Mopping Up: Modeling Wikipedia Promotion Decisions

Moira Burke and Robert Kraut
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
{moira, robert.kraut}@cmu.edu

ABSTRACT
This paper presents a model of the behavior of candidates for promotion to administrator status in Wikipedia. It uses a policy capture framework to highlight similarities and differences in the community's stated criteria for promotion decisions to those criteria actually correlated with promotion success. As promotions are determined by the consensus of dozens of voters with conflicting opinions and unwritten expectations, the results highlight the degree to which consensus is truly reached. The model is fast and easily computable on the fly, and thus could be applied as a self-evaluation tool for editors considering becoming administrators, as a dashboard for voters to view a nominee's relevant statistics, or as a tool to automatically search for likely future administrators. Implications for distributed consensus-building in online communities are discussed.

Author Keywords
Wikipedia, administrators, management, collaboration, policy capture, promotion, organizational behavior

ACM Classification Keywords
H.5.3 [Information Interfaces]: Group and Organization Interfaces - Collaborative computing, Web-based interaction, Computer-supported cooperative work

INTRODUCTION
Wikipedia, the collaboratively edited online encyclopedia, had over 6.8 million registered users contributing to 2.3 million articles in the English version alone as of April 2008. Although Wikipedia is written and edited by volunteers and is not supervised by a professional staff, evidence suggests that the quality in Wikipedia is comparable to that of the Encyclopedia Britannica [8].

In the midst of exponential growth of both content and users [13], *administrators* help maintain this quality: they delete copyright violations, protect frequently vandalized pages, block malicious users, move pages when there are name conflicts, exclude bulk vandalism from the recent changes list, and edit the front page. Approximately 1500 editors have successfully passed the rigorous peer review associated with Wikipedia's Request for Adminship (RfA) process and been given administrator privileges. They are considered trusted custodians of the successful encyclopedia and its community of contributors.

Although CSCW scholars have examined many aspects of this highly successful decentralized environment, including conflict and coordination [14,20,21]; regulation, policymaking, and consensus-building [2,7]; and the transition of novice readers to committed community members [1], we know very little about how the community is managed and how it makes decisions. In particular, though Wikipedia members have explicit criteria they look for in candidates for promotion to administrator status, we do not know whether the dozens of individuals discussing a candidate's promotion actually use those criteria. Previous policy capture research has often found disconnects between the factors people cite for making decisions and the factors actually used in those decisions [19,23]. The promotion decision in Wikipedia has a number of characteristics ideal for policy capture research: the role of the Wikipedia administrator is fairly well defined, there are many judges involved in the promotion decision, and there is high transparency in the records of past action since all edits are recorded.

This paper presents a model that predicts who will be promoted to administrator status in Wikipedia. The model can be used to identify editors likely to be promoted, as a self-evaluation tool for potential admins, and as a dashboard of relevant behavior for RfA voters[1]. The model is lightweight, based on edit counts and brief edit summary text available in the public Wikipedia database or the user's contribution page, and does not require any full text

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CSCW'08, November 8–12, 2008, San Diego, California, USA.
Copyright 2008 ACM 978-1-60558-007-4/08/11...$5.00.

[1] Note that Wikipedians do not consider the RfA process a strict "vote," but rather a consensus-building activity. Participants in the RfA discussion are referred to as "voters" in this paper for simplicity.

analysis of articles or talk pages. Thus, it can be run quickly to search the entire population of editors, or allow editors to calculate their own likelihood of success on the fly without taxing the server. It also identifies areas in which more in-depth methods—such as conversation analysis of Arbitration Committee pages—should be applied.

Most theories of human-motivation, such as the classic expectancy-value model, hold that people will work hard if they think that doing so will lead to outcomes they value [22] and emphasize the importance of feedback to help people achieve their performance goals [15]. Providing editors with feedback about the strengths and weaknesses in their edit histories may prompt those desiring to become admins to behave in ways valued by the community or may prompt editors who had not considered applying for administrator status to step up. Wikipedia implemented a simple version of this idea, maintaining a list of users with high edit counts[2]. However, sheer number of edits does not make an RfA nominee likely to pass; "editcountitis" among the voters is frowned upon because some users rack up high counts by making thousands of minor typographical changes, some with automated tools.

