Quantification of Topic Propagation Using Percolation Theory:
A Study of the ICWSM Network

Ali Azimi Bolourian
Yashar Moshfeghi
C. J. “Keith” van Rijsbergen

University of Glasgow
Computing Science

ICWSM 2009, San Jose, California, USA
Research Question

In a social network, how many people should know about a topic before it is considered prevalent?
Research Question

Part of an ongoing research to:

1. **Quantify** propagation; define a threshold for “Global” topics.

2. **Predict**, in an early stage, which topics have the potential to become global.
Why?
Why?

• Setting a goal for online ad campaigns
Why?

• Setting a goal for online ad campaigns

• Search engine resource allocation
Why?

• Setting a goal for online ad campaigns

• Search engine resource allocation

• Better understanding of social network dynamics
Why?

• Setting a goal for online ad campaigns
• Search engine resource allocation
• Better understanding of social network dynamics

This would in turn improve all services that have currently failed to capitalize on social networking dynamics.
MySpace and Friends Need to Make Money. And Fast.

By Kevin Kelleher  03.24.08

The numbers are amazing. MySpace's membership has ballooned from 20 million people in 2005 to 225 million today, an average annual growth rate of 513 percent. Rival Facebook grew at 550 percent a year during the same period. LinkedIn's rate was 182 percent.

Yet one social networking metric is distinctly underwhelming: the one with a dollar sign. Lookery, an ad network specializing in social media, offers display ads on MySpace, Facebook, and Bebo for only 13 cents per thousand times the ad is served (CPM); Yahoo's CPM is three times higher. If it is true that advertising generally pays for content, MySpace might have a problem.
MySpace, we have a problem!

On the money in social networking
By Matthew Garrahan in Los Angeles
Published: January 28 2009 02:00 | Last updated: January 28 2009 02:00

Chris DeWolfe is dashing around his Beverly Hills office. The co-founder of MySpace is preparing to go to Davos, where he will rub shoulders with leaders of the world economy, including his boss, Rupert Murdoch.

Davos this year is packed with gloomy-sounding sessions on the collapse of global capitalism but Mr DeWolfe is in an upbeat mood. MySpace has fine-tuned its advertising model and as it celebrates its fifth birthday he believes the site can prove its critics wrong about the durability and profit-making potential of social networking.

With their millions of users, social media has long been seen as a panacea of online advertising. But sites such as Facebook and MySpace, which is part of News Corp, have been unable to turn those users into significant profits. MySpace narrowly missed a $1bn revenue target last year, while Facebook has preferred to concentrate on building a large base of users.

By Mr DeWolfe says a relentless focus on profits at MySpace is starting to pay off. "From day one we have always been focused on building a
Networking sites seen hitting web ad growth

By Carlos Granda, Marking Correspondent
Published: July 11 2007 07:28 | Last updated: July 11 2007 07:28

The rise of social networking websites such as Facebook and MySpace could bring the rapid growth of UK internet advertising down a gear this year, according to media forecasters at WPP, the marketing services group.

GroupM, the umbrella forecasting group for WPP's media buying teams, cites the increasing time that audiences are spending on such websites as one reason to anticipate a slowdown in the "warp" speed expansion of web advertising.

Any deceleration in internet growth would be relative, especially compared with the sluggish state of UK broadcast and press advertising.

In a report, GroupM estimates UK internet advertising revenues will still increase by 34 per cent in 2007 before adjusting for inflationary factors 3\% per cent. Following the

Editor's Choice
Ads and labels deal double blow to music sites - Mar-07
Phorm delays internet trial - Sep-04
Analysis: Advertisers try the soft sell as TV drifts online - Mar-07

Senior Fixed Income Managers
UBA Capital
Chief Executive Officer
Cleanbrite
Business Planning Director - Film
Media / Advertising
Head of Planning & Analysis
International TV Network
RECRUITERS
FT.com can deliver talented individuals at industries around the world
Post a job now
Propagation in the context of Blogs

- Blogs are not only for publishing, but a place for online debates and discussions

- For a topic to reach the *tipping point*, what *critical fraction* of bloggers that need to be *actively aware* of it?
Propagation in the context of Blogs

• Blogs are not only for publishing, but a place for online debates and discussions

• For a topic to reach the **tipping point**, what **critical fraction** of bloggers that need to be **actively aware** of it?
Propagation in the context of Blogs

- Blogs are not only for publishing, but a place for online debates and discussions.

