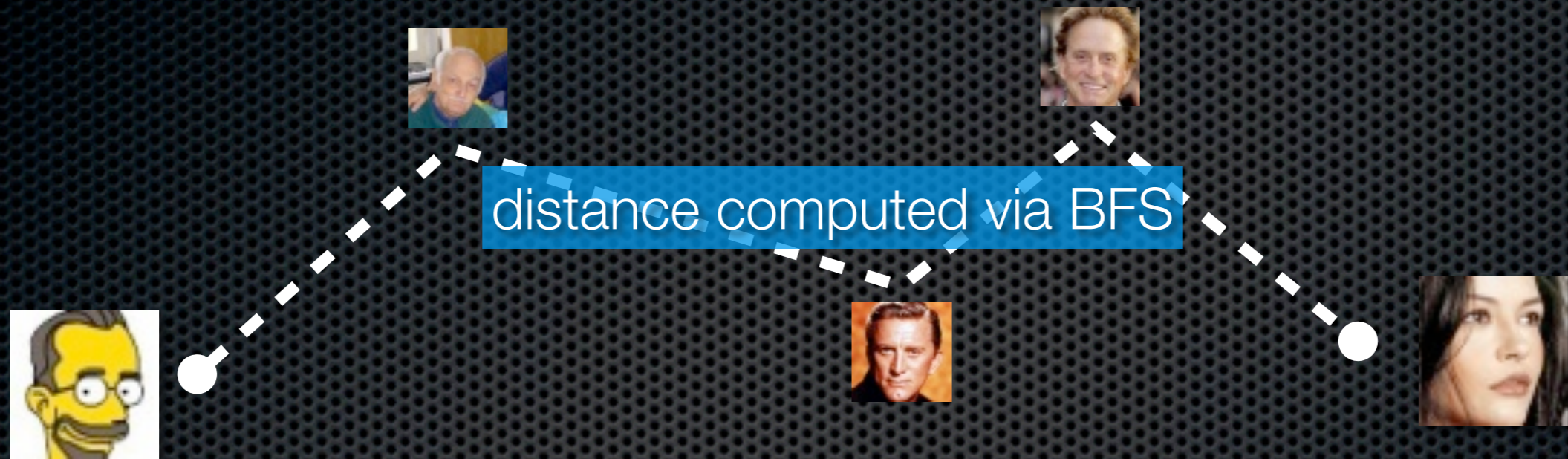
line lines lips little living look love makeup man
myway mywinner nature naturesfinest new night nikon nikor
outstandingshots park pencil people pink play portrait portret red
roumanie russian schita sea searchthebest self sexy shieldofexcellenc
smile soe spring street summer sun sunset superbmasterpiece supershot
ultimateshot up water white woman women xoxoxo yellow you young



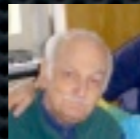
Similarity vs. social distance



Similarity vs. social distance



Similarity vs. social distance



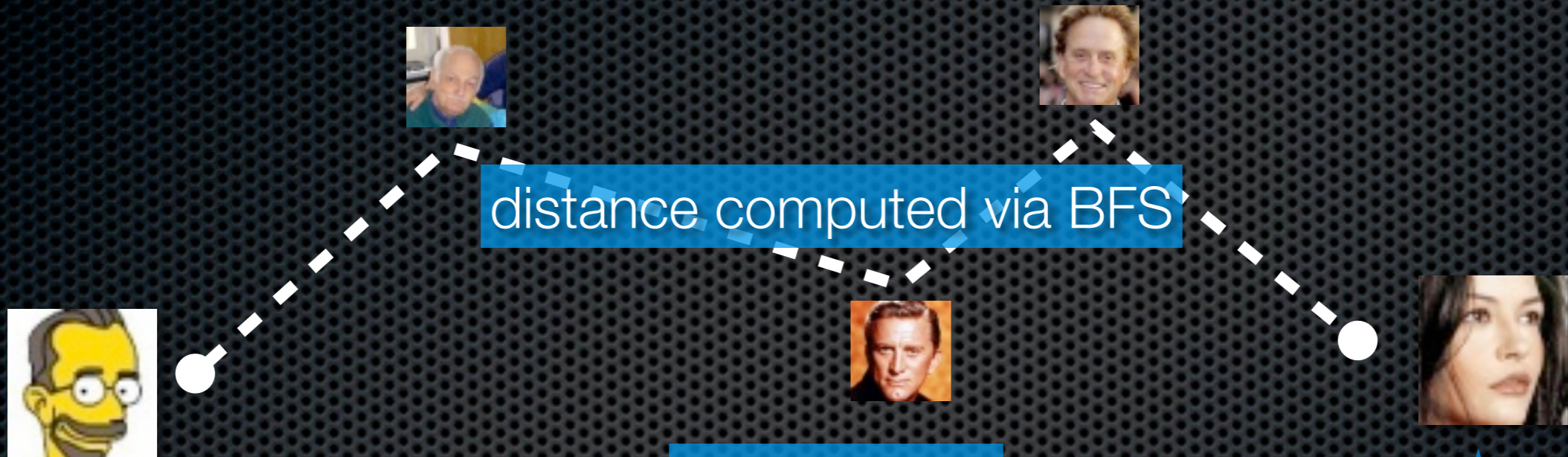
distance computed via BFS

similarity
computed via
matching or
cosine

line lines lips little living look love makeup man me memories mountains
myway mywinner nature naturesfinest new night nikon nikor
outstandingshots park pencil people pink play portrait portret red
roumanie russian schita sea searchthebest self sexy shieldofexcellenc
smile soe spring street summer sun sunset superbmasterpiece supershot
ultimateshot up water white woman women xoxoxo yellow you young

kitchen kushishabu la library littletokyo lomo love
ui meant mom museum nature newborn not1000 oc ocean
mid polaroid portrait puppy reading red sad santabarbara
sleep summer sun sunset sunshine sx70 tattoo the
tiger toes tree trip usa vacation wild woman wool wow yard

