Trajectories of Blocked Community Members:

Jonathan P. Chang
and
Cristian Danescu-Niculescu-Mizil
Cornell University
Trajectories of Blocked Community Members:

Jonathan P. Chang
and
Cristian Danescu-Niculescu-Mizil
Cornell University
Can we tell which path will be taken?
Can we tell which path will be taken?

User characteristics?
Account age, amount of interaction, etc.

Ribeiro et al. (2018)
Cheng et al. (2017)
D-N-M et al. (2013)
Halfaker et al. (2011)
and more...
Can we tell which path will be taken?

- User characteristics?
  - Account age, amount of interaction, etc.
  - Ribeiro et al. (2018)
  - Cheng et al. (2017)
  - D-N-M et al. (2013)
  - Halfaker et al. (2011)
  - and more...

- Mod action context?
  - How severe was the moderator’s action? How does the user react?
  - Corbett-Davies et al. (2017)
  - Tonry (2008)
  - Makkai & Braithwaite (1994)
  - Grasmik & Bryjak (1980)
  - and more...
Redemption
Recidivism
Departure
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”

Moderator response: blocking
Prevents user from making edits (except own talk page)
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”

Moderator response: **blocking**
Prevents user from making edits (except own talk page)

52% Redemption
Recidivism 18%
Departure 30%
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”

Moderator response: blocking
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log

52% Redemption
Recidivism 18%
Departure 30%
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”

Moderator response: blocking
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log
Limit to antisocial behavior: incivility, harassment, edit warring, or disruptive editing

52% Redemption
Recidivism 18%
Departure 30%
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”

Moderator response: blocking
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log
Limit to antisocial behavior: incivility, harassment, edit warring, or disruptive editing
Combine with Wikiconv dataset (Hua et al., 2018)
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”

Moderator response: blocking
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log
Limit to antisocial behavior: incivility, harassment, edit warring, or disruptive editing
Combine with Wikiconv dataset (Hua et al., 2018)
Filter out bots/spam accounts with activity filter

52% Redemption

Recidivism 18%

Departure 30%
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”

Moderator response: blocking
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log
Limit to antisocial behavior: incivility, harassment, edit warring, or disruptive editing
Combine with Wikiconv dataset (Hua et al., 2018)
Filter out bots/spam accounts with activity filter
Final data size: 6,026 blocked users
Making meaningful comparisons

Departing during block vs. Staying on Wikipedia
Making meaningful comparisons

“Natural” departure
Departing during block vs. Staying on Wikipedia
Matching lets us make nontrivial comparisons between departure and redemption...
Matching lets us make nontrivial comparisons between departure and redemption...

“Natural” departure vs. Staying on Wikipedia

D
(departs during block)
Making meaningful comparisons

Matching lets us make nontrivial comparisons between departure and redemption...

Experimental Pair

- **D**: (departs during block)
- **Reformed user**: (blocked around same time as D)

```
"Natural" departure

Departing during block

vs.

Staying on Wikipedia
```
Making meaningful comparisons

Matching lets us make nontrivial comparisons between departure and redemption...

...and a second level of matching ensures those comparisons are specific to block departures.
Matching lets us make nontrivial comparisons between departure and redemption...

...and a second level of matching ensures those comparisons are specific to block departures.

Experimental Pair
- **D** (departs during block)
- Reformed user (blocked around same time as D)

Control Pair
- **N** (never-blocked user, departs around same time as D)

“Natural” departure
- Departing during block

Staying on Wikipedia
Matching lets us make nontrivial comparisons between departure and redemption...

...and a second level of matching ensures those comparisons are specific to block departures

**Experimental Pair**

- **D** (departs during block)
- Reformed user (blocked around same time as **D**)

**Control Pair**

- **N** (never-blocked user, departs around same time as **D**)
- Clean user (active at/beyond time of **N**’s departure)

“Natural” departure

Departing during block

Staying on Wikipedia
Matching lets us make nontrivial comparisons between departure and redemption...

...and a second level of matching ensures those comparisons are specific to block departures.

Experimental Pair

D (departs during block) vs. Reformed user (blocked around same time as D)

Control Pair

N (never-blocked user, departs around same time as D) vs. Clean user (active at/beyond time of N’s departure)

"Natural" departure vs. Staying on Wikipedia

(analogous process for recidivist vs reformed users)
Redemption

Recidivism

Can we tell which path will be taken?

User characteristics?
- Account age, amount of interaction, etc.

Mod action context?
- How severe was the moderator’s action? How does the user react?

Ribeiro et al. (2018)
Cheng et al. (2017)
D-N-M et al. (2013)
Halfaker et al. (2011)
and more...

Corbett-Davies et al. (2017)
Tonry (2008)
Makkai & Braithwaite (1994)
Grasmik & Bryjak (1980)
and more...
Can we tell which path will be taken?

User characteristics?
Account age, amount of interaction, etc.
Ribeiro et al. (2018)
Cheng et al. (2017)
D-N-M et al. (2013)
Halfaker et al. (2011)
and more...

