Social Computing in Blogosphere
Opportunities and Challenges

Nitin Agarwal*
Arizona State University

(Joint work with Huan Liu, Sudheendra Murthy, Arunabha Sen, Lei Tang, Xufei Wang, and Philip S. Yu)

* Nitin Agarwal will join University of Arkansas at Little Rock as Assistant Professor from Fall 2009
Social Media & Web 2.0

- Blogs
  - Blogger
  - Wordpress
  - Twitter
- Wikis
  - Wikipedia
  - Wikiversity
- Social Networking Sites
  - Facebook
  - Myspace
- Digital media sharing websites
  - Youtube
  - Flickr
- Social Tagging (folksonomies)
  - Del.icio.us
Top 20 Most Visited Websites

- Internet traffic report by Alexa on April 26th 2009

<table>
<thead>
<tr>
<th>Rank</th>
<th>Website</th>
<th>Rank</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yahoo!</td>
<td>11</td>
<td>Orkut</td>
</tr>
<tr>
<td>2</td>
<td>Google</td>
<td>12</td>
<td>RapidShare</td>
</tr>
<tr>
<td>3</td>
<td>YouTube</td>
<td>13</td>
<td>Baidu.com</td>
</tr>
<tr>
<td>4</td>
<td>Windows Live</td>
<td>14</td>
<td>Microsoft Corporation</td>
</tr>
<tr>
<td>5</td>
<td>MSN</td>
<td>15</td>
<td>Google India</td>
</tr>
<tr>
<td>6</td>
<td>Myspace</td>
<td>16</td>
<td>Google Germany</td>
</tr>
<tr>
<td>7</td>
<td>Wikipedia</td>
<td>17</td>
<td>QQ.Com</td>
</tr>
<tr>
<td>8</td>
<td>Facebook</td>
<td>18</td>
<td>EBay</td>
</tr>
<tr>
<td>9</td>
<td>Blogger</td>
<td>19</td>
<td>Hi5</td>
</tr>
<tr>
<td>10</td>
<td>Yahoo! Japan</td>
<td>20</td>
<td>Google France</td>
</tr>
</tbody>
</table>

- 40% of the top 20 websites are social media sites
Social Media Characteristics

- Power of the Long Tail
- Rich Internet Applications
- User generated contents
- User enriched contents
- User developed widgets (Mashups)
- Collaborative environment: Participatory Web, Citizen journalism
Challenges

- **Time Challenge:** Dynamic environment
  - Data gets stale too soon

- **Size Challenge:** Phenomenal growth
  - Difficult to follow

- **Sparse link structure**
  - Nature of the Long Tail

- **Information Quality**
  - Colloquial, often misspelled, slang text
  - Lots of off-topic chatter/noise

- **Evaluation Challenge**
  - Absence of ground truth

ICWSM’09, WSDM’08, SIGKDD’08, ICWE’08, ICCCD’08 NGDM’07
Identifying Influential Bloggers

WSDM’08
http://videolectures.net/wsdm08_agarwal_iib/
Blogosphere Growth

- Technorati is indexing 133 million blog records currently
- 2 blogs or 18.6 blog posts per second
Influential Sites and Bloggers

- Power law distribution
- Short Head blogs
  - Influential sites
  - Search engines
  - Information Diffusion [Gruhl et al. 2004; Kempe et al. 2003; Richardson and Domingos 2002; Java et al. 2006]
- Long Tail blogs [Anderson 2006]
  - Inordinately many
  - Less popular
  - Cater to niche interests
- Extremely challenging to study all these blogs
- Influential bloggers as representatives
Real and Virtual World

Real World

Virtual World

Domain Expert

Friends

Online Community
Influential Bloggers

• Inspired by the analogy between real-world and blog communities, we answer:
  Who are the influentials in Blogosphere?
  Can we find *them*?

  \[ \text{Active Bloggers} = \text{Influential Bloggers} \]

• Active bloggers may not be influential
• Influential bloggers may not be active
Searching for the Influentials

- Active bloggers
  - Easy to define
  - Often listed at a blog site
  - Are they necessarily influential?
- How to define an influential blogger
  - Influential bloggers have influential posts
  - Subjective
  - Collectable statistics
  - How to use these statistics
Intuitive Properties

• Social Gestures (statistics)
  – **Recognition**: Citations (incoming links)
    – An influential blog post is recognized by many. The more influential the referring posts are, the more influential the referred post becomes.
  – **Activity Generation**: Volume of discussion (comments)
    – Amount of discussion initiated by a blog post can be measured by the comments it receives. Large number of comments indicates that the blog post affects many such that they care to write comments, hence influential.
  – **Novelty**: Referring to (outgoing links)
    – Novel ideas exert more influence. Large number of outlinks suggests that the blog post refers to several other blog posts, hence less novel.
  – **Eloquence**: “goodness” of a blog post (length)
    – An influential is often eloquent. Given the informal nature of Blogosphere, there is no incentive for a blogger to write a lengthy piece that bores the readers. Hence, a long post often suggests some necessity of doing so.