- For a topic to reach the **tipping point**, what **critical fraction** of bloggers that need to be **actively aware** of it?
Propagation in the context of Blogs

- Blogs are not only for publishing, but a place for online debates and discussions

- For a topic to reach the tipping point, what critical fraction of bloggers that need to be actively aware of it?
Propagation in the context of Blogs

- Blogs are not only for publishing, but a place for online debates and discussions.
- For a topic to reach the tipping point, what critical fraction of bloggers need to be actively aware of it?
Propagation in the context of Blogs

- Blogs are not only for publishing, but a place for online debates and discussions
- For a topic to reach the tipping point, what critical fraction of bloggers that need to be actively aware of it?
Active Awareness

- Participate in online discussions and debates
- Use hyperlinks during the course of their activity
- Expressed by **direct linking, commenting** and **trackbacks**

<table>
<thead>
<tr>
<th>in-links</th>
<th>out-links</th>
</tr>
</thead>
<tbody>
<tr>
<td>both</td>
<td>none</td>
</tr>
</tbody>
</table>

Interaction is important
Active Awareness

- Participate in online discussions and debates
- Use hyperlinks during the course of their activity
- Expressed by **direct linking, commenting** and **trackbacks**

<table>
<thead>
<tr>
<th>in-links</th>
<th>out-links</th>
</tr>
</thead>
<tbody>
<tr>
<td>both</td>
<td>none</td>
</tr>
</tbody>
</table>
Active Awareness

- Participate in online discussions and debates
- Use hyperlinks during the course of their activity
- Expressed by **direct linking, commenting** and **trackbacks**

<table>
<thead>
<tr>
<th>in-links</th>
<th>out-links</th>
</tr>
</thead>
<tbody>
<tr>
<td>both</td>
<td>none</td>
</tr>
</tbody>
</table>

Interaction is important

Not showing active participation in any online debate or discussion
Active Awareness

- Participate in online discussions and debates
- Use hyperlinks during the course of their activity
- Expressed by **direct linking, commenting** and **trackbacks**

<table>
<thead>
<tr>
<th>in-links</th>
<th>out-links</th>
</tr>
</thead>
<tbody>
<tr>
<td>both</td>
<td>none</td>
</tr>
</tbody>
</table>

Not showing active participation in any online debate or discussion
Blogosphere Link Structure

Differentiate between Friendship (blogroll) links and blog post links

**BL: Blogroll Links**
(friendship, readership)

**EL: Entry Links**

Directed graph, bloggers are vertices
EL Graph
EL Graph

\[ B (Blog) \]

- Component 1
- Component 2
- Component 3

\[ \ldots \]

\[ d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, \ldots, d_m \]
EL Graph

5 blogs

component 1

component 2

component 3

B (Blog)

T (time)

d_1 d_2 d_3 d_4 d_5 d_6 d_7 d_8 d_9 d_{10} d_{11} d_{12} d_{13} \ldots d_m

root

5 blogs

component 1

component 2

component 3

B_1 B_2 B_3 B_4 B_5 B_6 B_7 B_8 B_9 B_{10} B_{11} B_{12} B_{13} B_{14}
EL Graph

5 blogs

component 1

12 blogs

component 2

component 3
EL Graph

B (Blog)

root

component 1

5 blogs

component 2

12 blogs

component 3

3 blogs

B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8, B_9, B_{10}, B_{11}, B_{12}, B_{13}, B_{14}, ..., B_n

d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, ..., d_m

T (time)
Percolation Theory
Percolation Theory

- Theorized in 1957 by Broadbent and Hammersley;
Percolation Theory

- Theorized in 1957 by Broadbent and Hammersley;
- deals with critical phenomena, when properties of system components reach a “tipping point”, and the system goes through major changes
Percolation Theory

- Theorized in 1957 by Broadbent and Hammersley;
- deals with critical phenomena, when properties of system components reach a “tipping point”, and the system goes through major changes

**Examples:**
- small scattered forest fire changes into uncontrollable fire engulfing the whole forest
- change of metal from normal to superconducting form during a cooling process
- spread of rumors in a society
Percolation Theory

Systems tackled by Percolation Theory share a few common features:

• Their physical structures are varied and irregular

• Despite that, in the vicinity of a critical point, all display a similar behaviour