In addition to the practical benefits to the Wikipedia community, this paper also speaks to long-standing concerns of organizational scholars who have asked what causes employees to get ahead in their jobs—for example, their experiences and skills, social networks, or job-irrelevant attributes like gender or attractiveness (see [17] for review). Despite protestations in Wikipedia that admins are lowly janitors "mopping up," in many ways election to administrator is a promotion, distinguishing an elite core group from the larger mass of editors. The research described here uses policy capture [19] to compare the attributes Wikipedia guidelines state are important in selecting administrators to those that are actually associated with promotion. The behavioral data here is rarely available in conventional organizational settings.

PROMOTION TO ADMINISTRATOR STATUS

The RfA Processes and Stated Promotion Criteria

To become an administrator, an editor must undergo a week of scrutiny during which the community builds consensus about the candidate's experience and trustworthiness. Administrator tools are not granted lightly. An inexperienced, biased, or ill-intentioned administrator could cause significant damage, reducing the encyclopedia's credibility or demotivating other editors. The damage comes less from the new tools that admins gain, since most damage could be reversed by other admins, but instead from administrators' privileged social status and capacity to represent Wikipedia to external audiences. Yet many inexperienced editors seek to "level up" to admin status within a few months of joining Wikipedia. In the Guide to RfAs[3], the community describes criteria many RfA evaluators look for in nominees, including:

- **Strong edit history** with plenty of material contributions to Wikipedia articles.
- **Varied experience**. RfAs where an editor has mainly contributed in one way (little editing of articles, or little or no participation in [Articles for Deletion], or little or no participation in discussions about Wikipedia policies and processes, for example) have tended to be more controversial than those where the editor's contributions have been wider.
- **User interaction**. Evidence of you talking to other users, on article talk or user talk pages. These interactions need to be helpful and polite.
- **Trustworthiness**. General reliability as evidence that you would use administrator rights carefully to avoid irreversible damage, especially in the stressful situations that can arise more frequently for administrators.
- **Helping with chores**. Evidence that you are already engaging in administrator-like work and debates such as RC Patrol and articles for deletion.
- **High quality of articles**. A good way to demonstrate this is contributing to getting articles featured, although good articles are also well-regarded.
- **Observing consensus**. A track record of working within policy, showing an understanding of consensus.
- **Edit summaries**. Constructive and frequent use of edit summaries is a quality some RfA contributors want to see. Some expect use of edit summaries to approach 100% of the time.

However, the Guide points out that community members have differing requirements and unwritten expectations. This paper examines which of these criteria are most predictive of success, and which are only nominally used by the community in choosing its administrators.

The Request for Administration Process

The RfA process consists of three parts: an introductory nomination statement, the nominee's answers to questions about past and future behavior, and statements of support, opposition, or neutrality by community members. Any

[2] http://en.wikipedia.org/wiki/Wikipedia:NA

[3] http://en.wikipedia.org/wiki/Wikipedia:GRFA

registered Wikipedia member can voice an opinion, but the RfA is not a strict vote: at the end of a week, a bureaucrat—one of approximately 25 editors with privileges greater than administrators—reads through the opinions and decides whether consensus was reached. Candidates with more than 75% support are generally successful, though bureaucrats weigh voters' reasons, not just their votes. Votes by suspected sockpuppets (multiple identities held by the same person) or meatpuppets (new users recruited by a voter to back up the voter's opinion) are discounted. Nominees may withdraw at any time, and the "snowball clause" allows any editor to close a nomination early to avoid wasting the community's time if the nominee is so inexperienced as to not stand a "snowball's chance in hell" of passing.

RfA candidates answer three sets of standard questions as part of their nomination: (1) What chores do you intend to help with? (2) What are your best contributions, and why? and (3) Have you been involved in conflicts over editing or have other users caused you stress? Nominees answer these and ad hoc questions posed by other community members, citing records of past events to demonstrate their competence and handling of controversy.

RfA evaluators typically many look for an answer to the first question that demonstrates the candidate is already behaving like an administrator, helping with chores such as monitoring recent changes, welcoming newcomers, or participating in debates at Requests for Comment (RfC) or Articles for Deletion (AfD). Editors are encouraged to first help the wiki using their existing privileges, such as reverting vandalism, organizing collections of articles in a subject area (WikiProjects), and reducing the non-administrator backlog, before considering promotion to administrator. Editors are encouraged to seek administrator status only if they need the privileges that go with the role, for example, if their efforts to fight vandalism were hampered by frequently needing to wait for administrators to block users they identified.

The second RfA question demonstrates the significance of a candidate's contribution, and many successful candidates discuss their work on articles that reached high levels of quality, such as Featured Article status.

