Yes, There is a Correlation — From Social Networks to Personal Behavior on the Web

Defended by Mykell Miller
Who else thinks this is important?

- Microsoft wanted to fund it.
- Google says 11 papers cited it, even though it’s only 1 year old.
Has it been done before?

- A lot of papers have been written on similar topics:
  - “Planetary-Scale Views on a Large Instant-Messaging Network” presented at the same conference
- However, this is the only one I could find that studied peoples’ similar interests instead of just similar demographics
Is it useful?

- Advertising: ads you’re more likely to click on
- Search engines: more relevant results
- Instant messaging: finding new chat partners
- Other online social networks
- More research: topological, temporal data
- More topics they didn’t think of
  - Online dating
  - Professional networking
Theory

$P_{\Phi}(U|R) =$ probability that users have trait $U$ in common, given that they have relationship $R$ in common.

$P_{\Phi}(R|U) =$ probability that users have relationship $R$, given that they have trait $U$ in common.

Relationship $R$ is usually the “we talk” relationship.
Theory

\[ P_\Phi (R | S) = P_\Phi (S | R) \ast P_\Phi (R) / P_\Phi (S) \]

- Bayes’ Theorem, which you hopefully already know
- Example \( S \): all 20 year old male users in zip 60201

\[ P_\Phi (S | R) = P_\Phi (S_{A_1} | R) \ast P_\Phi (S_{A_2} | R) \ast \ldots \ast P_\Phi (S_{A_m} | R) \]

- Assumes all attributes are independent
  - A lot of attributes are close to independent
  - Impossible to tell how all attributes depend on each other
  - It’s OK to be inaccurate because “[their] itention is not to calculate \( P_\Phi (R | S) \)”
- Example: \( S_{A_1} \) is all 20 year olds, \( S_{A_2} \) is all males, \( S_{A_3} \) is all residents of 60201.
The Dataset – where it came from

• They have all the MSN Messenger logs and Windows Live search data
  ◦ This is a lot of data they are very lucky to have.

• They are missing data on AIM, Google, etc. However…
  ◦ MSN and Windows Live are big and diverse
  ◦ They aren’t gods. They can’t get amazing data from everything.
The Dataset – Kinds of Data

- Demographics
  - Age, gender, zip

- Search queries
  - Not everybody uses Windows Live search
    - Had to throw out sessions where only 1 or 0 partners had search data
      - Otherwise, the data was useless because the paper is about common interests as revealed through search data.
  - Put into categories and sub-categories
Computing the Similarities

- Each attribute is indivisible – you’re either the same or you’re not.
- For searches, there two broader attributes for similar but not identical search queries.
- Throughout this paper, they compare how similar users are given different kinds of relationships.
Establishing the Correlation

- To do this, we compare the “baseline” and “messenger” probabilities for various attributes.
- Baseline: probability that any possible pair of users share an attribute.
- Messenger: probability that any pair of users that talk to each other share an attribute. In other words, probability that any pair of users with an edge between them share an attribute.
Basic results

Table 4: Similarities (%) comparing random pairs and messenger pairs

<table>
<thead>
<tr>
<th></th>
<th>Word</th>
<th>Query</th>
<th>Main Category</th>
<th>Sub Category</th>
<th>Zip</th>
<th>Age Group</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.51</td>
<td>0.09</td>
<td>15.26</td>
<td>6.23</td>
<td>0.81</td>
<td>34.40</td>
<td>51.67</td>
</tr>
<tr>
<td>Messenger</td>
<td>1.00</td>
<td>0.62</td>
<td>16.68</td>
<td>7.59</td>
<td>13.00</td>
<td>64.19</td>
<td>48.74</td>
</tr>
</tbody>
</table>

Figure 1: Similarities(%) comparing random pairs and messenger pairs: query attributes(left) and personal attributes(right)
Basic results – Search Queries

• If you talk, you’re almost 7x as likely to search for the same query – much smaller difference for main category and sub category.
 ◦ Me and someone who researches computational complexity both like to search for computer science papers, but we probably don’t talk
  ◦ Me and the authors of this paper both like to search for papers on online social networks, but we also probably don’t talk
  ◦ Me and you both like to search for papers by Alex Kuzmanovic, and we probably do talk.
Basic results – Personal Attributes

- If you talk, you’re about 16x as likely to live in the same zip code
  - There are a lot of zip codes, but a lot of zip codes have tight-knit communities in them, like a high school.

- If you talk, you’re about twice as likely to be in the same age group
  - Some contacts, like friends, are likely to be in the same age group
  - Others, like your boss, are likely to be in a different age group

- If you talk, you’re slightly less likely to be the same gender
  - One use of MSN is for flirting, and most people flirt with people of different genders.
Varying the Talk-time: Total Time

- Some people have close relationships, some people rarely talk. Divide relationships into 5 bins based on how often they talk.
- Is having bins with the same number of pairs the best solution?
- 5 bins is a lot.
- Any contact at all -> huge jump in zip, age.
- Bin with most contact -> huge jump in query.
Varying the Talk-time: Total Time

Figure 2: Variation in similarities(%) with total talk duration: query attributes(left) and personal attributes(right)
Varying the Talk-time: Time Per Message

- Longer time to send one message implies a more distant relationship
  - Not that interested in the conversation
  - Careful because you want to leave a good impression on a boss, hot piece of ass

- Age similarity does not seem to have any trend with increasing time spent per message – surprising
Varying the Talk-time: Time Per Message

Figure 4: Variation in similarities(%) with average time spent per message: query attributes(left) and personal attributes(right)
Conditioning on Personal Attributes

- Given you already have the same age, zip, and/or gender, how big of an effect does talking have on your similar interests?
  - A significant amount

- Why study based on same gender when people are more likely to talk if they are not the same gender?
  - People of the same gender are more likely to have similar interests.
Conditioning on Personal Attributes

Figure 6: Similarities(%) of query attributes conditioning on all personal attributes being same
Effect of Indirect Links

- Are friends of friends likely to be similar?
  - Yes!
- Triangles -> pair of users belong to both the 1-hop network and the 2-hop network.
- They do not correct for triangles. Is this a big flaw?
  - No! The 2-hop network is about 10x as big as the 1-hop network, so only about 10% of the similarities found are due to the similarities found in the 1-hop network.
- They only sample about 10% of the 2-hop network. Is this a big flaw?
  - No!
  - The sampling is uniform – likely to be a representative sample
  - Did several iterations and they were invariant.
Effect of Indirect Links

Figure 10: Similarities(%) in a 2-hop network: query attributes(left) and personal attributes(right)