Social Search in “Small World” Experiments

Speakers: Jing-Kai Lou & Jun-Yi Zheng

Sharad Goel, Roby Muhamad, and Duncan Watts
Proceedings of the 18th international conference on World wide web (2009)
what is the small world experiment?

- Six degrees of separation [Milgram, 1967]
- Basic Procedure
  - Random people in Omaha, Nebraska were asked to send letters to a stranger in Boston.
  - Letters can only be passed to acquaintances. People were asked to choose one that they believed to have higher chance to succeed (i.e. find the stockbroker)
- Results
  - Only 64 out of 296 letters reached the goal (20%)
  - The average steps is 6 to reach the destination
questions

• In “small world experiments”, most chains fail to complete (about 80%)
  ➜ biasing estimates of true chain length!

• How can we estimate “true” chain length??
two interpretations of Milgram's “small world experiment”

• for a randomly chosen pair of individuals,
  – topological version
    • are connected by short paths
  – algorithmic version (stronger)
    • are connected by short paths
    • can navigate the short paths themselves

• examples:
  – To spread the sexually-transmitted disease
    ➔ Topological interpretation
  – To get a new job
    ➔ Algorithmic interpretation
contributions of this paper

• Providing support for the algorithmic hypothesis

• Findings for “true” chain
  – half of all chains can be completed in 6~7 steps
    (supports for “small world”)
  – mean estimated chain lengths are much longer
    (for some of the population, the world is NOT small)
previous work for topological interpretation

- [Leskovec and Horvitz, 2008] a study of a network of 180M IM users,
  - found that users were separated by a mean of 6.6 steps and a median of 7 steps.
  - a great evidence for topological interpretation!

- How is the previous work for algorithmic interpretation?
previous work for algorithmic interpretation

• A few anecdotal examples
  – loss of generality

• Mathematical and simulation models
  – homogeneity assumptions that treat all individuals as equivalent

• The previous works are not good enough!
summary of small world experiments

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average length of completed chain</td>
<td>7</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Completion rate</td>
<td>20 %</td>
<td>13 %</td>
<td>0.5 %</td>
</tr>
</tbody>
</table>

- bring us that
  - completed chains tend short
  - vast majority of chains never reach the targets

- Why?
potential reasons

• **Reason: social capital**
  – Some **resourceful individuals** can construct **short** paths to distant targets

• **Reason:**
  – Real world may consists of **many loosely connected small worlds**
attrition

• probability of failure passing
• [White 1970, Dodds 2003]
  – Estimator for the length distribution of chains

\[
\hat{p}_l = \frac{\tilde{p}_l}{\prod_{j=0}^{l-1} (1 - r_j)}
\]

The number of P_1 would be

\[
\frac{3}{(1-1/4)(1-1/2)} = 8
\]
experimental dataset

• Email Network
    • sources: 98,865 people (from 168 countries)
    • targets: 18 people (in 13 countries)
    • numbers of initial chains: 106,295

    • sources: 85,621 people (from 163 countries)
    • targets: 21 people (in 13 countries)
    • numbers of initial chains: 56,033
mostly participants of the experiments

• Common status:
  – from the United States and Western Europe,
  – white and Christian,
  – young,
  – college-educated,
  – middle-class
  – Professionals

• This would be a baseline group
first look at the experiments results

• Completion Rate
  – Exp. 1: 491 / 106,295 chains (0.5%)
  – Exp.2: 61 / 56,033 (0.1%)

• Reference: for 100,000 chains,

<table>
<thead>
<tr>
<th>Constant Attrition Rate</th>
<th>Survival Chains in Six Removes</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>1,7797</td>
</tr>
<tr>
<td>67%</td>
<td>129</td>
</tr>
</tbody>
</table>
To improve the model

• previous estimators do not account for individual-level heterogeneity
  – e.g. socioeconomic status, education, ages, …

• However, the majority of people who did not continue chains never came to the experiment website;

• as a substitute, authors decide to estimate the next-step continuance
"next-step continuance" probability

- given an individual A who forwards the message to B (not the ultimate target),
- estimating the probability that B continues the chain by.
Dataset for Search-ability

• The authors analyzed
  – 88,875 sender-recipient pairs, of which 32% of pairs comprised recipients who forwarded messages (continued links),
  – and 68% comprised recipients who did not (terminated links).
Searchability Model

• estimating the **next-step continuance probability** by **logistic multilevel regression**
  – a standard statistical tool for modeling data with group structure

• simplify statement,
  – Input: **social status**
  – Output: **next-step continuance probability**
\[ P(y_i = 1) = \logit^{-1}(\gamma + \beta_{\text{nonwhite}} X_{\text{nonwhite},i} + \beta_{\text{female}} X_{\text{female},i} + \alpha_{j1[i]} + \alpha_{j9[i]}) \]

- \( r \) is the intercept,
- the two \( \beta \) terms are fixed effects for female and nonwhite participants,
- the \( jk[i] \) correspond to the nine group effects.