Similarity vs. social distance



line lines lips little living look love makeup man me memories mountains
 myway mywinner nature naturesfinest new night nikon nikor
 outstandingshots park pencil people pink play portrait portret red
 roumanie russian schita sea searchthebest self sexy shieldofexcellenc
 smile soe spring street summer sun sunset superbmasterpiece supershot
 ultimateshot up water white woman women xoxoxo yellow you young

similarity
 computed via
 matching or
 cosine

kitchen kushishabu la library littletokyo lomo love
 ui meant mom museum nature newborn not1000 oc ocean
 id polaroid portrait puppy reading red sad santabarbara
 sleep summer sun sunset sunshine sx70 tattoo the
 tiger toes tree trip usa vacation wild woman wool wow yard

$$\sigma_{tags}(u, v) = \frac{\sum_t f_u(t) f_v(t)}{\sqrt{\sum_t f_u(t)^2} \sqrt{\sum_t f_v(t)^2}}$$

lexical: shared tags

topical: shared groups

$$\sigma_{groups}(u, v) = \frac{\sum_g \delta_u^g \delta_v^g}{\sqrt{n_g(u) n_g(v)}}$$

Correlation \neq Causation

- ✦ We expect a purely statistical (spurious) correlation just because of assortative biases

Correlation \neq Causation

- ✦ We expect a purely statistical (spurious) correlation just because of assortative biases
- ✦ Need a proper **null model**:
 - ✦ Same social structure
 - ✦ Shuffle tags and groups, preserving \mathbf{k} , \mathbf{n}_t , \mathbf{n}_g , and \mathbf{a}
 - ✦ No local topical or lexical alignment other than from statistical mixing patterns



dog

dog

cat

$$\mathbf{n}_t = 2$$
$$\mathbf{a} = 3$$



cat

fish

$$\mathbf{n}_t = 2$$
$$\mathbf{a} = 2$$

Correlation \neq Causation

- ✦ We expect a purely statistical (spurious) correlation just because of assortative biases
- ✦ Need a proper **null model**:
 - ✦ Same social structure
 - ✦ Shuffle tags and groups, preserving \mathbf{k} , \mathbf{n}_t , \mathbf{n}_g , and \mathbf{a}
 - ✦ No local topical or lexical alignment other than from statistical mixing patterns