• Local packets of change grow to create a very large one

• At this point, change percolates throughout the system

• Behaviour at the critical point is independent of their structure
Percolation Theory
Percolation Theory

**Lattices** and *regular* geometric structures are used to model percolation in *irregular* systems.
Percolation Theory

**Lattices** and *regular* geometric structures are used to model percolation in *irregular* systems.

Percolation is defined by assigning a probability $p$ of a property to exist in a lattice vertex (*site percolation*), or probability of a bond or link to exist between two lattice vertices (*bond percolation*).
Percolation Theory

**Lattices** and *regular* geometric structures are used to model percolation in *irregular* systems.

Percolation is defined by assigning a probability \( p \) of a property to exist in a lattice vertex (*site percolation*), or probability of a bond or link to exist between two lattice vertices (*bond percolation*).

The value of the critical probability \( (P_c) \) can be calculated for various lattices, either precisely or using numerical and computational methods.
Percolation Theory

(3,12²)  (4,6,12)  (4,8²)  Honeycomb
Kagome  (3,4,6,4)  Square  (3,6)
(3,4,3,4)  (3,4²)  Triangular
Percolation

Bethe Lattice is chosen because:

• It has a root, resembling spread of information from one blog outwards to the rest;

• It has a configurable structure. **Coordination number** ($z$) can be set according to requirements

• Its statistical properties are **solvable**.
In other words:

- We use the regular Bethe lattice to estimate the spread of information in a complicated BL graph
Percolation

Percolation Threshold in Bethe lattice:

\[ P_c = \frac{1}{z - 1} \]

If probability \( p \) of bond connectivity between vertices exceeds \( P_c \), then information will percolate throughout the lattice.

In other words:

\( P_c \) denotes a “tipping point” for the propagation of a topic, and when \( p > P_c \), the topic is said to have percolated throughout the entire community.
Percolation

• The fraction of blogs associated with each component in EL is $p$

• A topic is considered **prevalent or “Global”** if the fraction of bloggers actively aware of it reaches $P_c$
Experiment

with the spinn3r/ICWSM dataset

1. Identification of unique blog URLs
2. Finding blogroll links
3. Creating the BL graph
4. Extracting communities
5. Calculating Percolation Threshold
Experiment

1. Identification of unique blog URLs

2,627,194

Assumption: Each URL represents a person (blogger)
2. Finding blogroll links (by crawling home pages)

- 1.7% failure during crawls: no home page data for 46,418 blogs
- Side note: would be useful to have home page dump in the dataset
  - in regular intervals (e.g. every week or fortnight)
- Around 30% of blogs (778,628) are hosted by LiveJournal, which has a different way of expressing blogroll links
Experiment

2. Finding blogroll links (by crawling home pages)
Experiment

2. Finding blogroll links (by crawling home pages)
Experiment

2. Finding blogroll links (by crawling home pages)

http://jsp.typepad.com
with 50,085 in-links
has highest in-degree
3. Creating the BL graph

<table>
<thead>
<tr>
<th>Vertices</th>
<th>Non-singleton vertices</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,627,194</td>
<td>1,089,223</td>
<td>11,839,763</td>
</tr>
</tbody>
</table>

**Table 4:** Basic statistics of the logroll graph

- **Vertices:** 2,627,194
- **Non-singleton vertices:** 1,089,223
- **Edges:** 11,839,763

![In-degree distribution](image)

In-degree distribution head is scale-free with exponent $\alpha = 1.03$
### Experiment

#### 4. Extracting communities & Calculating Percolation Threshold

<table>
<thead>
<tr>
<th>Community size</th>
<th>Average in-degree</th>
<th>$z$</th>
<th>$P_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,046,063</td>
<td>11.2630</td>
<td>11</td>
<td>0.1000</td>
</tr>
<tr>
<td>1,166</td>
<td>3.3490</td>
<td>3</td>
<td>0.5000</td>
</tr>
<tr>
<td>117</td>
<td>8.5213</td>
<td>9</td>
<td>0.1250</td>
</tr>
<tr>
<td>83</td>
<td>13.1445</td>
<td>13</td>
<td>0.0833</td>
</tr>
<tr>
<td>72</td>
<td>22.6666</td>
<td>23</td>
<td>0.454</td>
</tr>
<tr>
<td>64</td>
<td>2.6093</td>
<td>3</td>
<td>0.5000</td>
</tr>
<tr>
<td>60</td>
<td>23.60</td>
<td>24</td>
<td>0.0434</td>
</tr>
<tr>
<td>56</td>
<td>5.750</td>
<td>6</td>
<td>0.2000</td>
</tr>
<tr>
<td>50</td>
<td>44.22</td>
<td>44</td>
<td>0.0232</td>
</tr>
<tr>
<td>49</td>
<td>6.2653</td>
<td>6</td>
<td>0.2000</td>
</tr>
</tbody>
</table>
## Experiment