Answers to the third question highlight incidents in which the nominee has dealt with interpersonal conflict, a common occurrence in an encyclopedia edited by thousands of people with differing viewpoints yet striving for neutrality. Previous work has shown conflict on even seemingly neutral topics like chocolate [20], and that administrators often serve as both official and unofficial mediators in edit wars on controversial topics, such as euthanasia and evolution [14]. All members of the Arbitration Committee, a formal body for resolving conflict, are administrators. In answering the third question, the nominee demonstrates how he has dealt with controversy or uncivil comments from other editors, linking to evidence on talk pages. Candidates involved in heated "edit wars" are unlikely to be well received by the community, though candidates who make full disclosures of previous mistakes and have recent histories of good behavior are more likely to succeed.

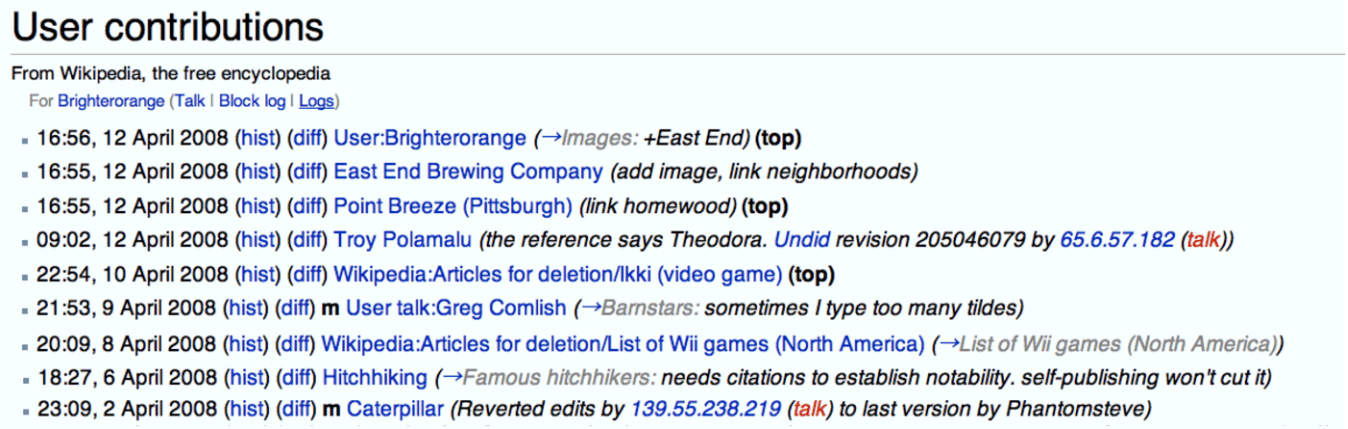
Approximately 2700 editors have been nominated for adminship since 2001 with an overall success rate of 53%. However, the process has gradually grown more rigorous, dropping from a 75.5% success rate through 2005 to 42% in 2006 and 2007, and some early administrators have expressed doubt that they would pass muster if their RfA were held today [7]. The process once called "no big deal" by the founder of Wikipedia has become a fairly big deal[4].

MODELING SUCCESSFUL ADMIN CANDIDATES

We use the technique known as policy capture [10] to compare the criteria the community states it uses in making promotion decisions and the behavioral data that are actually correlated with promotion. We examined all 1551 Requests for Adminship from January 2006 through October 2007, with 49 RfAs removed for being multiple attempts by the same candidate in one month (all of which failed), bots, sockpuppets, or because the nominee's edit history prior to the RfA was not available.

Stated criteria come from the categories in the Guide to RfAs listed above. Behavioral data come from the user's contribution history. For each RfA, the nominee's contribution history page up to the month before the RfA was parsed, counted, and grouped according to the categories described in the Guide to RfAs. We used an informal analysis of the discussions of the RfA evaluators to determine which data fell into which category, as well as the descriptions of behavior in the Guide itself. Contribution histories include date and time stamps, the namespace (indicating the type of page, such as an article, policy, or discussion), a link to the page itself, and an optional free-text summary left by the user. The same features are also available in the public database download. Figure 1 shows a sample contribution history page and Table 1 provides summary statistics. Features applicable to multiple categories were placed in a single category as described below, and two categories from the Guide to RfAs—trustworthiness and high quality of articles—were excluded from this analysis because they could not be captured from simple edit counts. These two categories are considered further in the discussion section.