Mod action context?
How severe was the moderator’s action? How does the user react?
Corbett-Davies et al. (2017)
Tonry (2008)
Makkai & Braithwaite (1994)
Grasmik & Bryjak (1980)
and more...
What characteristics matter?

Prior work: norm violations correlate with level of involvement in community

Simple measure: number of talk page comments ("activity level")
What characteristics matter?

Prior work: norm violations correlate with level of involvement in community

Simple measure: number of talk page comments ("activity level")

contribution
What characteristics matter?

Prior work: norm violations correlate with level of involvement in community

Simple measure: number of talk page comments (“activity level”)

contributed

received
Activity level vs departure

Users who depart tend to have lower activity level
- Intuitive interpretation: less involvement → less reason to stay
Activity level vs departure

Users who depart tend to have lower activity level
- Intuitive interpretation: less involvement → less reason to stay
Beyond activity level

Activity level measures *amount* of involvement, but not *nature* of involvement.

For the latter, need *activity spread*.
Activity Spread

Example User has written 100 comments
Activity Spread

Example User has written 100 comments
Activity Spread

Example User has written 100 comments
Activity Spread

Example User has written 100 comments

\[
\text{contributed activity spread} = \frac{\text{# comments written}}{\text{# other user talk pages written to}}
\]
Activity Spread

*Example User* has received 100 comments

\[
\text{received activity spread} = \frac{\# \text{ comments received}}{\# \text{ unique comment authors}}
\]
Activity spread vs departure

Users who depart tend to have higher received activity spread

- Possible intuitive interpretation: less tightly integrated into a social circle
Activity spread vs departure

Users who depart tend to have higher received activity spread
- Possible intuitive interpretation: less tightly integrated into a social circle
Engagement features: predicting trajectories

Can the engagement measures be used to predict a blocked user’s future trajectory?

Methodology: use engagement measures as features to SVM

- Separate models for predicting departure and recidivism

Baselines: block reason, block duration

Also consider how long user has been active (“community age”)

- Basic measure of engagement
Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data. *s indicate significant (p < 0.05) improvement over best baseline in column.

<table>
<thead>
<tr>
<th></th>
<th>Departure</th>
<th>Recidivism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: block reason</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline: block duration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engagement features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engagement + Age</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data
*s indicate significant (p < 0.05) improvement over best baseline in column

<table>
<thead>
<tr>
<th></th>
<th>Departure</th>
<th>Recidivism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: block reason</td>
<td>59.0</td>
<td></td>
</tr>
<tr>
<td>Baseline: block duration</td>
<td>56.7</td>
<td></td>
</tr>
<tr>
<td>Community age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engagement features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engagement + Age</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data. *s indicate significant (p < 0.05) improvement over best baseline in column.

<table>
<thead>
<tr>
<th></th>
<th>Departure</th>
<th>Recidivism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: block reason</td>
<td>59.0</td>
<td></td>
</tr>
<tr>
<td>Baseline: block duration</td>
<td>56.7</td>
<td></td>
</tr>
<tr>
<td>Community age</td>
<td>58.6</td>
<td></td>
</tr>
<tr>
<td>Engagement features</td>
<td>61.4*</td>
<td></td>
</tr>
<tr>
<td>Engagement + Age</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Engagement features: predicting trajectories**

All accuracies are computed via leave-one-out CV on balanced (paired) data; *'s indicate significant ($p < 0.05$) improvement over best baseline in column.

<table>
<thead>
<tr>
<th></th>
<th>Departure</th>
<th>Recidivism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: block reason</td>
<td>59.0</td>
<td></td>
</tr>
<tr>
<td>Baseline: block duration</td>
<td>56.7</td>
<td></td>
</tr>
<tr>
<td>Community age</td>
<td>58.6</td>
<td></td>
</tr>
<tr>
<td>Engagement features</td>
<td>61.4*</td>
<td></td>
</tr>
<tr>
<td>Engagement + Age</td>
<td>66.2*</td>
<td></td>
</tr>
</tbody>
</table>
### Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data. *

*s indicate significant ($p < 0.05$) improvement over best baseline in column.

<table>
<thead>
<tr>
<th></th>
<th>Departure</th>
<th>Recidivism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: block reason</td>
<td>59.0</td>
<td>51.9</td>
</tr>
<tr>
<td>Baseline: block duration</td>
<td>56.7</td>
<td>43.8</td>
</tr>
<tr>
<td>Community age</td>
<td>58.6</td>
<td></td>
</tr>
<tr>
<td>Engagement features</td>
<td>61.4*</td>
<td></td>
</tr>
<tr>
<td>Engagement + Age</td>
<td>66.2*</td>
<td></td>
</tr>
</tbody>
</table>
Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data, *s indicate significant (p < 0.05) improvement over best baseline in column.

<table>
<thead>
<tr>
<th></th>
<th>Departure</th>
<th>Recidivism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: block reason</td>
<td>59.0</td>
<td>51.9</td>
</tr>
<tr>
<td>Baseline: block duration</td>
<td>56.7</td>
<td>43.8</td>
</tr>
<tr>
<td>Community age</td>
<td>58.6</td>
<td>56.3*</td>
</tr>
<tr>
<td>Engagement features</td>
<td>61.4*</td>
<td>59.1*</td>
</tr>
<tr>
<td>Engagement + Age</td>
<td>66.2*</td>
<td>58.8*</td>
</tr>
</tbody>
</table>
Can we tell which path will be taken?