dog

cat

cat

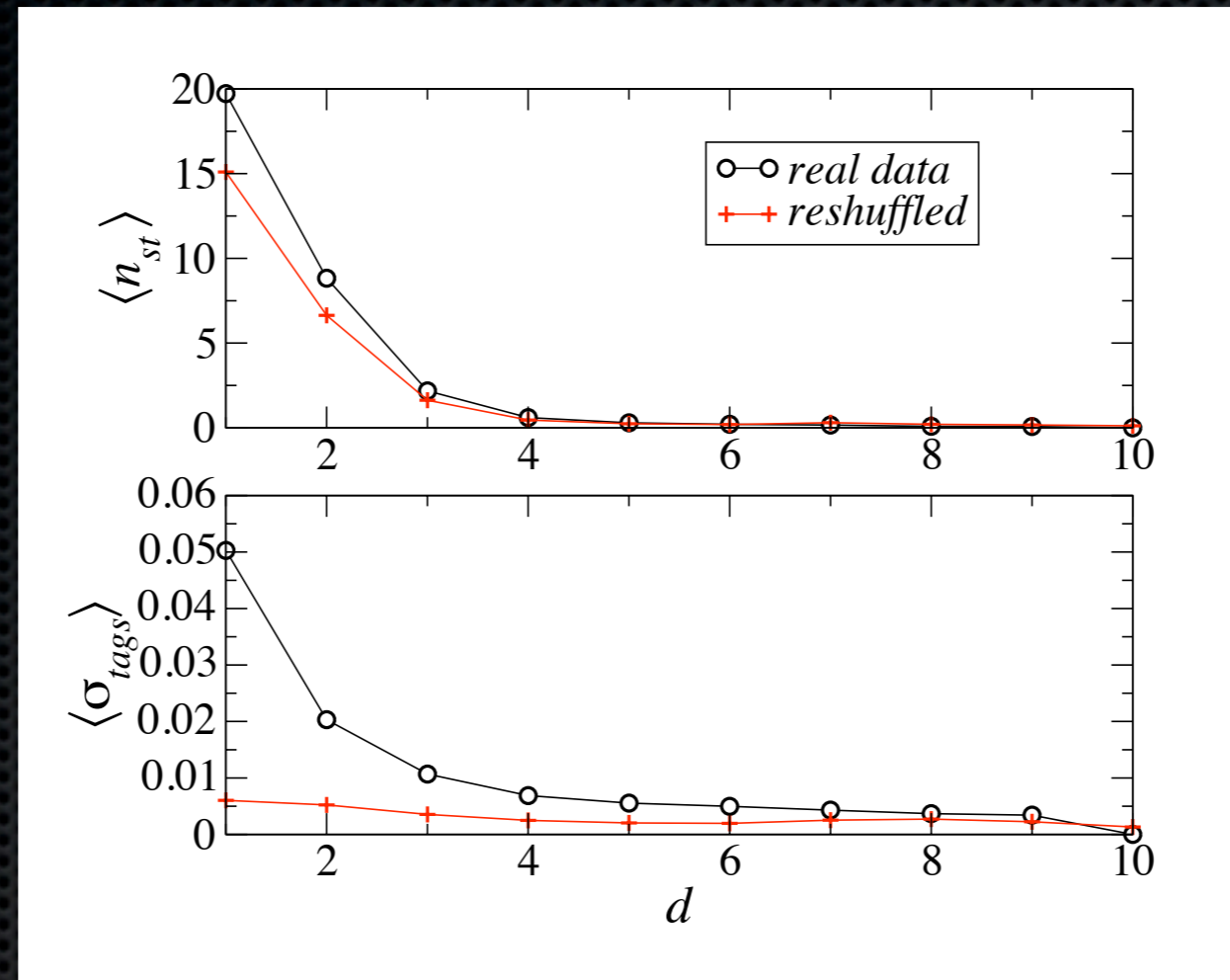
$$\mathbf{n}_t = 2$$
$$\mathbf{a} = 3$$



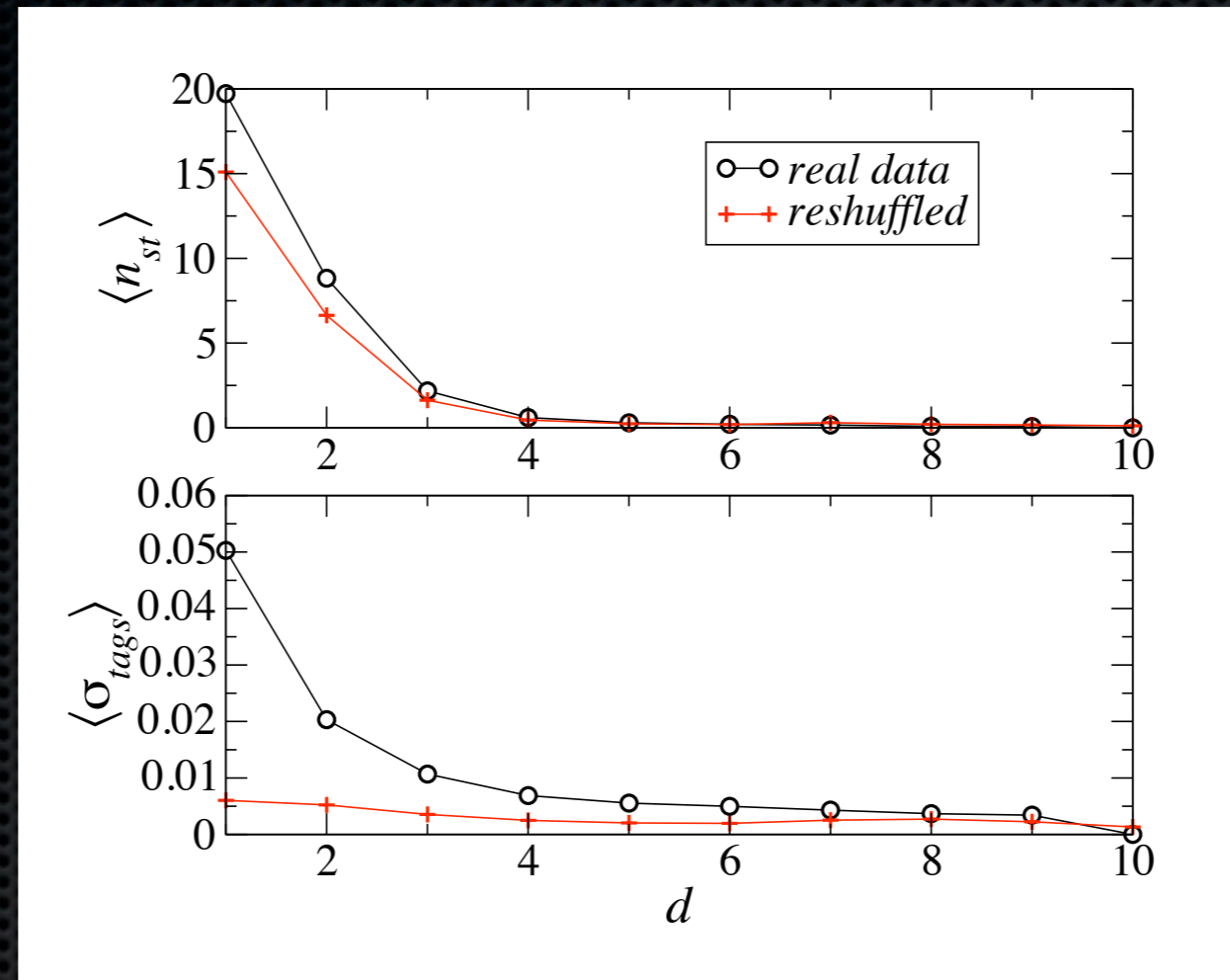
dog

fish

