All who wander:
On the Prevalence and Characteristics of Multi-community Engagement

Chenhao Tan
Cornell University
Joint work with Lillian Lee
We have many chances to engage with many communities

A variety of organizations, and social circles exist on a university campus
We have many chances to engage with many communities

A variety of organizations, and social circles exist on a university campus

Computer science conferences (DBLP)
We have many chances to engage with many communities

A variety of organizations, and social circles exist on a university campus

Computer science conferences (DBLP)

Online communities, e.g., reddit.com

Babybumps
Existing work

Single community setting: e.g., predicting user survival (churn prediction) 

Objects of interest: user trajectories across communities

An example from a user on Reddit
How much do users explore new communities?

#communities

Age
First 50 posts on Reddit and DBLP

The average time to accumulate 50 contributions is 456.0 days on Reddit, 15.6 years on DBLP.
Lifetime on Reddit and DBLP

Reddit

DBLP

Error bars (tiny) show standard error.
Main dataset: reddit

• Many sub-communities (subreddits)
• An active platform where users submit posts, make comments and rate posts

• All 76.6M posts ever submitted to Reddit from its inception until Jan 2014
• 157K “50+” posters who first posted between Jan 2008 and Jan 2012 [Danescu-Niculescu-Mizil et al. 2013]

Link: https://chenhaot.com/pages/multi-community.html
Three aspects of user trajectories

• (How) does the wandering pattern change over time?

• Do people adapt their language in each community over time?

• Do people receive better evaluations over time?

Can these be used to predict future user activity?
A framework for understanding properties of the trajectory

Split the trajectory into windows of the same size (w=10 in main experiments)

Define a function \( (F) \) on a window to capture different properties and obtain a time series

\[ \text{Community}_t, \text{words}_t, \text{evaluations}_t \]
Do users “jump” more over time?

\[ F: \text{Count of } \text{Community}_t \neq \text{Community}_{t-1} \text{ in a window (w=10)} \]

![Diagram showing transitions between different communities.](image-url)
Do users “jump” more over time?

\[ F: \text{Count of } Community_t \neq Community_{t-1} \text{ in a window (w=10)} \]

\[ \begin{align*}
Jumps &= 1 \\
\text{pentagon } \rightarrow \text{ triangle } \rightarrow \text{ pentagon } \rightarrow \text{ triangle } \rightarrow 6 \text{ more } \cdots
\end{align*} \]

\[ \begin{align*}
Jumps &= 3 \\
\text{pentagon } \rightarrow \text{ triangle } \rightarrow \text{ pentagon } \rightarrow \text{ triangle } \rightarrow 6 \text{ more } \cdots
\end{align*} \]
Users “jump” more over time

\[ F: \text{Count of } Community_t \neq Community_{t-1} \text{ in a window } (w=10) \]

Error bars (tiny) show standard error.
Users “jump” more over time; future departing users less so

Future departing users: stopped posting in the entire reddit in the last 6 months (44K)

Future staying users: stay active in the last 6 months (76K)

Error bars (tiny) show standard errors.
Users get more adventurous over time; future departing users less so

Many more different perspectives on the wandering pattern:
- number of unique communities, \(\uparrow\)
- level of concentration, \(\downarrow\)
- visible community size, \(\downarrow\)
- community similarity, \(\downarrow\)

In our data, people do not settle down at all!
Users keep adopting each community’s language; future departing users more so.

**F**: average cross entropy of $\text{words}_t$ vs language in $\text{Community}_t$.

A larger value indicates larger dissimilarity.

Users stay young:
Different from “users get old” in single community setting

[Danescu-Niculescu-Mizil et al. 2013]
Users get more positive evaluations over time; future departing users less so

\[ F: \text{fraction of } \text{evaluations}_t \text{ that outperform the median in } \text{Community}_t \]

A larger value indicates better evaluations.
Using the first 50 posts to predict future departing status

• Feature sets
  – Wandering pattern
  – Language
  – Evaluations
  – Combination of the above features

• 30 randomized train-test samples, logistic regression, F1 on departing users for evaluation

• Measure performance using the first x posts
Features from trajectories outperform time-gap baseline

Users are destined to leave from the beginning!

Features from first 10 posts outperform baseline with all 50 posts.

All differences along x-axis are significant (p<0.001) according to Wilcoxon signed rank test.
Is recent information more important or how you start more important?

First 10, 20, 30, 40, 50 vs Last 10, 20, 30, 40, 50

All differences along x-axis are significant (p<0.001) according to Wilcoxon signed rank test.

How you start is more important!
Do people speak differently in different communities?

Focus on non-content words for language style


<table>
<thead>
<tr>
<th>$V$</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>parts of speech</td>
<td>62.5%</td>
</tr>
<tr>
<td>most frequent 100 words</td>
<td>56.0%</td>
</tr>
<tr>
<td>most frequent 500 words</td>
<td>61.4%</td>
</tr>
</tbody>
</table>
Summary

• Users’ multi-community engagement is an interesting problem; lots of room for future work

• Design implications:
  – First impressions matter
  – Give people choices to move to

• Life lessons: People, unlike trees, thrive by relocation

人挪活，树挪死

Thank you!

Contact: chenhao@cs.cornell.edu, @ChenhaoTan
Data: https://chenhaot.com/pages/multi-community.html