Examples of online social network analysis
Social networks

• Huge field of research

• Data: mostly small samples, surveys

• Multiplexity

• Longitudinal data

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/
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
Map of Online Communities

And Related Points of Interest

Geographic area represents estimated size of membership

Spring 2007

(Note: not a complete survey. Sizes based on best figures I could find but involved some guesswork. Do not use for navigation.)
Map of Online Communities

Size on map represents volume of social activity (posts, chatters, etc.) based on data gathered over the spring and summer of 2010.

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 communities - that is, how much talking, playing, sharing, or other socializing happens there. This means some communities of articles and images, but it did my best and tried to be consistent.

Experiences are based on the best numbers I could find. But included a great deal of guesstimating. Experience, honking, slamming, incapable to guess, a 20,000-cell spreadsheet, dramas, carrots, sex, cat, reading, golf, sacrifices, and gut instinct (i.e. making things up).
An example: Dunbar number on twitter

Fraction of reciprocated connections as a function of in-degree

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

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
Mixing patterns

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

average activity of nearest neighbors as a function of own activity
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 global 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

Real data vs null model

Average number of common books

Average normalized similarity measure

\[ \langle n_{cb} \rangle \]

\[ \langle \sigma_b \rangle \]

Distance between users on social network

\[ d \]

\[ \Rightarrow \text{Genuine HOMOPHILY effect, not only due to assortativity w.r.t. amount of activity} \]
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?

Origin of homophily?

selection vs influence

Need to observe temporal evolution
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

Preferential attachment dynamics of new nodes

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

Distance between u and v on social network before creation of link (u,v)
Dynamics: selection or influence?

<table>
<thead>
<tr>
<th></th>
<th>$\langle n_{cb} \rangle$</th>
<th>$\sigma_b$</th>
<th>$\langle n_{cg} \rangle$</th>
<th>$\sigma_g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All $u,v$ such that $d_{uv}=2$</td>
<td>9.5 (0.2)</td>
<td>0.02</td>
<td>1.12 (0.61)</td>
<td>0.05</td>
</tr>
<tr>
<td>Simple closure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>($u \rightarrow v$ with $d_{uv}=2$)</td>
<td>18.2 (0.09)</td>
<td>0.04</td>
<td>1.81 (0.45)</td>
<td>0.1</td>
</tr>
<tr>
<td>Double closure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>($u \leftrightarrow v$ with $d_{uv}=2$)</td>
<td>23.4 (0.03)</td>
<td>0.05</td>
<td>2.2 (0.36)</td>
<td>0.12</td>
</tr>
</tbody>
</table>

New links between already present users

![Diagram of new links between users](attachment:diagram.png)

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

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

$P(0) \sim 1e^{-4}$
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

<table>
<thead>
<tr>
<th>Demographic Feature</th>
<th>feed</th>
<th>no feed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>51.6%</td>
<td>51.4%</td>
</tr>
<tr>
<td>Male</td>
<td>46.7%</td>
<td>47.0%</td>
</tr>
<tr>
<td>Unspecified</td>
<td>1.5%</td>
<td>1.5%</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 OR YOUNGER</td>
<td>12.8%</td>
<td>13.1%</td>
</tr>
<tr>
<td>18-25</td>
<td>36.4%</td>
<td>36.1%</td>
</tr>
<tr>
<td>26-35</td>
<td>27.2%</td>
<td>26.9%</td>
</tr>
<tr>
<td>36-45</td>
<td>13.0%</td>
<td>12.9%</td>
</tr>
<tr>
<td>46 OR OLDER</td>
<td>10.6%</td>
<td>10.9%</td>
</tr>
<tr>
<td>Country (top 10 &amp; other)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>28.9%</td>
<td>29.1%</td>
</tr>
<tr>
<td>Turkey</td>
<td>6.1%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Great Britain</td>
<td>5.1%</td>
<td>5.2%</td>
</tr>
<tr>
<td>Italy</td>
<td>4.2%</td>
<td>4.1%</td>
</tr>
<tr>
<td>France</td>
<td>3.8%</td>
<td>3.9%</td>
</tr>
<tr>
<td>Canada</td>
<td>3.7%</td>
<td>3.8%</td>
</tr>
<tr>
<td>Indonesia</td>
<td>3.7%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Philippines</td>
<td>2.1%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Germany</td>
<td>2.3%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Mexico</td>
<td>2.0%</td>
<td>2.1%</td>
</tr>
<tr>
<td>226 Others</td>
<td>37.5%</td>
<td>37.7%</td>
</tr>
</tbody>
</table>

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

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
A small-world network
Clustering spectrum

![Graph showing the average clustering coefficient against degree. The graph includes a solid line representing the mean and a dashed line representing the 5/95th percentile.]
Degree correlations
Activity-degree correlations

(logins during 28 days)
Age homophily

- Age 20
- Age 30
- Age 40
- Age 50
- Age 60
- Random edge
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
Results

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

![Graph showing the probability of sharing vs. the number of sharing friends. The graph has two conditions: feed and no feed. The probability increases as the number of sharing friends increases for both conditions.](image)
Results

Stronger ties carry more influence
Results

weak ties are collectively more influential
it’s complicated
(but interesting!)