$$\mathbf{n}_t = 2$$
$$\mathbf{a} = 2$$



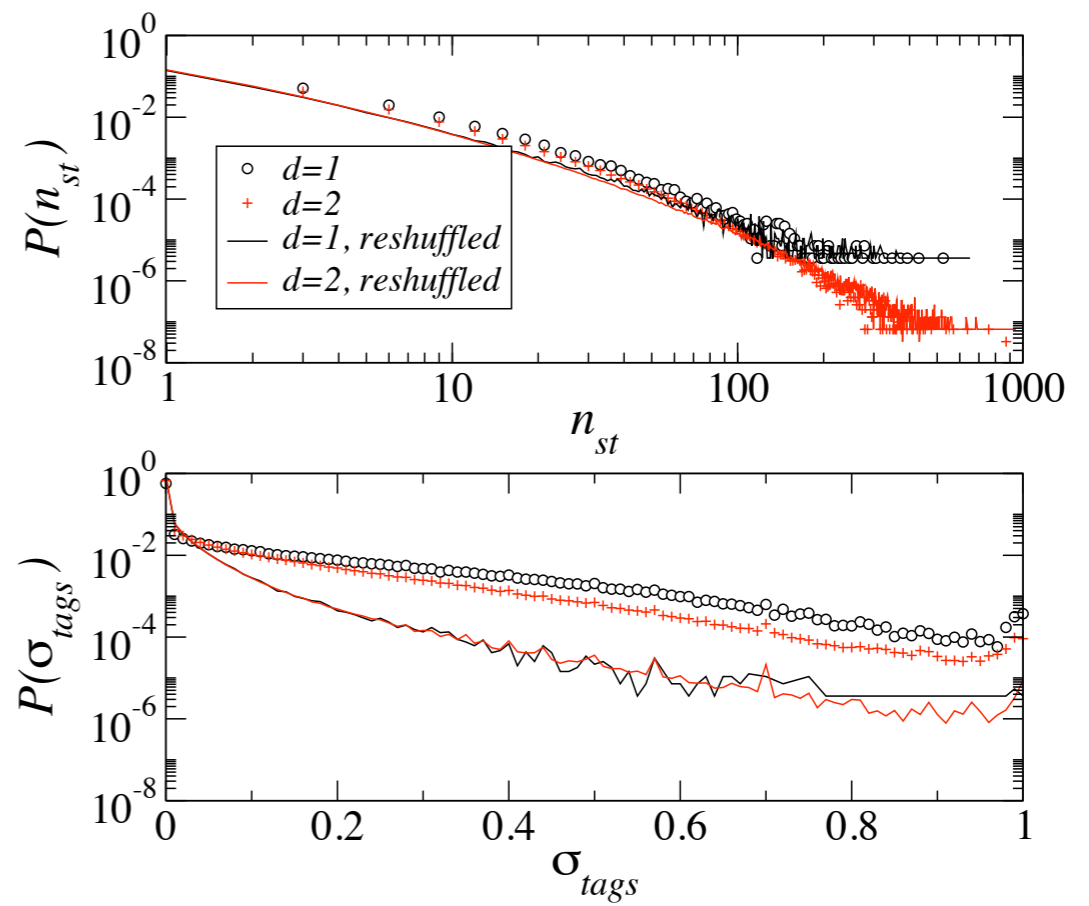
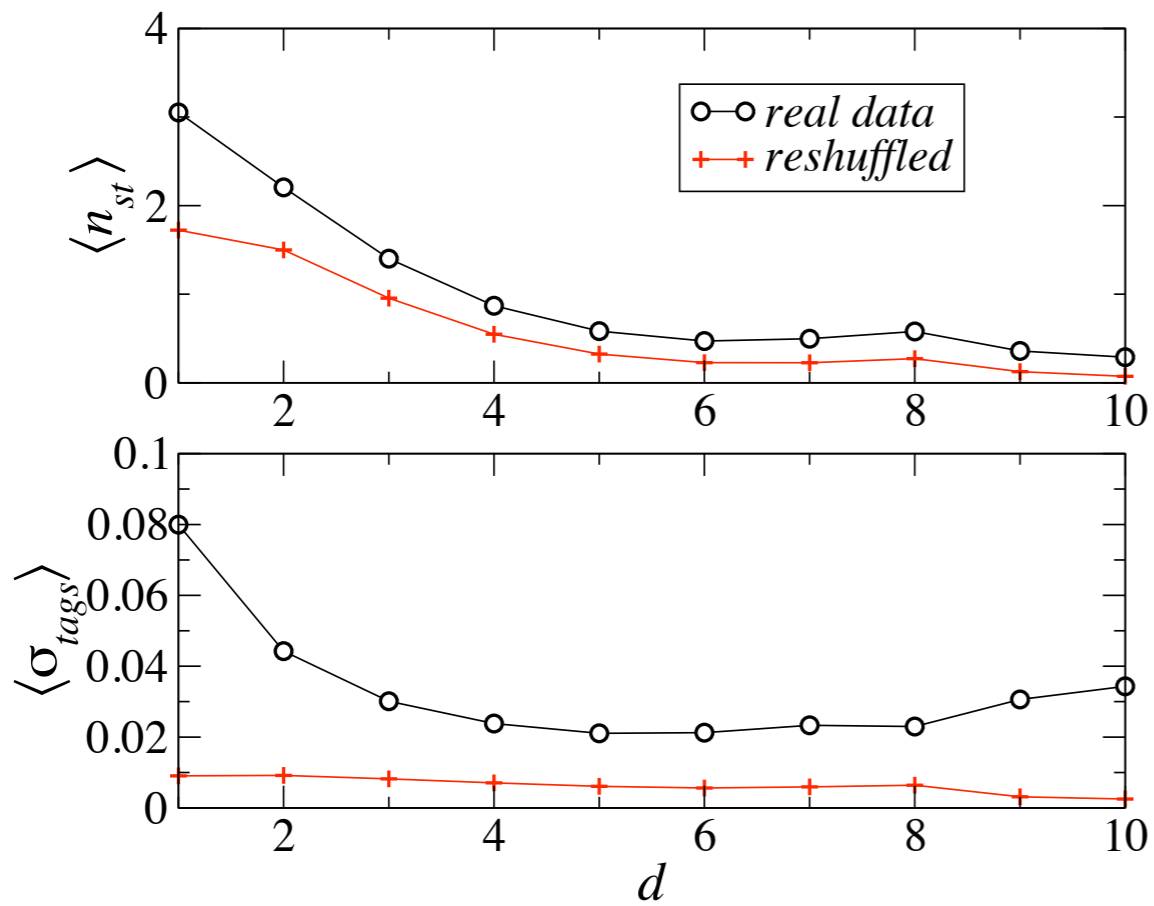
Similarity vs. social distance