User characteristics?
Account age, amount of interaction, etc.
Ribeiro et al. (2018)
Cheng et al. (2017)
D-N-M et al. (2013)
Halfaker et al. (2011) and more...

Mod action context?
How severe was the moderator's action? How does the user react?
Corbett-Davies et al. (2017)
Tonry (2008)
Makkai & Braithwaite (1994)
Grasmik & Bryjak (1980) and more...
Block context: motivation

How do properties of first block affect likelihood of another block?
Block context: motivation

How do properties of first block affect likelihood of another block?

Unique to recidivism (as opposed to getting blocked in general)
Block context: motivation

How do properties of first block affect likelihood of another block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance
Block context: motivation

How do properties of first block affect likelihood of another block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

    Likelihood of repeat offense depends on...
Block context: motivation

How do properties of first block affect likelihood of another block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

  Likelihood of repeat offense depends on...

    ...severity of the punishment

Grasmick and Bryjack (1980); Klepper and Nagin (1989)
How do properties of first block affect likelihood of another block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

- Likelihood of repeat offense depends on...
  - severity of the punishment
  - offender’s perception of punishment as fair

Grasmick and Bryjack (1980); Klepper and Nagin (1989)
Makkai and Braithwaite (1994); Paternoster et al. (1997); Williams (2005); Tonry (2008)
Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Unique to recidivism (as opposed to getting blocked in general)

*Loosely* motivated by studies in offline law compliance

2 views of offline recidivism:

- Likelihood of repeat offense depends on...
  - severity of the punishment
  - offender’s perception of punishment as fair
Block context: measuring perceived fairness

2 angles: blocked user’s perspective and admin’s perspective

What can blocked users do to signal that they view block as (un)fair?

What can admins do to signal to blocked users that rules are fair?
Block context: measuring perceived fairness

2 angles: blocked user’s perspective and admin’s perspective

What can blocked users do to signal that they view block as (un)fair?
  ● Talk page comments

What can admins do to signal to blocked users that rules are fair?
  ● Unblocks
Admin’s perspective: Unblocks

- Recidivist: 45.4%
- Reformed: 55.6%

Not unblocked (4,696 total)
Admin’s perspective: Unblocks

<table>
<thead>
<tr>
<th></th>
<th>Unblocked (498 total)</th>
<th>Not unblocked (4,696 total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recidivist</td>
<td>37.6%</td>
<td>45.4%</td>
</tr>
<tr>
<td>Reformed</td>
<td>62.4%</td>
<td>55.6%</td>
</tr>
</tbody>
</table>
User’s perspective: Perceived fairness

3 linguistic indicators of user’s perception of fairness in comments:

Apologizing (suggests user acknowledges fairness of block)
- e.g., “I am deeply sorry for not understanding the whole situation, and ask for your forgiveness.”

Direct questioning (hostile; suggests user is fighting back)
- e.g., “So what policy, precisely have I violated?”

Explicit mentions of “unfairness” and related phrases
- e.g., “I have alerted another administrator about your blatant [sic] and unwarranted abuse of power”
User’s perspective: Perceived fairness

Likelihood of recidivism is lower for users who apologize

- 37.0% Recidivist (Apologizes, 165 total)
- 45.9% Recidivist (Does not apologize, 2,029 total)
- 63.0% Reformed (Apologizes, 165 total)
- 54.1% Reformed (Does not apologize, 2,029 total)
User’s perspective: Perceived fairness

Likelihood of recidivism is higher for users who use unfairness phrases or direct questioning.
How do properties of first block affect likelihood of another block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

Likelihood of repeat offense depends on...

- ...severity of the punishment
- ...offender’s perception of punishment as fair
Block context: motivation

How do properties of first block affect likelihood of another block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

- Likelihood of repeat offense depends on...
  - severity of the punishment
  - offender's perception of punishment as fair
Concluding Thoughts
Limitations / Future work

What **not** to do: deploy this classifier in production setting

- Not trained for a realistic setting!
- Even with really “good” classifier, would require careful analysis of potential biases and other risks
Limitations / Future work

What **not** to do: deploy this classifier in production setting
- Not trained for a realistic setting!
- Even with really “good” classifier, would require careful analysis of potential biases and other risks

Drawbacks of setup
- “Unfairness lexicon” is incomplete, crude measure
- Departure is not binary!
- Lack of another block doesn’t necessarily mean lack of reoffense
Limitations / Future work

What **not** to do: deploy this classifier in production setting
- Not trained for a realistic setting!
- Even with really “good” classifier, would require careful analysis of potential biases and other risks

Drawbacks of setup
- “Unfairness lexicon” is incomplete, crude measure
- Departure is not binary!
- Lack of another block doesn’t necessarily mean lack of reoffense
Blocks have consequences!
Blocks have consequences! Moderators should pay close attention to how their actions might be perceived.
Questions?

Code and data available via ConvoKit (convokit.cornell.edu)