How much can social metrics actually help in content distribution?

Ioannis Stavrakakis
National & Kapodistrian University of Athens

Based on works with: Merkouris Karaliopoulos, Eva Jaho, Panagiotis Pandazopoulos, et.al.

May 18, 2011

Focus on two key social metrics

Interest similarity

centrality
Interest Similarity

- Groups in online social networks are currently formed based on acquaintance, family relationships, social status, educational/professional background

...yet interests/preferences of group members are not always similar

There is value in assessing and using interest similarity in groups

Define and measure Interest Similarity

- assess similarity in the interests of existing social groups
- identify further interest-based structure within those groups

**ISCoDE framework**

1st step: user profiles \(\rightarrow\) weighted graph

2nd step: weighted graph \(\rightarrow\) interest-similar groups

---

**Similarity metrics: PS vs. InvKL (Kullback Leibler)**

- **Proportional Similarity (PS)**
  - $PS(F_i, F_j) \in [0, 1]$

- **Inverse symmetrized KL divergence**
  - $InvKL(F_i, F_j) = \frac{1}{\sum_{m} E_m \log \frac{E_m}{F_m} + \sum_{m} E_m \log \frac{F_m}{E_m}}$

$E_{mn}$, $1 \leq n \leq N$, $1 \leq m \leq M$ : distribution of node $n$ over interest class $m$

Example with $M=2$ interest classes and $N=2$ nodes

**Resolution performance**

(a) Similar nodes

<table>
<thead>
<tr>
<th>$Q$</th>
<th>$C$</th>
<th>$PS$ partition</th>
<th>$Q$</th>
<th>$C$</th>
<th>$InvKL$ partition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0215</td>
<td>2</td>
<td>{1.18} {39.80}</td>
<td>0.6740</td>
<td>5</td>
<td>{1.14} {15.28} {29.44} {45.61} {62.80}</td>
</tr>
</tbody>
</table>

(b) Dissimilar nodes

<table>
<thead>
<tr>
<th>$Q$</th>
<th>$C$</th>
<th>$PS$ partition</th>
<th>$Q$</th>
<th>$C$</th>
<th>$InvKL$ partition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7860</td>
<td>10</td>
<td>{1.8} {73.80}</td>
<td>0</td>
<td>1</td>
<td>{1.80}</td>
</tr>
</tbody>
</table>

$InvKL$ can identify smaller communities than $PS$, in a highly similar network

$PS$ can identify smaller communities than $InvKL$, in a highly dissimilar network

(could argue that this is not very useful)
Can Interest similarity improve network protocols?

- Gain of cooperation for content replication in a group of nodes

- Gain under cooperation

  - $T$ : tightness metric ($= \text{mean invKL}$), measuring interest similarity across group members

  - Percentage of cooperative nodes

  - E. Jaho, M. Karaliopoulos, I. Stavrakakis, “Social similarity as a driver for selfish, cooperative and altruistic behavior”, in Proc. AOC 2010 (extended version submitted to IEEE TPDS)

Can Interest similarity improve network protocols?

- Content dissemination in opportunistic networks

  - Protocols A,B,C are push protocols exercising interest-based forwarding

**Betweenness Centrality (BC)**

**Content (service) Migration / Placement**

Can BC help provide for a low-complexity, distributed, scalable solution?

- Destination-aware vs destination unaware BC
- Ego-centric vs socio-centric computation of BC

---

**CBC: the “destination-aware” counterpart to BC**

**Betweenness Centrality** ($u$): portion of all pairs shortest paths of $G$ that pass through node $u$

$$BC(u) = \sum_{s=1}^{|V|} \sum_{t=1}^{s-1} \frac{\sigma_{st}(u)}{\sigma_{st}}$$

**Conditional Betweenness Centrality** ($u$, $t$): portion of all shortest paths of $G$ from node $u$ to target $t$, that pass through node $u$

$$CBC(u; t) = \sum_{s \in V, u \neq t} \frac{\sigma_{ut}(u)}{\sigma_{ut}}$$
The content placement problem

Deploy scalable and distributed mechanisms for publishing, placing, moving UG Service facilities / content within networking structures

Optimal content / service placement in a Graph $\leftrightarrow$ k-median

Only distributed, scalable, solutions are relevant

- Use local information to migrate towards a better location
- Use locally available limited information to solve repeatedly small-scale k-median and repeat

(*)


Centrality-based service migration

Consider set of nodes with highest CBC values

- Solve *iteratively small-scale* k-mediants on subgraphs $G_i \in G$, around the current facility location of host $i$ containing the top nodes based on CBC values

- Map the outside demand properly on nodes in subgraphs $G_i$

P. Pandazopoulos, M. Karaliopoulos, I. Stavrakakis, “Centrality-driven scalable service migration”, 23rd International Teletraffic Congress (ITC), Sept. 6-9, 2011, San Francisco, USA.
**Ego-centric vs socio-centric computation of BC**