• **Influence Score** = f(Social Gestures)
Proposed Model

\[ \text{InfluenceFlow}(p) = \sum_{m=1}^{\lvert I \rvert} I(p_m) - \sum_{n=1}^{\lvert J \rvert} I(p_n) \]

\[ I(p) \propto w_{\text{comm}} \gamma_p + \text{InfluenceFlow}(p) \]

\[ I(p) = w(\lambda) \times (w_{\text{comm}} \gamma_p + \text{InfluenceFlow}(p)) \]

\[ \text{iIndex}(B) = \max(I(p_i)) \]

**Link adjacency matrix:** \( A \)

\[ A_{ij} = 1; \ p_i \rightarrow p_j \]

\[ A_{ij} = 0; \text{otherwise} \]

\[ \vec{\lambda} = (w(\lambda_{p_1}), ..., w(\lambda_{p_N}))^T, \]

\[ \vec{\gamma} = (\gamma_{p_1}, ..., \gamma_{p_N})^T, \]

\[ \vec{I} = (I(p_1), ..., I(p_N))^T, \]

\[ \vec{f} = (f(p_1), ..., f(p_N))^T \]

\[ \vec{f} = w_{\text{in}} A^T \vec{I} - w_{\text{out}} A \vec{I} = (w_{\text{in}} A^T - w_{\text{out}} A) \vec{I} \]

\[ \vec{I} = \vec{\lambda}(w_c \vec{\gamma} + \vec{f}) \]

\[ \vec{I} = \vec{\lambda}(w_c \vec{\gamma} + (w_{\text{in}} A^T - w_{\text{out}} A) \vec{I}) \]
First Look: Analytics for iPhone
by Dave Caolo on Feb 5th 2009 at 8:00AM

Google Analytics is a popular and quite useful set of tools for monitoring a web site's traffic and performance. Set up is a snap and the reports are easy to read and flexible. You can create goals, monitor traffic and so on. What more could you want? On-the-go reports via your iPhone? All of your target statistics in your pocket? Oh, all right.

Earlier this week, Michael D Jensen of Inblosam LLC released Analytics App, which presents everything you’d ever want from Google Analytics on your iPhone. It is exhaustive.

When you first launch Analytics App, you’re asked for your Google login (you must have a pre-existing Analytics account). From there, a list of all the sites you’re monitoring appears. Click any one and view nearly 30 reports, including traffic, visitors, content ... even events tracking you’ve set up and your own customized reports. It’s speedy over Wi-Fi and EDGE.

For example, Analytics App’s traffic reports include referring sites, search engines, keywords, AdWords campaigns and more. Set the date range of any report to sort by day, week or month. The Dashboard provides an overview complete with easy-to-read graphs.

For $5.99US, this application is a keeper. Up-to-date stats from all of your sites, available nearly anywhere, makes our geeky little hearts go pitter-pat.
Active & Influential Bloggers

- Active and Influential Bloggers
- Inactive but Influential Bloggers
- Active but Non-influential Bloggers

- We don’t consider “Inactive and Non-influential Bloggers”, because they seldom submit blog posts. Moreover, they do not influence others.
Temporal Patterns

- Long term Influentials
- Average term Influentials
- Transient Influentials
- Bourgeoing Influentials
Verification of the Model

• Challenges
  – No training and testing data
  – Absence of ground truth
  – How to do it?

• We use another Web 2.0 website, Digg as a reference point.
• “Digg is all about user powered content. Everything is submitted and voted on by the Digg community. Share, discover, bookmark, and promote stuff that’s important to you!”
• The higher the digg score for a blog post is, the more it is liked.
• A not-liked blog post will not be submitted thus will not appear in Digg
Digg - *Power of Web 2.0*
Findings w.r.t. Digg

• Digg records top 100 blog posts obtained through Digg Web API.

• Top 5 influential and top 5 active bloggers were picked to construct 4 categories

• For each of the 4 categories of bloggers, we collect top 20 blog posts from our model and compare them with Digg top 100.

<table>
<thead>
<tr>
<th>Bloggers</th>
<th>Active</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influential</td>
<td>S1: 17</td>
<td>S2: 7</td>
</tr>
<tr>
<td>Non-influential</td>
<td>S3: 3</td>
<td>S4: 0/1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bloggers</th>
<th>Active</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influential</td>
<td>S1: 71</td>
<td>S2: 14</td>
</tr>
<tr>
<td>Non-influential</td>
<td>S3: 8</td>
<td>S4: 7</td>
</tr>
</tbody>
</table>

• Distribution of Digg top 100 and TUAW’s 535 blog posts
Relative Importance of Parameters

• Observe how much our model aligns with Digg.
• Compare top 20 blog posts from our model and Digg.
• Considered six months

Considered all configuration to study relative importance of each parameter.

**Recognition (Inlinks)** > **Activity Generation (Comments)** > **Novelty (Outlinks)** > **Eloquence (Blog post length)**
Identifying Familiar Strangers

ICWSM’09, NGDM’07
Who are Familiar Strangers?