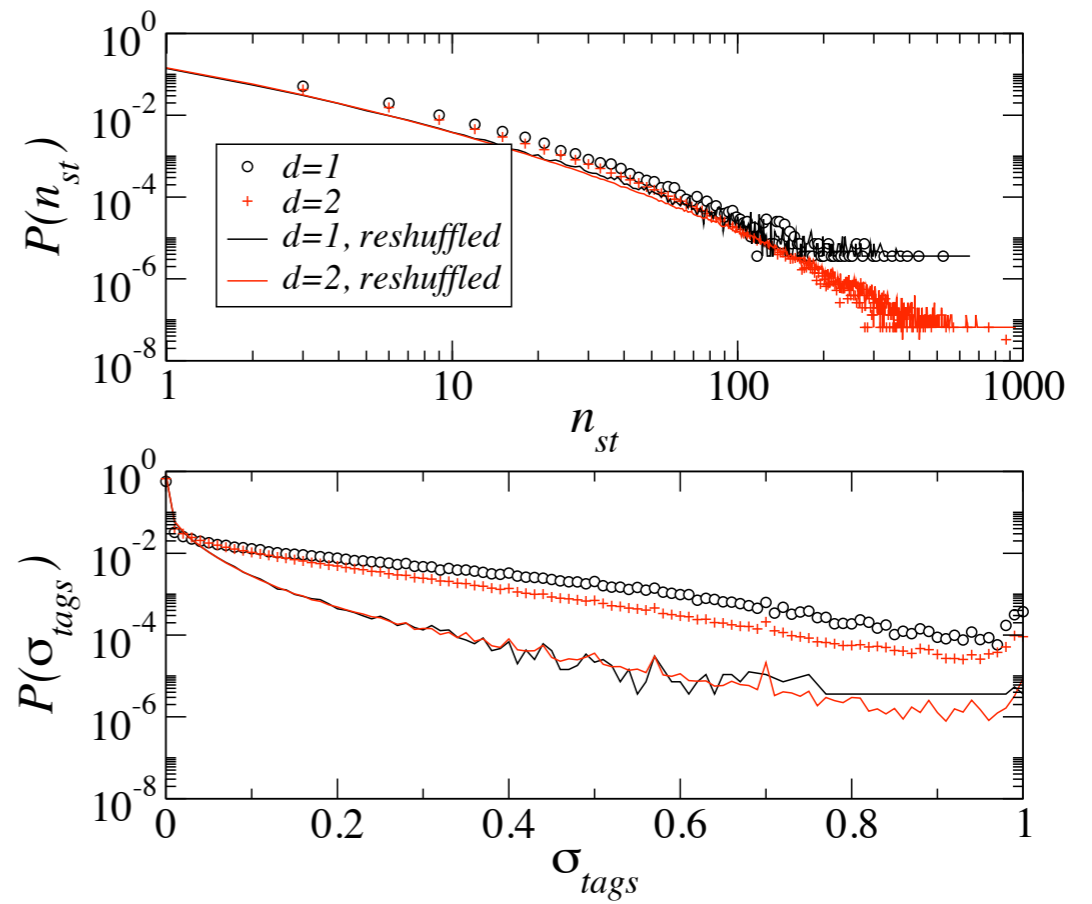
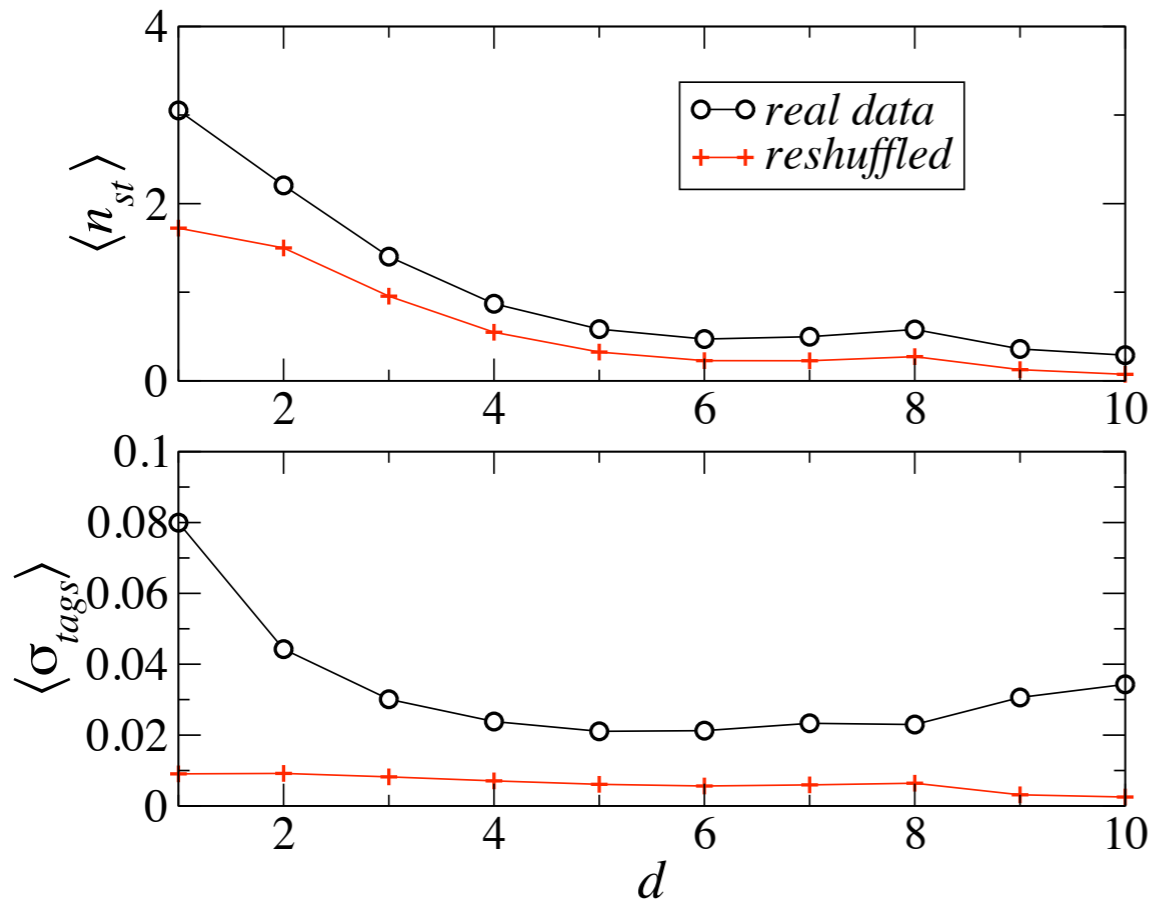
Similarity vs. social distance

local lexical alignment is real (using cosine)

local topical alignment is weaker



Similarity vs. social distance



Similarity vs. social distance

Last.fm results similar to Flickr

Part 2

- **Can we predict social links from lexical similarity?**
 - Semantic similarity measures based on annotations (Markines et al. HT'08, WWW'09, HT'09)
 - Information-theoretic extensions of various similarity measures, such as *Jaccard*, *Dice*, *cosine*, etc.
 - 2 aggregation methods: *distributional* and *collaborative*
 - User-user?