**Very high rank correlation (Spearman coefficient) !!!**

Ego- and socio-centric metrics identify same subsets

---

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AS/AS number</th>
<th>&lt;G&lt;</th>
<th>Diameter</th>
<th>Size</th>
<th>&lt;level&gt;</th>
<th>Spearman ρ</th>
<th>BC vs. ego-BL</th>
<th>Spearman ρ</th>
<th>BC vs. ego-CBL</th>
<th>Spearman ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td></td>
<td>56</td>
<td>106</td>
<td>2.23</td>
<td>0.9670</td>
<td>0.9906</td>
<td>0.9805</td>
<td>0.9707</td>
<td>0.9879</td>
<td>0.9868</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>33</td>
<td>11</td>
<td>3.53</td>
<td>0.9560</td>
<td>0.9705</td>
<td>0.9882</td>
<td>0.8480</td>
<td>0.9280</td>
<td>0.9030</td>
</tr>
<tr>
<td>13</td>
<td>TeoNet(3383)</td>
<td>338</td>
<td>12</td>
<td>3.73</td>
<td>0.9870</td>
<td>0.9705</td>
<td>0.9882</td>
<td>0.8480</td>
<td>0.9280</td>
<td>0.9030</td>
</tr>
<tr>
<td>14</td>
<td>Sprint(3392)</td>
<td>339</td>
<td>16</td>
<td>3.13</td>
<td>0.9900</td>
<td>0.9705</td>
<td>0.9882</td>
<td>0.8480</td>
<td>0.9280</td>
<td>0.9030</td>
</tr>
</tbody>
</table>

Betweeness Centrality (BC)

Centrality-driven routing in opportunistic nets

(SimBetTS and BubbleRap use BC values of encounters for content forwarding)

How is performance of centrality-based routing affected by?

- Adding or not, destination awareness to BC (BC vs CBC)
- Working with ego-centric vs socio-centric BC values
- Type of contact graph (unweighted vs. weighted) ? Not discussed here

P. Pantazopoulos, et.al. “How much off-center are centrality metrics for opportunistic routing?”, under submission

Datasets

5 well-known iMote-based real traces available from the Haggle Project at CRAWDAD.

<table>
<thead>
<tr>
<th>Characteristics of employed datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Configuration</strong></td>
</tr>
<tr>
<td>Device type</td>
</tr>
<tr>
<td>Network type</td>
</tr>
<tr>
<td>Duration (days)</td>
</tr>
<tr>
<td>Scan time (sec)</td>
</tr>
<tr>
<td>Granularity (sec)</td>
</tr>
<tr>
<td>Mobile Devices</td>
</tr>
<tr>
<td>Stationary Dev.</td>
</tr>
<tr>
<td>External Dev.</td>
</tr>
<tr>
<td>Average internal contacts/pair/day</td>
</tr>
<tr>
<td># of Contacts</td>
</tr>
</tbody>
</table>
BC vs CBC

- opt ➔ optimal routing through knowledge of contact sequences.
- BC/CBC ➔ up to 30% of messages never reach their destination ➔ about 5 times more hops and 1 day of additional delay

**BC outperforms CBC in delay** (due to zero CBC values when destination in an unconnected cluster)

**CBC outperforms BC in hops** (up to 50% shorter paths, due to selecting more proper nodes to forward to)

![BC vs CBC Diagram]

---

socio- vs ego-metrics

### Table 2: Probability of delivery

<table>
<thead>
<tr>
<th>DataSet</th>
<th>Probability of delivery (6h window)</th>
<th>(egoBC_u)</th>
<th>(socBC_u)</th>
<th>(egoBC_{uw})</th>
<th>(socBC_{uw})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambridge</td>
<td></td>
<td>82.68</td>
<td>84.50</td>
<td>82.14</td>
<td>77.24</td>
</tr>
<tr>
<td>Infocom’05</td>
<td></td>
<td>95.21</td>
<td>96.25</td>
<td>88.06</td>
<td>86.40</td>
</tr>
<tr>
<td>Content</td>
<td></td>
<td>65.55</td>
<td>71.84</td>
<td>69.18</td>
<td>69.20</td>
</tr>
<tr>
<td>Infocom’06</td>
<td></td>
<td>80.93</td>
<td>85.08</td>
<td>89.33</td>
<td>89.73</td>
</tr>
</tbody>
</table>
**socio- vs ego-metrics**

strong positive correlation of socio- and ego – metrics
(Intel / Content data)

![Graphs showing correlation values](image)

**Conclusions**

Focused on exploring the impact on two key social metrics on content distribution

- Interest Similarity
- Centrality

**Interest similarity metrics**

- Highly similar groups can yield high gains in content replication.
- Interest similarity –based forwarding improves performance
- Worth assessing interest similarity in groups – framework for doing that

**Destination-aware BC :**

- Very effective in content placement (BC is totally ineffective)
- Decreases hop count in opp nets (energy) substantially. Can increase delay

**Ego-centric centrality variants (BC/CBC)**

- Highly rank correlated → no performance degradation in content placement / centrality-driven content forwarding.
- Easier to compute