Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion

Sanmay Das*, Allen Lavoie*, Malik Magdon-Ismail**

*Virginia Tech→Washington University in St. Louis, **Rensselaer Polytechnic Institute
Why administrators?

- Rule enforcement
  - Block, protect, delete
Why administrators?

- Rule enforcement
  - Block, protect, delete
- Conflict mediation
  - “Uninvolved administrator”
  - Informal social role

Barack Obama

From Wikipedia, the free encyclopedia
(Redirected from Obama)

This article is semi-protected to promote compliance with the policy on biographies of living people.
Plans for manipulation

for now I don’t touch this any more until we have maybe 20 editors who can fight and two “uninvolved” admins on our side . . .

-Email to a CAMERA Wikipedia discussion group

- Incentive for groups to abuse administrator status
  - CAMERA emails: plans to install an “uninvolved” administrator
Plans for manipulation

for now I don’t touch this any more until we have maybe 20 editors who can fight and two “uninvolved” admins on our side ...

-Email to a CAMERA Wikipedia discussion group

- Incentive for groups to abuse administrator status
  - CAMERA emails: plans to install an “uninvolved” administrator

- Suspicious administrator behavior changes on Wikipedia?
Plans for manipulation

for now I don’t touch this any more until we have maybe 20 editors who can fight and two “uninvolved” admins on our side . . .

-Email to a CAMERA Wikipedia discussion group

- Incentive for groups to abuse administrator status
  - CAMERA emails: plans to install an “uninvolved” administrator

- Suspicious administrator behavior changes on Wikipedia?
  - Yes
Outline

- Motivation
- Methods for quantifying behavior
- Validation
- Administrator behavior changes on Wikipedia
How do we define focused controversial editing?
How do we define focused controversial editing?

- Focus on user’s local edit graph
Controversy score

Problem: administrators expected to deal with controversy (see Welser et al., 2011; Burke and Kraut, 2008)

Allen Lavoie (Virginia Tech→Wash. U.)  Manipulation among the arbiters of collective intelligence
Controversy score

- Science
- Global Warming
- Glacier
- Carbon Tax
- Methane
- Politics
- Agriculture
- U.S. Budget

- Edit-proportion-weighted controversy (user-level)

\[ \text{CScore} = \sum_{k} p_k c_k \]
Controversy score

- Edit-proportion-weighted controversy (user-level)

\[
\text{CScore} = \sum_k p_k c_k
\]

- Problem: administrators expected to deal with controversy (see Welser et al., 2011; Burke and Kraut, 2008)
Clustering score

\[
clust(k) = \sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j w_{ki} w_{kj} w_{ij}
\]
Clustering score

- Measure topical focus

User-level clustering score

\[ \text{Clust}(k) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j w_{ki} w_{kj} w_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j w_{ki} w_{kj}} \]
Clustering score

- Measure topical focus
- Generalization of clustering coefficient (Kalna and Higham, 2007)

\[
\text{clust}(k) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j w_{ki} w_{kj} w_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j w_{ki} w_{kj}}
\]
Clustering score

- Measure topical focus
- Generalization of clustering coefficient (Kalna and Higham, 2007)

\[
\text{clust}(k) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j w_{ki} w_{kj} w_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j w_{ki} w_{kj}}
\]

- User-level clustering score \( \sum_k p_k \text{clust}(k) \)
Clustered controversy

Define page impact $\nu(k) = p_k c_k$

$$\text{clust}(k) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \nu(i) \nu(j) w_{ki} w_{kj} w_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} \nu(i) \nu(j) w_{ki} w_{kj}}$$
Clustered controversy

Define page impact $\iota(k) = p_k c_k$

$$\text{clust}(k) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \iota(i) \iota(j) w_{ki} w_{kj} w_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} \iota(i) \iota(j) w_{ki} w_{kj}}$$

User-level CCSScore $\sum_k \iota(k) \text{clust}(k)$
Page controversy

- Predict number of revisions with `{{controversial}}` tag (Kittur et al, 2007)
  - Features: reverts, page edits, talk edits, anonymous edits, etc.
Natural language topic model (LDA), compare topic distributions
Natural language topic model (LDA), compare topic distributions

- Wikipedia has other good options (e.g. Li et al, 2011)
Outline

- Motivation
- Methods for quantifying behavior
- Validation
- Administrator behavior changes on Wikipedia
High and low CC scores visualized
Validation: manipulative users

- Can the scores find manipulative users?
Validation: manipulative users

- Can the scores find manipulative users?
- Dataset of active users blocked for manipulative behavior
  - Sock puppetry
  - Biographies of living persons
  - Edit warring/three revert rule
Validation: manipulative users

- Can the scores find manipulative users?
- Dataset of active users blocked for manipulative behavior
  - Sock puppetry
  - Biographies of living persons
  - Edit warring/three revert rule
- Compare with never-blocked users
Identifying manipulative users