### 4. Extracting communities & Calculating Percolation Threshold

<table>
<thead>
<tr>
<th>Community size</th>
<th>Average in-degree</th>
<th>$z$</th>
<th>$P_c$</th>
<th>identifier URL</th>
<th>identified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,046,063</td>
<td>11.2630</td>
<td>11</td>
<td>0.1000</td>
<td>jsp.typepad.com</td>
<td>blog</td>
</tr>
<tr>
<td>1,166</td>
<td>3.3490</td>
<td>3</td>
<td>0.5000</td>
<td>financeusedcar.wordpress.com</td>
<td>splog</td>
</tr>
<tr>
<td>117</td>
<td>8.5213</td>
<td>9</td>
<td>0.1250</td>
<td>sellandtellonline.com</td>
<td>?</td>
</tr>
<tr>
<td>83</td>
<td>13.1445</td>
<td>13</td>
<td>0.0833</td>
<td>megalizardfan.livejournal.com</td>
<td>blog</td>
</tr>
<tr>
<td>72</td>
<td>22.6666</td>
<td>23</td>
<td>0.454</td>
<td>cbkopf.exblog.jp</td>
<td>splog</td>
</tr>
<tr>
<td>64</td>
<td>2.6093</td>
<td>3</td>
<td>0.5000</td>
<td>ownlink228.cocolog-nifty.com/blog</td>
<td>splog</td>
</tr>
<tr>
<td>60</td>
<td>23.60</td>
<td>24</td>
<td>0.0434</td>
<td>beaconbade.livejournal.com</td>
<td>blog</td>
</tr>
<tr>
<td>56</td>
<td>5.750</td>
<td>6</td>
<td>0.2000</td>
<td>brummitt89.livejournal.com</td>
<td>blog</td>
</tr>
<tr>
<td>50</td>
<td>44.22</td>
<td>44</td>
<td>0.0232</td>
<td>tm675345.livejournal.com</td>
<td>splog</td>
</tr>
<tr>
<td>49</td>
<td>6.2653</td>
<td>6</td>
<td>0.2000</td>
<td>debbie-g.livejournal.com</td>
<td>blog</td>
</tr>
</tbody>
</table>
What Next?
What Next?

◦ Set up of the EL graph
What Next?

- Set up of the EL graph
- Extract WCCs and count number of blogs involved in each
What Next?

- Set up of the EL graph
- Extract WCCs and count number of blogs involved in each
- Compare the fraction of blogs to $P_c$
What Next?

- Set up of the EL graph
- Extract WCCs and count number of blogs involved in each
- Compare the fraction of blogs to $P_c$
Discussion
Discussion

‣ Evaluation

‣ Correlation between our measurements and outcomes of existing systems

‣ How much information we’re losing by not considering content?

‣ Separate components discussing the same topic!
Discussion

- Evaluation
  - Correlation between our measurements and outcomes of existing systems
  - How much information we’re losing by not considering content?
  - Separate components discussing the same topic!

- How precise is:
  - Estimation using the Bethe lattice?
  - Using average in-degree as coordination number?
Future Work
Future Work

- **Prediction**
  - Find, in an early stage, the topics that might become global

  "Topic $T_1$ will become 85% global with probability 0.73"
Future Work

- Prediction
  - Find, in an early stage, the topics that might become global
    
    “Topic $T_1$ will become $85\%$ global with probability $0.73$”

- Topics
  - What are the actual topics?
Future Work

‣ Prediction
  ▶ Find, in an early stage, the topics that might become global
    
    “Topic $T_1$ will become 85% global with probability 0.73”

‣ Topics
  ▶ What are the actual topics?

‣ Non-blog domain
  ▶ YouTube video, academic citations
Thank You!

Acknowledgements

University of Glasgow

Microsoft Research Cambridge

spinn3r For the dataset