No country for old members: User lifecycle and linguistic change in online communities

Cristian Danescu-Niculescu-Mizil
Robert West
Dan Jurafsky
Jure Leskovec
Christopher Potts

Stanford University
Max Planck Institute SWS
User-community relation

Our high-level goal:
Analyzing the relation between an user and an online community
User-community relation

Our high-level goal:
Analyzing the relation between an user and an online community

Joining
User-community relation

Our high-level goal:
Analyzing the relation between an user and an online community
Our high-level goal:
Analyzing the relation between an user and an online community

Joining — Abandoning
User-community relation

Concrete questions we address:

→ How does a user become member of a community?
User-community relation

Concrete questions we address:

→ How does a user become member of a community?

→ How do user & community practices co-evolve?
User-community relation

Concrete questions we address:

→ How does a user become member of a community?

→ How do user & community practices co-evolve?

→ Can we predict when a user will leave the community?
Main insight: linguistic change

Language practices (norms, etiquette, ... )
→ build collective identity
→ foster individual expression
Main insight: linguistic change

Language practices (norms, etiquette, ...) → build collective identity
→ foster individual expression

Linguistic change allows us to capture
→ relation between members and their community
Our approach: linguistic change

→ Statistical framework for tracking linguistic change
Our approach: linguistic change

→ Statistical framework for tracking linguistic change
→ Measures of user reaction to linguistic change
Our approach: linguistic change

→ Statistical framework for tracking linguistic change

→ Measures of user reaction to linguistic change

→ Features predicting when user will leave the community
Longitudinal data

Complete linguistic record of two online communities:

data available at http://snap.stanford.edu/data/
Longitudinal data

Complete linguistic record of two online communities:

**Beeradvocate**

2001 (inception)  
2011

10 years of complete linguistic record  
1,600,000 posts  
33,000 users

data available at http://snap.stanford.edu/data/
Longitudinal data

Complete linguistic record of **two** online communities:

![RateBeer logo]

10 years of complete linguistic record
3,000,000 posts
30,000 users

Data available at http://snap.stanford.edu/data/
Linguistic change: one example
Linguistic change: one example

... Aroma: Buttery, slightly spicy malt notes ...
Linguistic change: one example

... Aroma: Buttery, slightly spicy malt notes ...

![Graph showing the usage of 'Aroma' convention over years from 2001 to 2011]
... **Aroma**: Buttery, slightly spicy malt notes ...

... **S**: Great nose of ginger, honey, perfume ...

![Graph showing convention usage over years](image)
### Users joining in 2003

<table>
<thead>
<tr>
<th>Year</th>
<th>'Aroma' convention</th>
<th>'Smell' convention</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>2002</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>2003</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>2004</td>
<td>6%</td>
<td>6%</td>
</tr>
<tr>
<td>2005</td>
<td>8%</td>
<td>8%</td>
</tr>
<tr>
<td>2006</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>2007</td>
<td>12%</td>
<td>12%</td>
</tr>
<tr>
<td>2008</td>
<td>14%</td>
<td>14%</td>
</tr>
<tr>
<td>2009</td>
<td>16%</td>
<td>16%</td>
</tr>
<tr>
<td>2010</td>
<td>18%</td>
<td>18%</td>
</tr>
<tr>
<td>2011</td>
<td>20%</td>
<td>20%</td>
</tr>
</tbody>
</table>

### Users joining in 2005

<table>
<thead>
<tr>
<th>Year</th>
<th>'Aroma' convention</th>
<th>'Smell' convention</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>2002</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>2003</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>2004</td>
<td>6%</td>
<td>6%</td>
</tr>
<tr>
<td>2005</td>
<td>8%</td>
<td>8%</td>
</tr>
<tr>
<td>2006</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>2007</td>
<td>12%</td>
<td>12%</td>
</tr>
<tr>
<td>2008</td>
<td>14%</td>
<td>14%</td>
</tr>
<tr>
<td>2009</td>
<td>16%</td>
<td>16%</td>
</tr>
<tr>
<td>2010</td>
<td>18%</td>
<td>18%</td>
</tr>
<tr>
<td>2011</td>
<td>20%</td>
<td>20%</td>
</tr>
</tbody>
</table>

---

... **Aroma**: Buttery, slightly spicy malt notes ...

... **S**: Great nose of ginger, honey, perfume ...