[4] http://en.wikipedia.org/wiki/WP:DEAL


User contributions

From Wikipedia, the free encyclopedia

For Brighterorange (Talk | Block log | Logs)

- 16:56, 12 April 2008 (hist) (diff) User:Brighterorange *(→Images: +East End)* **(top)**
- 16:55, 12 April 2008 (hist) (diff) East End Brewing Company *(add image, link neighborhoods)*
- 16:55, 12 April 2008 (hist) (diff) Point Breeze (Pittsburgh) *(link homewood)* **(top)**
- 09:02, 12 April 2008 (hist) (diff) Troy Polamalu *(the reference says Theodora. Undid revision 205046079 by 65.6.57.182 (talk))*
- 22:54, 10 April 2008 (hist) (diff) Wikipedia:Articles for deletion/Ikki (video game) **(top)**
- 21:53, 9 April 2008 (hist) (diff) **m** User talk:Greg Comlish *(→Barnstars: sometimes I type too many tildes)*
- 20:09, 8 April 2008 (hist) (diff) Wikipedia:Articles for deletion/List of Wii games (North America) *(→List of Wii games (North America))*
- 18:27, 6 April 2008 (hist) (diff) Hitchhiking *(→Famous hitchhikers: needs citations to establish notability. self-publishing won't cut it)*
- 23:09, 2 April 2008 (hist) (diff) **m** Caterpillar *(Reverted edits by 139.55.238.219 (talk) to last version by Phantomsteve)*

Figure 1. A sample contribution history page.

We acknowledge that the simple metrics described below do not capture the full scope of RfA voters' criteria. For example, when voters refer to a candidate's interaction with others, they are referring to many subtle qualities such as the tone of the candidate's posts in discussion venues, ability to defuse conflict when challenged, or even the variety of other users she interacts with. Instead, we use easy-to-measure proxies, such counts of edits to talk pages. We did so in part because previous research suggests that people overestimate the subtlety and sophistication of the decision rules they use to assess other people [16,5]. We also did so to ensure that the models could run quickly enough to be the basis of a practical decision aid. Despite its simplicity, the model we built is moderately accurate. We discuss improvements to the model and ways that it can augment more high-level human consensus-building in the discussion section.

Behavioral Data

Strong edit history

These metrics includes the number of edits the nominee made to articles and the number of months between the nominee's first edit and the RfA. Total edit count (Mean=5010.5, Std. Dev=5818.8) is included in the baseline analysis but is replaced by counts of edits in the individual namespaces (e.g. articles, article talk pages, user talk pages) in the final model to avoid multicollinearity.

Varied experience

This measures the breadth of edits across the community. Editors distribute their work across individual namespaces, including the Wikipedia namespace (pages in a subsection of the encyclopedia focusing mainly on policy, with wikiproject edits counted separately), and User pages.

The *breadth* score, a proxy for diverse experience, is the number of different areas in which the user has participated, from the set {article, article talk, Wikipedia, Wikipedia talk, user, user talk, articles/categories/ templates for deletion (XfD), (un)deletion review, other RfAs, village pump, admin intervention against vandalism (AIV), requests for protection (RfP), administrators' noticeboard, arbitration committee, mediation committee, and wikiprojects}. Thus, a user who has edited articles, edited her own user page, and posted once at the Village Pump would have a breadth score of 3. The number of edits in each of these areas is accounted for in the following categories to determine their relative importance.

User interaction

This includes edits to all talk pages, participation on arbitration or mediation committee pages or wikiquette alerts (an early stage in dispute resolution), posting "welcome" on user talk pages, and including common variants of "please" (including "pls" and "plz") or "thanks" (including "thx") in summary text. More complicated ways of measuring user interaction, such as analysis of social networks or talk page text, although outside the scope of our research, could improve the model.

Helping with chores

This measures a demonstrable "need for the tools" and includes reversion of vandalism (noted by "revert" or "rv" in the summary) requesting administrator intervention for specific vandals ("AIV"), requesting protection for a frequently vandalized page, neutrality fixes (noted by "pov" or "npov" for "neutral point of view"), requesting administrator attention (e.g. for inappropriate usernames), and participating in deletion discussions, including articles/categories/templates for deletion ("XfD") and (un)deletion reviews. It also includes the percent of the user's total edits marked as minor (designated with an "m" in the contribution history page), which is used for spelling or small formatting changes.

Observing consensus

This includes participation in other editors' RfAs, posting to the Village Pump (a forum for technical and policy discussions), or discussing articles to be deleted or rewritten. Though there are better ways to measure consensus building, they involve natural language processing and thus may be prohibitively time-consuming at Wikipedia scale. Thus, these simple behavioral measures are lightweight proxies for consensus.