Social link prediction

- ✦ Both Flickr and Last.fm data sets allow to test prediction by comparing with explicit social links
 - ✦ Similar results

Social link prediction

- ✦ Both Flickr and Last.fm data sets allow to test prediction by comparing with explicit social links
 - ✦ Similar results
- ✦ Let us focus on Last.fm data set
 - ✦ “Broad” folksonomy
 - ✦ Stronger baseline: **neighbor** recommendations

Maximum Information Path

line lines lips little living look love makeup man me memories mountains
myway mywinner nature naturesfinest new night nikon nikor
outstandingshots park pencil people pink play portrait portret red
roumanie russian schita sea searchthebest self sexy shieldofexcellence
smile soe spring street summer sun sunset superbmasterpiece supershot
ultimateshot up water white woman women xoxoxo yellow you young

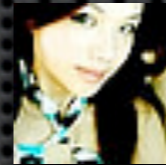
kitchen kushishabu la library littletokyo lomo love
ai meant mom museum nature newborn not1000 oc ocean
bid polaroid portrait puppy reading red sad santabarbara
sleep summer sun sunset sunshine sx70 tattoo the
tiger toes tree trip usa vacation wild woman wool wow yard



Maximum Information Path

line lines lips little living look **love** makeup man **me** memories mountains
myway mywinner nature naturesfinest new night **nikon nikon**
outstandingshots park **pencil** people pink play **portrait** portret red
roumanie russian **schita** sea searchthebest self sexy shieldofexcellence
smile **soe** spring street **summer** sun sunset superbmasterpiece **supershot**
ultimateshot up water **white** **woman** women xoxoxo yellow you young

kitchen **kushishabu** **la** library littletokyo **lomo** **love**
ai meant mom **museum** nature newborn not1000 **oc** ocean
bid **polaroid** portrait **puppy** reading red sad **santabarbara**
sleep **summer** **sun** sunset sunshine **sx70** tattoo the
tiger toes tree **trip** usa vacation wild **woman** wool wow yard



Maximum Information Path

line lines lips little living look **love** makeup man **me** memories mountains
myway mywinner nature naturesfinest new night **nikon** nikor
outstandingshots park **pencil** people pink play **portrait** portret red
roumanie russian schita sea searchthebest self sexy shieldofexcellenc
smile **soe** spring street **summer** sun sunset superbmasterpiece supershot
ultimateshot up water **white** **woman** women xoxoxo yellow you young

kitchen **kushishabu** **la** library littletokyo **lomo** **love**
ai meant mom **museum** nature newborn not1000 **oc** ocean
bid **polaroid** portrait puppy reading red sad **santabarbara**
sleep **summer** **sun** sunset sunshine **sx70** tattoo the
tiger toes tree **trip** usa vacation wild **woman** wool wow yard



$$\sigma(x_1, x_2) = \frac{2 \times \log(\min_{y \in X_1 \cap X_2} [p(y)])}{\log(\min_{y \in X_1} [p(y)]) + \log(\min_{y \in X_2} [p(y)])}$$

Semantic similarity measures

$$\text{Matching } \sigma(x_1, x_2) = -\sum_{y \in X_1 \cap X_2} \log p(y)$$

$$\text{Jaccard } \sigma(x_1, x_2) = \frac{\sum_{y \in X_1 \cap X_2} \log p(y)}{\sum_{y \in X_1 \cup X_2} \log p(y)}$$

$$\text{Dice } \sigma(x_1, x_2) = \frac{2 \sum_{y \in X_1 \cap X_2} \log p(y)}{\sum_{y \in X_1} \log p(y) + \sum_{y \in X_2} \log p(y)}$$

$$\text{Overlap } \sigma(x_1, x_2) = \frac{\sum_{y \in X_1 \cap X_2} \log p(y)}{\max[\sum_{y \in X_1} \log p(y), \sum_{y \in X_2} \log p(y)]}$$

$$\text{Cosine } \sigma(x_1, x_2) = \frac{X_1 \cdot X_2}{\|X_1\| \|X_2\|}$$

Applying semantic similarity measures to users

- Example: Max Info Path (MIP)










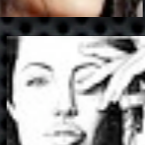
Aggregation	Distributional	Collaborative
Across items	$\frac{2 \log(\min_{t \in T_1 \cap T_2} p[t])}{\log(\min_{t \in T_1} p[t]) + \log(\min_{t \in T_2} p[t])}$	$\sum_i \frac{2 \log(\min_{t \in T_1^i \cap T_2^i} p[t i])}{\log(\min_{t \in T_1^i} p[t i]) + \log(\min_{t \in T_2^i} p[t i])}$
Across tags	$\frac{2 \log(\min_{i \in I_1 \cap I_2} p[i])}{\log(\min_{i \in I_1} p[i]) + \log(\min_{i \in I_2} p[i])}$	$\sum_t \frac{2 \log(\min_{i \in I_1^t \cap I_2^t} p[i t])}{\log(\min_{i \in I_1^t} p[i t]) + \log(\min_{i \in I_2^t} p[i t])}$

Evaluation

1. Select set of users
 - A. most active
 - B. most connected
 - C. random











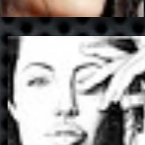

Evaluation

1. Select set of users
 - A. most active
 - B. most connected
 - C. random
2. Sort pairs by similarity

σ	u_1	u_2	?	TP	FP
0.7			y	1/3	0
0.6			n	1/3	1/3
0.4			y	2/3	1/3
0.3			y	3/3	1/3
0.1			n	3/3	2/3
0.0			n	3/3	3/3

Evaluation

1. Select set of users
 - A. most active
 - B. most connected
 - C. random
2. Sort pairs by similarity
3. Construct ROC plot, compare AUC

σ	u_1	u_2	?	TP	FP
0.7			y	1/3	0
0.6			n	1/3	1/3
0.4			y	2/3	1/3
0.3			y	3/3	1/3
0.1			n	3/3	2/3
0.0			n	3/3	3/3



User pair sampling procedure

- Because of sparsity of neighbor and similarity matrices, we biased the selection of user pairs in favor of neighbors — a **conservative** choice!

```
repeat:  
  pick next u by sorting criterion  
  R ← set of 60 neighbors of u  
  for each n from R:  
    if n is active:  
      P ← (u,n)  
      stop when |P| = M
```

