Community Structure and Information Flow in Usenet: Improving Analysis with a Thread Ownership Model

Mary McGlohon, Carnegie Mellon*
Matthew Hurst, Microsoft

*work completed at Microsoft
Motivation

• Comparing communities of online social networks may lend insight into how groups form and thrive
• We would also like to understand how information diffuses between groups
Why Usenet?

• We delve into these questions by analyzing data from Usenet
• Public
• Can be analyzed over a long time period
• Has pre-defined, hierarchical community structure
• Two main goals:
  ◦ Compare different group activity (size, reciprocity)
  ◦ Observe diffusion between groups
Data

- Posts from 200 politically-oriented newsgroups (bulletin boards)
  - “polit” in name
- January 2004–June 2008
- Several countries, state/provinces, and topics.
- 19.6 million unique articles, 6.2 million cross-posted
Cross-posting

• A large percentage of articles are cross-posted to multiple groups.
• Somebody reading one group may “reply-to-all”, such that all groups see it.

Major issue: many are cross-posted to multiple groups. Where is conversation truly occurring?
Outline

• Motivation
• Data description
• Structural Analysis
  ▫ Size
  ▫ Reciprocity
  ▫ Similarity
• Ownership model
  ▫ Effects of Cross-posting
  ▫ Information Flow based on Ownership
  ▫ Similarity
Structural Analysis

• We hope to compare the structure of communities by answering the following questions:

• How do edges form?
• How does the reciprocity of groups compare?
• How can we measure similarity?
Sizes of groups

- How do edges form?

- To answer, we make a network of authors for each group

- If \( a_1 \) has replied to \( a_2 \) at any point, there is an edge from \( a_1 \) to \( a_2 \)
Sizes of groups

- Power law-like relationship between number of authors and number of edges.

- Similar to densification law [Leskovec+05], only with individual networks instead of snapshots of a network over time.
Reciprocity

• Which groups have highest reciprocity?
• Reciprocity: percentage of reply-edges that are mutual
• Top 10 were European newsgroups (up to 0.58):
  ▫ hun.politika
  ▫ relcom.politics
  ▫ hsv.politics
  ▫ italia.modena.politica
  ▫ se.politik
  ▫ it.discussioni.leggende.metropolitane
  ▫ ukr.politics
  ▫ yu.forum.politika
  ▫ ni.politics
  ▫ swnet.politik
• Lowest reciprocity occurred in tw.bbs.* (<0.1)
Similarity

• How can we measure similarity between groups?
• Use Jaccard coefficient for cross-posts:

\[
\text{# Shared articles (cross-posts) between 2 groups} \over \text{Total number of articles in groups}
\]

• Can do the same with shared authors
• Highest similarity \( \sim 0.54 \) (bc.politics and ont.politics)
Similarity

• Each group is a node
• Edge drawn if similarity > 0.1 (thick edge >0.2)
• Form clusters: parties, US regional, countries, alt.politics subgroups
Parties/topics

- alt.politics.india.communist
- alt.politics.radical-left
- alt.politics.libertarian
- alt.politics.socialism
- alt.politics.org.nsa
- alt.politics.usa.constitution.gun-rights
- alt.politics.org.fbi
- talk.politics.libertarian
- alt.politics.org.cia
- alt.politics.congress
- talk.politics.guns
- uk.politics.guns
- talk.politics.drugs
- uk.politics.drugs
US States
English-speaking countries
alt.politics.*
Outline

• Motivation
• Data Description
• Structural Analysis
  ▫ Size
  ▫ Reciprocity
  ▫ Similarity
• Ownership model
  ▫ Information Flow based on Ownership
  ▫ Similarity
Problem: Excessive cross-posting

- We just saw that there is significant overlap between groups in terms of articles.
- However, cross-posting occurs often between unrelated groups ("edges below threshold").
- We would like to find out in which group the activity is truly occurring.