- Observe repeatedly, but do not know each other
- Real World
  - E.g., Individuals observe each other daily on a train
  - Discover the latent pattern: going to same workplace,
- Blogosphere
  - What you write is who you are…
  - Have similar blogging behavior, interests (Movie, Games, Technology, Politics, etc.)
  - Not in each others social network
Aggregating Familiar Strangers

• Together they form a critical mass
  – understanding of one blogger gives a sensible and representative glimpse to others
  – better customization, personalization and recommendation
  – nuances among them present new business opportunities
  – predictive modeling and trend analysis
An Example

u: Given blogger
C_u: \{v_1, v_2, v_3, v_4\}

A_u: \{Exercise, History, Recreation\}
A_{v1}: \{Internet, News\}
A_{v2}: \{Blogging, Internet\}
A_{v3}: \{Blogging, Internet, Technology\}
A_{v4}: \{Recreation, Travel\}

Find T_u, given \(\gamma\): Sports=\{Exercise, Recreation\}
Searching for Familiar Strangers

- Given a node $u$, its attributes $A_u$
- Egocentric view of the network, $C_u = \{\text{adjacent nodes of } u\}$
- Familiar strangers, $T_u = \{v\}$
  - *Familiar*: $A_v \cap \gamma \neq \emptyset$, where $\gamma \subseteq A_u$
  - *Stranger*: $u$ and $v$ are non-adjacent
Social Identity Approach

- Social Identity: ability to cluster contacts into meaningful groups
- Search only relevant clusters of contacts
  - Prune the search space
- Desiderata
  - Small-world assumption
    - Power law degree distribution: \( f(x) \propto ax^{-\gamma} \)
    - High clustering coefficient:
    - Short average path length:
      \[
      l_G = \frac{1}{n(n-1)} \sum_{i,j} \sum_{i \neq j} d(v_i,v_j)
      \]
Social Identity Construction

• Offline clustering of contacts
• Contacts represented by
  – Tag vector
  – Content vector
• LSA transformation to concept vectors [Deerwester et al. 1990]

\[
X_{\text{tag}} = U_{\text{tag}} \Sigma_{\text{tag}} V_{\text{tag}}^T \quad X_{\text{con}} = U_{\text{con}} \Sigma_{\text{con}} V_{\text{con}}^T
\]

• \( S_{\text{tag}} \): Pairwise cosine similarity between row vectors of \( V_{\text{tag}} \)
• \( S_{\text{con}} \): Pairwise cosine similarity between row vectors of \( V_{\text{con}} \)
• \( S = \alpha S_{\text{tag}} + (1-\alpha) S_{\text{con}} \)
• \( k \)-means clustering
Alternative Approaches

- **Exhaustive Approach**
  - Search all the contacts
  - 100% accuracy
  - Exponential search cost: 
    \[
    \sum_{k=1}^{h} d^k
    \]

- **Random Approach**
  - Fraction of contacts (\(\sigma\)) propagate the search
  - \(\sigma = 1\) corresponds to Exhaustive approach
Evaluation

• Ground Truth - Global network view
  – Steiner tree based approach [Du and Hu 2008]
    • Lower bound on search space
• Compare with
  – Exhaustive approach
  – Random approach
• Datasets:
  – Blogcatalog (~24K bloggers)
  – DBLP (~35K authors)
Small-World Properties

• Blogcatalog
  – Power law degree dist.
  – Clustering Coefficient
    • 0.51 (actual)
    • 0.001±0.0002 (random)
  – Avg. path length
    • 2.37

• DBLP
  – Power law degree dist.
  – Clustering Coefficient
    • 0.69 (actual)
    • 0.001 ± 0.0002 (random)
  – Avg. path length
    • 5.08
## Results

<table>
<thead>
<tr>
<th>Approach (E)</th>
<th>Accuracy (%)</th>
<th>Search Space (edge traversals)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steiner Tree</td>
<td>100%</td>
<td>3,565 ± 23</td>
</tr>
<tr>
<td>Exhaustive</td>
<td>100%</td>
<td>4,531,967 ± 944</td>
</tr>
<tr>
<td>Random</td>
<td>1.0283% ± 0.928</td>
<td>1,823 ± 43</td>
</tr>
<tr>
<td>Social Identity</td>
<td>79.2908% ± 3.008</td>
<td>6,032 ± 46</td>
</tr>
</tbody>
</table>

**BlogCatalog:**

<table>
<thead>
<tr>
<th>Approach (E)</th>
<th>Accuracy (%)</th>
<th>Search Space (edge traversals)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steiner Tree</td>
<td>100%</td>
<td>4,752 ± 30</td>
</tr>
<tr>
<td>Exhaustive</td>
<td>100%</td>
<td>909,543 ± 403</td>
</tr>
<tr>
<td>Random</td>
<td>2.304% ± 0.355</td>
<td>58 ± 12</td>
</tr>
<tr>
<td>Social Identity</td>
<td>91.349% ± 2.107</td>
<td>12,182 ± 68</td>
</tr>
</tbody>
</table>

**DBLP:**
Looking Ahead

• Open source taxonomies to create social identity

• Multiple social dimensions

• Temporal dynamics of familiar strangers as social network evolves

• Affect of negative polarity on social ties

• “Strength of weak ties” and effects on communication topology