Results: ROC (M=1000 pairs)

- ✦ MIP consistently among top 3 measures
- ✦ MIP better than Last.fm's neighbor recommendations
- ✦ Best results:
 - ✦ most active users
 - ✦ aggregation across items (user = tag vector)

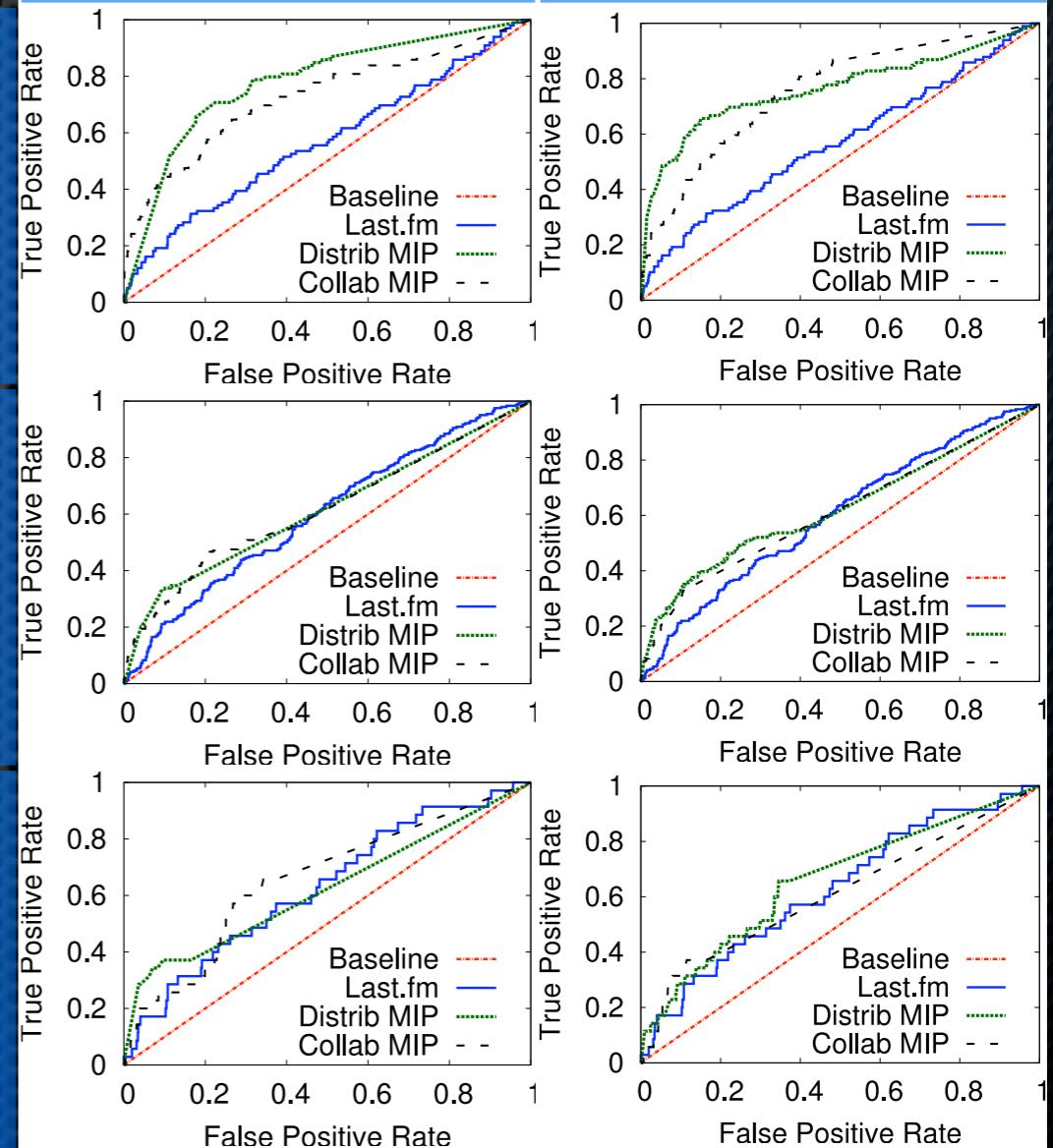
most active

most connected

random

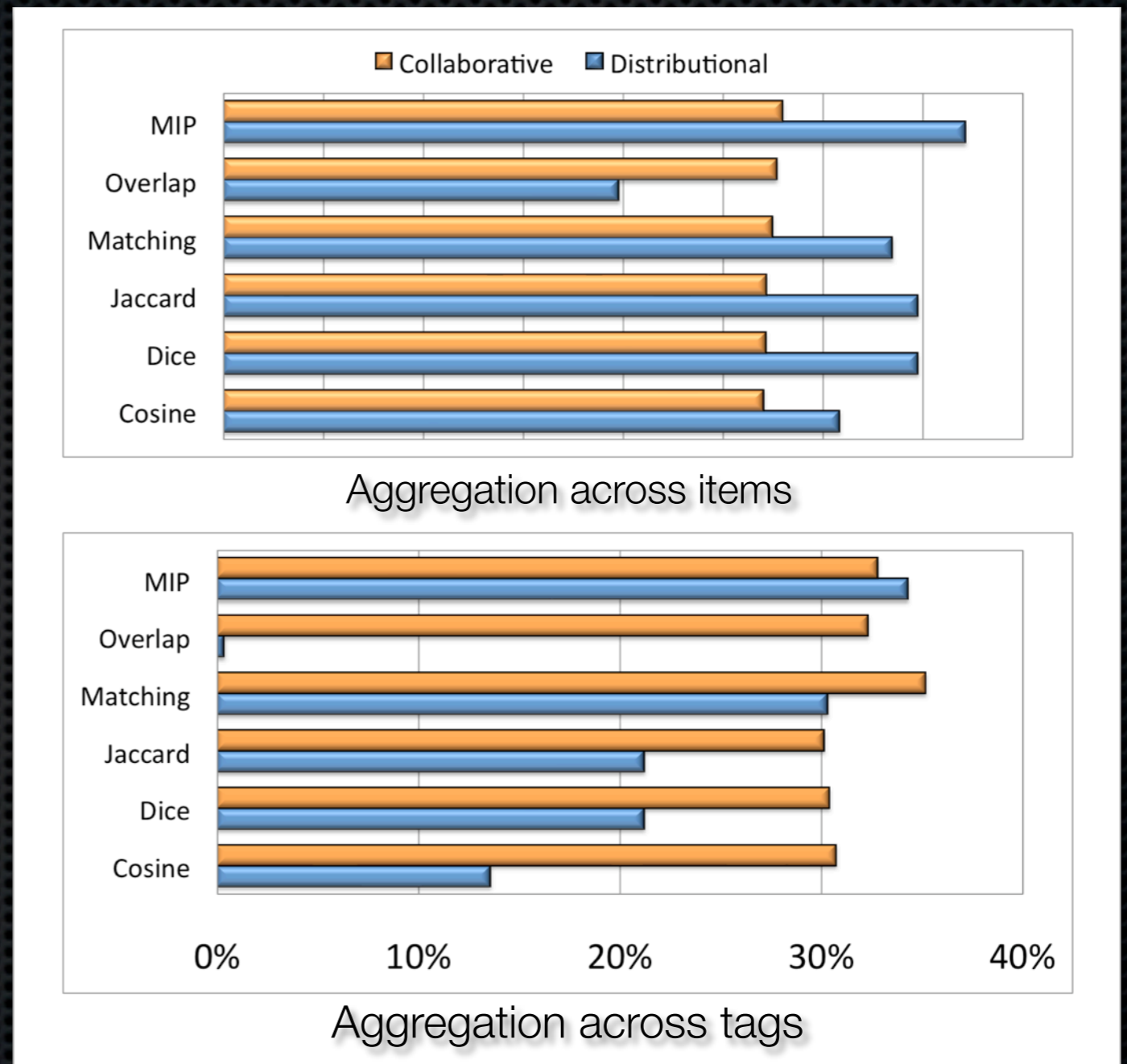
aggregation across items

aggregation across tags



Results: comparing measures (active users)

- All semantic similarity measures based on annotations outperform Last.fm's neighbor recommendations
- Distr. MIP aggregated across items is best overall
- Collaborative aggregation more helpful across tags



$$\text{AUC}(\sigma) / \text{AUC}(\text{Last.fm}) - 1$$

Summary

- ✦ Homophily: local alignment of tag usage for social neighbors
 - ✦ Null model allows to separate homophily from spurious correlations due to assortative mixing in social network, groups, and tagging activities
- ✦ User similarity based on annotations by active users is good predictor of social links (better than based on listening patterns)
 - ✦ Could be used to improve friend recommendations.
Eg, *tell Angelina to befriend Fil !...*

Related work

- Social link prediction based on node similarity (Liben-Nowell and Kleinberg, CIKM'03)
- Flickr friends seem to have higher vocabulary overlap: correlation or causality? (Marlow *et al.*, HT'06)
- Structure and evolution of online social networks (Kumar *et al.*, KDD'06; Mislove *et al.*, IMC'07, WOSN'08)
- Role of social contacts in shaping browsing patterns on Flickr (Lerman & Jones, ICWSM'07; van Zwol, WI'07)