---

Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, Potts

No country for old members
Community-level change and user-level change

Other examples of community-level changes:
Re-tweet convention on Twitter, slang in hip-hop forums
[Romero et al. 2011; Kooti et al. 2012; Garley and Hockenmaier 2012; inter alia]
Community-level change and user-level change

Joining Abandoning
"life stage"

2001 2011
Community-level change and user-level change

Example of user-level change:
Decrease in usage of 1st person pronouns (e.g., I, me, mine, myself)

A sign of increasing identification with the community [Pennebaker 2007; Sherblom 2009]

Joining — “life stage” — Abandoning

No country for old members
Community-level change and user-level change

The rest of this talk: relation between these two levels of change

Joining  “life stage”  Abandoning

2001  2011

Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, Potts

No country for old members
Distance from the community
Distance from the community

2001

2011

No country for old members
Distance from the community

2001

2011

Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, Potts

No country for old members
Distance from the community

Distance from the community over time from 2001 to 2011.
Distance from the community

Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, Potts

No country for old members
Language model of the community in JAN 2006 “snapshot language model” (SLM$_{\text{JAN 2006}}$)
Distance from the community

Language model of the community in JAN 2006 “snapshot language model” (SLM_{JAN 2006})

cross-entropy of \( p \) according to SLM_{JAN 2006}:

\[
H(p, SLM_{m(p)}) = -\frac{1}{N} \sum_i \log P_{SLM_{m(p)}}(b_i)
\]
Distance from the community

Life stage

Distance from the community

0% 25% 50% 75% 100%
Stage 1: user **assimilates** the language of the community

Distance from the community

<table>
<thead>
<tr>
<th>Life stage</th>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>45</td>
<td>40</td>
<td>35</td>
<td>30</td>
</tr>
</tbody>
</table>

**Distance from the community**
Stage 1: user assimilates the language of the community
Distance from the community

Stage 1: user **assimilates** the language of the community

Stage 2: User’s language **distances** itself from that of the community
Hypothesis 1: User moves away from the community by using innovative language
Hypothesis 1: User moves away from the community by using innovative language

Hypothesis 2: User **stops adapting** and gets out of tune with the changing community
User language stability

Hypothesis 1: User moves away from the community by using innovative language

Hypothesis 2: User stops adapting and gets out of tune with the changing community
User language stability

Hypothesis 1: User moves away from the community by using innovative language

Hypothesis 2: User stops adapting and gets out of tune with the changing community

Compare user language with her past language
User language stability

Compare user language with her past language
User language stability

Compare user language with her past language

![Graph showing user language stability over life stages.](image-url)
User language stability

Compare user language with her past language

Confirms Hypothesis 2: before abandoning, users stop adapting
Adoption of lexical innovation
Adoption of lexical innovation

Lexical innovation:
• new word that get picked up by the community
• about 100 lexical innovations each month

<table>
<thead>
<tr>
<th>Type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventions</td>
<td>S[mell], M[outhfeel], FLAVOR</td>
</tr>
<tr>
<td>Descriptive</td>
<td>sandalwood, gummy, rubbery</td>
</tr>
<tr>
<td>Other</td>
<td>verdict, mysterious, nothingness</td>
</tr>
</tbody>
</table>

• user “adopts lexical innovation” if uses such a word in the 3 months after its introduction
Adoption of lexical innovation

Lexical innovation:
• new word that get picked up by the community

Peak reaction to change
Puzzle answer

Users joining in 2005 still in their flexible-language stage
Puzzle answer

Users joining in 2005 still in their **flexible-language stage**

Users joining in 2003 in the **rigid-language phase**
User lifecycle (summary)

Online linguistic lifecycle

0% User joins the community

**Stage 1: adaptation to community norms**
30% Peak receptivity to community norms

**Stage 2: linguistic patterns rigidify**
100% User abandons the community
User lifecycle (summary)

Online linguistic lifecycle
0% User joins the community
**Stage 1: adaptation to community norms**
30% Peak receptivity to community norms
**Stage 2: linguistic patterns rigidify**
100% User abandons the community

Offline linguistic lifecycle [Labov, 1966]
Birth Individual joins the community
**Stage 1: linguistic assimilation**
17 years Peak receptivity to community norms
**Stage 2: “adult language stability”**
Individual leaves the community
User lifecycle (summary)

Online linguistic lifecycle

- 0% User joins the community
- **Stage 1: adaptation to community norms**
  - 30% Peak receptivity to community norms
- **Stage 2: linguistic patterns rigidify**
  - 100% User abandons the community

Offline linguistic lifecycle [Labov, 1966]

- Birth Individual joins the community
- **Stage 1: linguistic assimilation**
  - 17 years Peak receptivity to community norms
- **Stage 2: “adult language stability”**
  - Individual leaves the community

Absolute time-frame, assumed biological effect
User lifecycle (summary)

