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
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

<table>
<thead>
<tr>
<th></th>
<th>$n_t$</th>
<th>$n_g$</th>
<th>$a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>0.349</td>
<td>0.482</td>
<td>0.268</td>
</tr>
<tr>
<td>$n_t$</td>
<td>0.429</td>
<td>0.753</td>
<td></td>
</tr>
<tr>
<td>$n_g$</td>
<td></td>
<td>0.304</td>
<td></td>
</tr>
</tbody>
</table>
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)
Similarity vs. social distance
Similarity vs. social distance

distance computed via BFS
Similarity vs. social distance

distance computed via BFS

similarity computed via matching or cosine
Similarity vs. social distance

distance computed via BFS

similarity computed via matching or cosine

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

lexical: shared tags

\[
\sigma_{\text{groups}}(u, v) = \frac{\sum_g \delta_g^u \delta_g^v}{\sqrt{n_g(u)n_g(v)}}
\]

topical: shared groups
Correlation ≠ Causation

- We expect a purely statistical (spurious) correlation just because of assortative biases
Correlation ≠ 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 $k$, $n_t$, $n_g$, and $a$
  - No local topical or lexical alignment other than from statistical mixing patterns.


\[
\begin{align*}
  n_t &= 2 \\
  a &= 3 \\
  n_t &= 2 \\
  a &= 2
\end{align*}
\]
Correlation ≠ 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 $k, n_t, n_g$, and $a$
  - No local topical or lexical alignment other than from statistical mixing patterns

$n_t = 2$
$a = 3$
$n_t = 2$
$a = 2$
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
Maximum Information Path
\[ \sigma(x_1, x_2) = \frac{2 \times \log(\min_{y \in X_1 \cap X_2} [p(y)])}{\log(\min_{y \in X_1} [p(y)]) + \log(\min_{y \in X_2} [p(y)])} \]
Semantic similarity measures

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

Jaccard:  \[ \sigma(x_1, x_2) = \frac{\sum_{y \in X_1 \cap X_2} \log p(y)}{\frac{1}{2} \sum_{y \in X_1 \cup X_2} \log p(y)} \]

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

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

Cosine:  \[ \sigma(x_1, x_2) = \frac{X_1 \cdot X_2}{||X_1|| \cdot ||X_2||} \]
Applying semantic similarity measures to users

- Example: Max Info Path (MIP)

<table>
<thead>
<tr>
<th>Aggregation</th>
<th>Distributional</th>
<th>Collaborative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Across items</td>
<td>$\frac{2 \log(\min_{t \in T_1 \cap T_2} p[t])}{\log(\min_{t \in T_1} p[t]) + \log(\min_{t \in T_2} p[t])}$</td>
<td>$\sum_i \frac{2 \log(\min_{t \in T_1 \cap T_2} p[t</td>
</tr>
<tr>
<td>Across tags</td>
<td>$\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])}$</td>
<td>$\sum_t \frac{2 \log(\min_{i \in I_1 \cap I_2} p[i</td>
</tr>
</tbody>
</table>
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
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

```
<table>
<thead>
<tr>
<th>σ</th>
<th>u1</th>
<th>u2</th>
<th>?</th>
<th>TP</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td><img src="image1" alt="u1" /></td>
<td><img src="image2" alt="u2" /></td>
<td>y</td>
<td>1/3</td>
<td>0</td>
</tr>
<tr>
<td>0.6</td>
<td><img src="image1" alt="u1" /></td>
<td><img src="image2" alt="u2" /></td>
<td>n</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>0.4</td>
<td><img src="image1" alt="u1" /></td>
<td><img src="image2" alt="u2" /></td>
<td>y</td>
<td>2/3</td>
<td>1/3</td>
</tr>
<tr>
<td>0.3</td>
<td><img src="image1" alt="u1" /></td>
<td><img src="image2" alt="u2" /></td>
<td>y</td>
<td>3/3</td>
<td>1/3</td>
</tr>
<tr>
<td>0.1</td>
<td><img src="image1" alt="u1" /></td>
<td><img src="image2" alt="u2" /></td>
<td>n</td>
<td>3/3</td>
<td>2/3</td>
</tr>
<tr>
<td>0.0</td>
<td><img src="image1" alt="u1" /></td>
<td><img src="image2" alt="u2" /></td>
<td>n</td>
<td>3/3</td>
<td>3/3</td>
</tr>
</tbody>
</table>
```
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!

```plaintext
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)
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 a 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