- Do tag-based or resource-based interest sharing in CiteULike and Connotea relate to participation in the same discussion group? (Santos-Neto *et al.*, HT'09)

Future

- ✦ Longitudinal analysis to assess causality: do friends or shared interests come first?
- ✦ Evaluation
 - ✦ Confirm (strengthen) results with neighbor-independent user pair sampling procedure via Last.fm tasteometer
 - ✦ New data sets from aNobii (books), Facebook (apps)

Future

- ✦ Longitudinal analysis to assess causality: do friends or shared interests come first?
- ✦ Evaluation
 - ✦ Confirm (strengthen) results with neighbor-independent user pair sampling procedure via Last.fm tasteometer
 - ✦ New data sets from aNobii (books), Facebook (apps)
- ✦ Applications
 - ✦ “Suggest friend” on GiveALink.org
 - ✦ Games based on tag, resource, and user similarity to incentivize annotations and make social recommendations
 - ✦ Link recommendation in mobile social networking



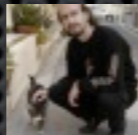
Thank you!

Q's?

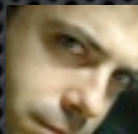
Rossano Schifanella



Alain Barrat



Ciro Cattuto



Benjamin Markines



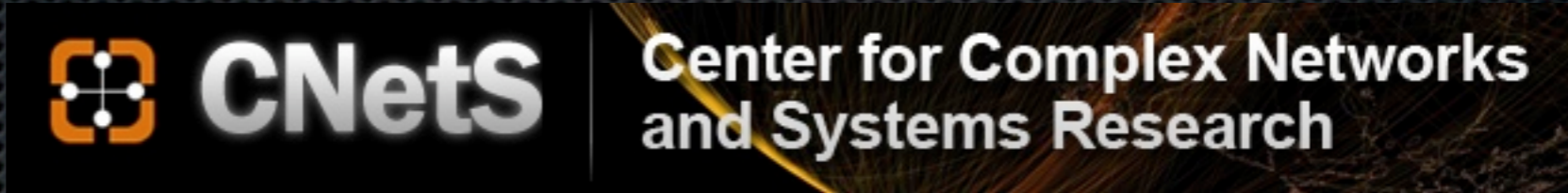
Filippo Menczer



Thanks also to Andrea Baldassarri, Andrea Capocci, Vittorio Loreto, Vito Servedio

Thank you!

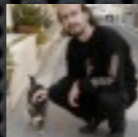
Q's?



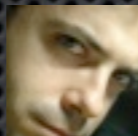
Rossano Schifanella



Alain Barrat



Ciro Cattuto



Benjamin Markines



Filippo Menczer



Thanks also to Andrea Baldassarri, Andrea Capocci, Vittorio Loreto, Vito Servedio