Online linguistic lifecycle

0% User joins the community

Stage 1: adaptation to community norms

30% Peak receptivity to community norms

Stage 2: linguistic patterns rigidify

100% User abandons the community

Offline linguistic lifecycle [Labov, 1966]

Birth Individual joins the community

Stage 1: linguistic assimilation

17 years Peak receptivity to community norms

Stage 2: “adult language stability”

Individual leaves the community

Relative time-frame, suggesting social effect

Absolute time-frame, assumed biological effect

Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, Potts
No country for old members
Elastic lifecycle

Lifecycle *stretches* according to the user’s ultimate lifespan
Elastic lifecycle

Lifecycle stretches according to the user’s ultimate lifespan

Users with vastly different lifespans
Elastic lifecycle

Lifecycle **stretches** according to the user’s ultimate lifespan

![Graph showing lifecycle stretches](image)

- **Prob. of adopting innovation**
- **Review number**
- **Break down by user ultimate lifespan**
Lifecycle stretches according to the user’s ultimate lifespan

Break down by user ultimate lifespan
Elastic lifecycle

Lifecycle stretches according to the user’s ultimate lifespan

Break down by user ultimate lifespan
Elastic lifecycle

Lifecycle *stretches* according to the user’s ultimate lifespan

→ Similar lifecycle in spite of vastly different lifespans
  “All users die old”
Elastic lifecycle

Lifecycle stretches according to the user’s ultimate lifespan

- Similar lifecycle in spite of vastly different lifespans
  “All users die old”

- End of Stage 1 is a function of the ultimate lifespan of the user
  (not tied to an absolute timeframe e.g., 60 reviews or 1 year)
Elastic lifecycle

Lifecycle stretches according to the user’s ultimate lifespan

→ Similar lifecycle in spite of vastly different lifespans
  “All users die old”

→ End of Stage 1 is a function of the ultimate lifespan of the user
  (not tied to an absolute timeframe e.g., 60 reviews or 1 year)

→ Level of receptivity is correlated with the ultimate lifespan of the user
Elastic lifecycle

Lifecycle *stretches* according to the user’s ultimate lifespan

- Similar lifecycle in spite of vastly different lifespans
  “All users die old”

- End of Stage 1 is a function of the ultimate lifespan of the user
  (not tied to an absolute timeframe e.g., 60 reviews or 1 year)

- Level of receptivity is correlated with the ultimate lifespan of the user
Elastic lifecycle

Lifecycle stretches according to the user’s ultimate lifespan

- Similar lifecycle in spite of vastly different lifespans
  “All users die old”

- End of Stage 1 is a function of the ultimate lifespan of the user
  (not tied to an absolute timeframe e.g., 60 reviews or 1 year)

- Level of receptivity is correlated with the ultimate lifespan of the user

Predict ultimate lifespan
Predicting user lifespan

Task: Given the first 20 posts, will the user abandon the community soon?
Predicting user lifespan

Task: Given the first 20 posts, will the user abandon the community soon?

Linguistic change features:
- distance from the community
- language stability
- adoption of lexical innovations
Predicting user lifespan

Task: Given the first 20 posts, will the user abandon the community soon?

Linguistic change features:
- distance from the community
- language stability
- adoption of lexical innovations

Baselines:
- post frequency ← previous work on churn prediction [Dror et al. 2012, Yang et al. 2010]
- post month ← accounts for community-wide changes
Predicting user lifespan

Task: Given the first 20 posts, will the user abandon the community soon?

Linguistic change features:
- distance from the community
- language stability
- adoption of lexical innovations

Baselines:
- post frequency ← previous work on churn prediction
  [Dror et al. 2012, Yang et al. 2010]
- post month ← accounts for community-wide changes

Logistic regression: One community for development, the other for test
## Predicting user lifespan

<table>
<thead>
<tr>
<th>Features</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>30.5</td>
</tr>
<tr>
<td>+ Distance from the community</td>
<td>37.4</td>
</tr>
<tr>
<td>+ Language stability</td>
<td>38.0</td>
</tr>
<tr>
<td>+ Adoption of lexical innovation</td>
<td>40.9</td>
</tr>
<tr>
<td>+ First person singular pronouns</td>
<td>41.2</td>
</tr>
<tr>
<td>+ Number of words</td>
<td>42.9</td>
</tr>
</tbody>
</table>

Results: Up to 12% absolute (40% relative) improvement
Conclusions

→ framework for tracking linguistic change
→ revealed an elastic two-stage lifecycle
→ exploited for predicting user abandonment
→ co-evolution of users and their communities
Thank you!

Data, slides, and more available at:
www.mpi-sws.org/~cristian