- How can we trace this?
Solution: Thread Ownership

- Answer: Assign “ownership” based on the authors of the posts
- First, assign authors to groups based on devotion
  - $Devotion(a, g)$: what percentage of an author $a$’s posts are exclusively posted to a given group $g$
- For each post, normalize devotion among groups where the post occurs.
  - Group with highest devotion score for the author has more “ownership” of a post
Example: Thread Ownership

- Suppose in the data authors have the following numbers of non-cross-posts in each group:

<table>
<thead>
<tr>
<th></th>
<th>alt.politics</th>
<th>us.politics</th>
<th>pa.politics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author 1</td>
<td>6</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Author 2</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Author 3</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

- Then, they form a thread:
Real thread

- Initially cross-posted to several groups (including talk.politics.misc), 38 groups in total
- Ownership concentrated in seattle.politics and or.politics
- Subject: “Kiss the National Parks Good-Bye”
Applications of thread ownership

- Ownership model aids in analyzing threads
  - Influence between groups: How are threads discovered and posted to new groups?
  - Similarity of groups: How can ownership help us more precisely state when two groups are similar?
Information flow between groups

- How are threads discovered and posted to new groups?
- Idea: Extend ownership to influence

- How often does an author in group 1 respond to a post they found in group 2?
  - Author finds parent post \( p_p \) by browsing group \( g_p \)
  - Author writes child post \( p_c \) to group \( g_c \)
  - Then, we say \( g_p \) influences \( g_c \)

\[
\text{Influence}(g_p, g_c) = \text{Devotion}(a, g_p) \times \text{Devotion}(a, g_c)
\]

- This helps pinpoint when an author decides to cross-post late in the thread
Example: Ownership-based influence

- Author 2 sees parent post
- Replies, adding **pa.politics**.
- Since Author 2 is not devoted to **alt.politics**, he was most likely influenced by **us.politics**
- $\text{Influence}(\text{us.politics}, \text{pa.politics}) = 1 \times 0.75 = 0.75$

<table>
<thead>
<tr>
<th></th>
<th>alt.politics</th>
<th>us.politics</th>
<th>pa.politics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author 1</td>
<td>6</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Author 2</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
Who influences whom?

- Information often diffuses from major to minor groups

<table>
<thead>
<tr>
<th>Influencer</th>
<th>Influencee</th>
</tr>
</thead>
<tbody>
<tr>
<td>swnet.politik</td>
<td>se.politik.diverse</td>
</tr>
<tr>
<td>de.soc.politik.misc</td>
<td>bln.politik.rassismus</td>
</tr>
<tr>
<td>can.politics</td>
<td>man.politics</td>
</tr>
<tr>
<td>can.politics</td>
<td>ab.politics</td>
</tr>
<tr>
<td>can.politics</td>
<td>bc.politics</td>
</tr>
<tr>
<td>can.politics</td>
<td>ont.politics</td>
</tr>
<tr>
<td>uk.politics.misc</td>
<td>uk.politics.constitution</td>
</tr>
<tr>
<td>uk.politics.misc</td>
<td>uk.politics.parliament</td>
</tr>
<tr>
<td>talk.politics.drugs</td>
<td>uk.politics.drugs</td>
</tr>
</tbody>
</table>
Ownership-based Similarity

• Q: How can ownership help us more precisely state when two groups are similar?
• A: Use “shared ownership” instead of shared posts

Western states

Eastern states
Applications and future work

• Potential Applications
  ▫ Link prediction
  ▫ Information retrieval and relevance
  ▫ Ownership for email lists

• Future Work
  ▫ Using comparative measures to predict whether group will continue
Related work: Discussion Groups

- Gomez, V.; Kaltenbrunner, A.; and Lopez, V. 2008. Statistical analysis of the social network and discussion threads in slashdot. WWW ’08
Related work: Information Diffusion

- Kossinets, G.; Kleinberg, J.; and Watts, D. 2008. The structure of information pathways in a social communication network. KDD’08
- Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. KDD ’05
Conclusions

• Case study of nearly 200 newsgroups, including 19 million unique posts
• Demonstrated “densification” law as applies to different groups
• Compared groups in terms of reciprocity and shared posts/authors
• Proposed thread ownership model to cut down on “noise” from cross-posts
• Applied ownership to diffusion between groups, group similarity
Contact info

- Mary McGlohon
- www.cs.cmu.edu/~mmcgloho
- Matthew Hurst
- datamining.typepad.com

- Special thanks to Christos Faloutsos, Michael Gamon, Kathy Gill, Christian Konig, Alexei Maykov, Purna Sarkar, Hassan Sayyadi, Marc Smith