Block ROC

- CC (AUC=0.68)
- Controversy (AUC=0.67)
- Clustering (AUC=0.54)
- Revert fraction (AUC=0.55)
Outline

- Motivation
- Methods for quantifying behavior
- Validation
- Administrator behavior changes on Wikipedia
Administrator elections (RfA)

- Nomination, followed by open “voting”
  - Support, oppose, or neutral with discussion
  - Have vote dataset
- ~70% support to become an administrator
Administrator elections (RfA)

- Nomination, followed by open “voting”
  - Support, oppose, or neutral with discussion
  - Have vote dataset
- ~70% support to become an administrator
- Interested in behavior changes after a successful RfA
## Suspicious behavior change anecdotes

<table>
<thead>
<tr>
<th>Before RfA</th>
<th>Admin 1</th>
<th>After RfA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>cc%</td>
<td>Article</td>
</tr>
<tr>
<td>Search engine optimization</td>
<td>48.7%</td>
<td>Homeopathy</td>
</tr>
<tr>
<td>Web 2.0</td>
<td>14.7%</td>
<td>Waterboarding</td>
</tr>
<tr>
<td>Kiev</td>
<td>12.3%</td>
<td>World Trade Center controlled demolition conspiracy theories</td>
</tr>
<tr>
<td>Zango (company)</td>
<td>2.5%</td>
<td>spacy theories</td>
</tr>
<tr>
<td>Wi-Fi</td>
<td>2.1%</td>
<td>Electronic voice phenomenon</td>
</tr>
<tr>
<td>Vanessa Fox</td>
<td>2.1%</td>
<td>Web 2.0</td>
</tr>
<tr>
<td>Scientology</td>
<td>1.6%</td>
<td>SS Edmund Fitzgerald</td>
</tr>
<tr>
<td>Gamma-ray burst</td>
<td>0.8%</td>
<td>Collapse of the World Trade Center</td>
</tr>
<tr>
<td>Search engine submission</td>
<td>0.8%</td>
<td>Naked short selling</td>
</tr>
<tr>
<td>Animal testing</td>
<td>0.8%</td>
<td>Joe Lieberman</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Before RfA</th>
<th>Admin 2</th>
<th>After RfA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>cc%</td>
<td>Article</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>10.9%</td>
<td>Abortion</td>
</tr>
<tr>
<td>Boolean algebra (structure)</td>
<td>9.3%</td>
<td>Support for the legalization of abortion</td>
</tr>
<tr>
<td>The Beatles</td>
<td>5.5%</td>
<td>Safe sex</td>
</tr>
<tr>
<td>Association football</td>
<td>3.3%</td>
<td>Condom</td>
</tr>
<tr>
<td>Philosophy</td>
<td>3.0%</td>
<td>Hippie</td>
</tr>
<tr>
<td>Irony</td>
<td>2.7%</td>
<td>Fox News Channel</td>
</tr>
<tr>
<td>Lysergic acid diethylamide</td>
<td>1.9%</td>
<td>Planned Parenthood</td>
</tr>
<tr>
<td>Hippie</td>
<td>1.3%</td>
<td>The Beatles</td>
</tr>
<tr>
<td>Bill O'Reilly (political commentator)</td>
<td>1.3%</td>
<td>Masturbation</td>
</tr>
<tr>
<td>Iraq War</td>
<td>1.2%</td>
<td>Lysergic acid diethylamide</td>
</tr>
</tbody>
</table>
Administrator behavior changes

- 76% more large behavior changes than expected
Administrator behavior changes

- 76% more large behavior changes than expected
  - Prior behavior indistinguishable from unsuccessful candidates
Administrator behavior changes

- 76% more large behavior changes than expected
  - Prior behavior indistinguishable from unsuccessful candidates
  - Both successful and unsuccessful have higher CC to start
76% more large behavior changes than expected
  ▶ Prior behavior indistinguishable from unsuccessful candidates
  ▶ Both successful and unsuccessful have higher CC to start

Most increase controversy, only some on a specific topic
Selection vs. causality

- Users change behavior because they become administrators
  - Matched sample of “almost successful” and “almost unsuccessful”
Predicting behavior changes

- Can we find behavior changes before they happen?
- Lots of data available
  - Prior behavior
  - Voting
Scoring candidates

- Percentage supporting candidate
Scoring candidates

- Percentage supporting candidate
- Predicted probability of success based on prior history (Burke and Kraut, 2008)
Scoring candidates

- Percentage supporting candidate
- Predicted probability of success based on prior history (Burke and Kraut, 2008)
- Weighted voter score (Ghosh et al, 2011)
Behavior changes by score

- Uncontroversial candidates change behavior at least as much...
Behavior changes by score

- Relevant information exists in the vote
  - Helps to consider who is voting
Summary and conclusions

- Some administrators change behavior in suspicious ways
  - Detectable automatically
Summary and conclusions

- Some administrators change behavior in suspicious ways
  - Detectable automatically
- Voting can predict behavior changes
  - Current system does not