Parameters (\( \alpha \)):

<table>
<thead>
<tr>
<th>age</th>
<th>education</th>
<th>work field</th>
</tr>
</thead>
<tbody>
<tr>
<td>work position</td>
<td>income</td>
<td>strength of relationship</td>
</tr>
<tr>
<td>reason for choosing recipient</td>
<td>relationship with recipient</td>
<td>relationship with target</td>
</tr>
</tbody>
</table>
# Effects of individual and relational attributes

<table>
<thead>
<tr>
<th>Age</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 17</td>
<td>0.01</td>
</tr>
<tr>
<td>18 – 29</td>
<td>0.03</td>
</tr>
<tr>
<td>30 – 39</td>
<td>0.02</td>
</tr>
<tr>
<td>40 – 49</td>
<td>-0.01</td>
</tr>
<tr>
<td>50 – 59</td>
<td>-0.02</td>
</tr>
<tr>
<td>Above 60</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relationship Strength</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely close</td>
<td>0.03</td>
</tr>
<tr>
<td>Very close</td>
<td>0.0</td>
</tr>
<tr>
<td>Fairly close</td>
<td>0.01</td>
</tr>
<tr>
<td>Casually</td>
<td>0.0</td>
</tr>
<tr>
<td>Not close</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education Level</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate school</td>
<td>0.04</td>
</tr>
<tr>
<td>Collage / University</td>
<td>0.0</td>
</tr>
<tr>
<td>High school</td>
<td>-0.03</td>
</tr>
<tr>
<td>Elementary school</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Work Field</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media/Advertising/Arts</td>
<td>0.02</td>
</tr>
<tr>
<td>Education/Science</td>
<td>0.01</td>
</tr>
<tr>
<td>IT/Telecommunication</td>
<td>0.0</td>
</tr>
<tr>
<td>Government</td>
<td>-0.01</td>
</tr>
<tr>
<td>Other</td>
<td>-0.02</td>
</tr>
</tbody>
</table>
short conclusion for this model

• It shows that **high status individuals**
  – get **higher next-step continuance probability**
  – are more likely to pass along messages to friends who again pass them along
the estimated attrition distribution
• Later, we introduce a rigorous formulation to estimate the “true chain length”
Proposed Estimation Procedure

selecting the particular path

Q (w)

the path could be observed or not?

• \( \Omega \): space of path \( w \)
• \( w \): a path in \( \Omega \)
• \( P(w) \): likelihood of selecting any particular path \( w \)
• \( Q(w) \): prob. of \( w \) be observed

\( \Rightarrow \) \( w \) is “\textit{observed to complete}” with probability \( P(w)Q(w) \)
expected chain length without attrition

- without attrition \(\Rightarrow\) all \(w\) are observed \(\Rightarrow\) \(Q(w) = 1\)
- the "true" expected chain length can be expressed as a weighted average over the space of all paths

\[
\mu = \sum_{\omega} f(\omega) P(\omega)
\]

Length of path \(\omega\)

- So, what if \(Q(w) < 1\)?
Theorem: An Unbiased Estimator

A general unbiased estimator of $\mu$, mean estimated chain length

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{m} \frac{f(X_{k_i})}{Q(X_{k_i})}$$

- $X_{k1}, \ldots, X_{km}$ are the $m$ observed, non-missing values.
- Total # of complete and incomplete chains in the sample
- # of observed complete chain
- probability that $X_{ki}$ is observed uncorrupted after it has been sampled

2009/10/13 Jing-Kai Lou & Jun-Yi Zheng
proof of theorem (1/2)

\[ \mu = \sum_{\omega} f(\omega) P(\omega) \quad \mu = \frac{1}{n} \sum_{i=1}^{n} \frac{\bar{f}(X_i)}{Q(X_i)} \]

\( \bar{f} \) (extensive function): union\{\Omega, NA\} \rightarrow union\{N, 0\}

where the sum is taken over all samples (including the missing values)

since, simples \( Xi \) are identically distributed

\[ \mathbb{E}[\hat{\mu}] = \mathbb{E}[\bar{f}(X_i)/Q(X_i)] \]
proof of theorem (2/2)

• Also, we know that
  \( w \) is \textit{observed to complete} (i.e. non-missing) with probability \( P(w)Q(w) \)
  \[ \tilde{f}(NA) = 0 \]

• Therefore,

\[
\mathbb{E}\left[ \frac{\tilde{f}(X_i)}{Q(X_i)} \right] = \sum_{\omega \in \Omega} \frac{f(\omega)}{Q(\omega)} P(\omega)Q(\omega) = \sum_{\omega \in \Omega} f(\omega)P(\omega) = \mu
\]
estimators for mean chain length

• estimated mean chain length (from theorem)

\[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{m} \frac{L(X_{k_i})}{Q(X_{k_i})} \]  ...(3)
estimators for mean chain length

- estimated chain length distribution
  - Set $f_i(w) = 1$ if $\text{len}(w) = i$, and $f_i(w) = 0$, o.w.
  - Then, $p_i = \sum_{\omega \in \Omega} f_i(\omega)P(\omega) = \sum_{\omega \text{ is of length } i} P(\omega)$
  - Applying to THM,

\[
\hat{p}_i = \frac{1}{n} \sum_{j=1}^{m} \frac{f_i(X_{k,j})}{Q(X_{k,j})} = \frac{1}{n} \left[ \sum_{X_k \text{ has length } i} \frac{1}{Q(X_k)} \right] ...	ext{(4)}
\]
we didn’t exactly introduce $Q(w)$ yet!
estimating chain lengths based on exp.

• applying the estimators (3) and (4) yield an estimated “true” mean and median

• bootstrap samples, $S_1, \ldots, S_k$, $k=10,000$
  – Bootstrap is sampling with replacement from a sample (complete and incomplete chains)
  – Bootstrap sample give more detail on the distribution of this mean, or probability of this mean
Homogeneous Attrition

- attrition $r$ is fixed

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Travers &amp; Milgram Exp.</th>
<th>Dodds Exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>attrition $r$</td>
<td>0.25</td>
<td>0.41 for initial 0.71 o.w</td>
</tr>
<tr>
<td>$Q(w)$</td>
<td>$(1-r)^{L(w)}$</td>
<td>$(1-r_0)(1-r)^{L(w)-1}$</td>
</tr>
<tr>
<td>mean of chain length</td>
<td>11.8</td>
<td>41.5</td>
</tr>
<tr>
<td>median</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

2009/10/13
Jing-Kai Lou & Jun-Yi Zheng
Heterogeneous Attrition

- Initial attrition is fixed, $r_{w_0 \rightarrow w_1} = 0.41$
- attrition of $i^{th}$ based on $(i-1)^{st}$’s attributes
- $Q(w) = (1-r_{w_0 \rightarrow w_1}) (1-r_{w_1 \rightarrow w_2}) \ldots (1-r_{w_{L(w)-1} \rightarrow w_{L(w)}})$
- mean chain length = 22
- median = 7

Fig 2: The estimated CDF of chain length under the heterogeneous attrition model.
Randomized Attrition

- attrition $R_i$ are randomly generated from the distribution in Figure 1
- $Q(w) = (1-R_{w_0})(1-R_{w_1}) \ldots (1-R_{w_{l(w)-1}})$
- mean chain length = 49
- median = 6

Fig 1: The estimated distribution of attrition over individuals. Average attrition 0.7
## Summary of “true” Average Distance

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean (95% CI)</th>
<th>Median (95% CI)</th>
<th>Attrition r</th>
<th>$Q(\omega)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homogeneous attrition</td>
<td>11.8 (8.5-15)</td>
<td>7 (6-7)</td>
<td>$r = 0.25$</td>
<td>$(1-r)^{L(\omega)}$</td>
</tr>
<tr>
<td>(Travers/Milgram Exp.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogeneous attrition</td>
<td>41.5 (20-68)</td>
<td>6 (6-6)</td>
<td>$r_0 = 0.41$</td>
<td>$(1-r_0)(1-r)^{L(\omega)-1}$</td>
</tr>
<tr>
<td>(Dodds et. al. Exp.)</td>
<td></td>
<td></td>
<td>$r_i = 0.70$ (i &gt; 0)</td>
<td></td>
</tr>
<tr>
<td>Heterogeneous attrition</td>
<td>22 (4.5-57.5)</td>
<td>7 (6-8.5)</td>
<td>$r_0 = 0.41$</td>
<td>$(1-r_w_0\rightarrow w_1) (1-$</td>
</tr>
<tr>
<td>(Dodds et. al. Exp.)</td>
<td></td>
<td></td>
<td></td>
<td>$r_w_1\rightarrow w_2) \ldots (1-$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$r_{w_{i-1}\rightarrow w_{i}}) )$</td>
</tr>
<tr>
<td>Randomized attrition</td>
<td>49 (37-63)</td>
<td>6 (6-6)</td>
<td>$R_i$</td>
<td>$(1-R_{w_0}) (1-R_{w_1})$</td>
</tr>
<tr>
<td>(Dodds et. al. Exp.)</td>
<td></td>
<td></td>
<td></td>
<td>$\ldots (1-R_{w_{i-1}})$</td>
</tr>
</tbody>
</table>
Conclusion of the Experiment

• Medians of “true” chains are 6 or 7 steps, but the Means of “true” chains are range from 11.8 to 49 steps

→ Possibly many, chains are much, much longer than the median
Open Problems

• Could you please design an experiment that we need not to model “next-step continuance" probability as a substitute.
Thanks for your listening!