Edit summaries

When making an edit, users have the option to include a brief summary. Summaries are both descriptions of changes and conversations between authors, often preempting objections or asking questions of each other [21]. Wikipedia also automatically generates some summary text, as when new sub-sections are created. This metric includes the percent of edits with a human-written summary (automatically generated summaries are not included), and the average length of the human-written summaries.

Results

To examine the impact of these behavioral factors on the likelihood of a candidate's promotion to administrator, we performed a probit regression on the binary dependent variable, RfA success. All variables were standardized, so the coefficients in Table 1 represent the change in probability of success when a continuous variable (such as the number of article edits) is increased by one standard deviation. Because many of the variables have long tails, with standard deviations greater than their means, we performed a similar analysis after first conducting a log transformation on the independent variables. The results were qualitatively similar, and we report results in terms of original units for ease of interpretation. To check for inflated standard errors due to multicollinearity between variables, we calculated variance inflation factors (VIFs). All VIFs are well below 10, indicating low collinearity between factors [11]. Multiple attempts by the same candidate in a single month were excluded, leaving only one attempt per month, and the candidate's number of previous RfA attempts (in other months) is included as a control variable; each subsequent attempt has an 11.8% lower chance of success than the previous one.

Table 1 presents two models: Model 1 contains category-level scales created by summing specific behaviors, and Model 2 uses the behaviors themselves as predictors. For Model 1, the inter-item correlations were calculated using Cronbach's alpha for variables within each category and are recorded in Table 1. Note that the strong edit history and edit summaries categories each had only two variables, too few to create reliable scales. RfA voters often explicitly mention the two variables comprising the strong edit history scale together—number of edits and length of time in Wikipedia—for example, *"Sorry, too new. Keep up the good work and try resubmitting when you get over 1,000 edits"* and *"Just over 200 edits and a membership time of not even two weeks is not nearly enough."* For the edit summary scale, it is not surprising that percentage of edits summarized and length of summaries are not highly correlated; editors doing minor edits may leave short summaries nearly every time. Thus, to address the issue of low reliability for two of the scales, Model 2 presents the data in terms of individual behavioral data.

Baseline performance would be 58.0% for a model that always predicted failure of RfA attempts. The category-level model 1 performs at 72.6% accuracy, while the behavioral-level model 2 improves the fit to 75.6%.

The results highlight both similarities and differences between the community's stated promotion criteria and those that actually correlate with RfA success. Criteria from the Guide to RfAs are discussed by category below, but in general, we find that, in line with the community's ideals, edit history and breadth of experience are good predictors of promotion, but contrary to these ideals, demonstrating a need for the tools by helping with chores, posting to the administrators' noticeboard, or taking conflict to appropriate venues for resolution does not help, and may even hurt a candidate's chances of promotion.

Strong edit history

In the RfA process, there is a tension between promoting editors with extensive experience and rewarding "editcountitis," and this is evident in the results. Merely adding the number of prior edits to the baseline model improved it from 58.0% to 66.2% accuracy. Successfully promoted editors had roughly twice as many edits (Mean=3037.7) as those who were not promoted (Mean=1604.1) ($p<.001$). However, it take several thousand more edits to increase one's likelihood of success Every additional 3804 edits increased chances of promotion by approximately 10%. Length of tenure in Wikipedia helped slightly; every eight additional months of experience increased a nominee's likelihood of promotion by approximately 3%.

Varied experience

Also consistent with the community's stated ideal of promoting editors with diverse experience, the breadth score is a strong predictor of promotion. Most nominees had made edits to many (Mean=11.0) different regions of the encyclopedia (e.g. articles, user talk pages, deletion discussions, and the village pump forum), and every additional 3.7 regions nominees participated in increased their chances of promotion by approximately 5%. However, though participation in policy discussions was in the right direction, the effect was not significant ($p=0.16$), and WikiProject participation did not increase the likelihood of promotion either.

			Change in probability of promotion	
	Mean	Standard Deviation	Model 1: Category scales	Model 2: Behavioral factors
Attempt number	1.2	0.6	-7.1% ***	-7.1%***
Strong edit history (α = 0.0)			11.4% ***	
Article edits	2611.1	3804.3		10.1%***
Months since first edit	9.6	8.0		2.9%*
Varied experience (α = 0.40)			14.8% ***	
Breadth score	11.0	3.7		4.6%*
Wikipedia policy edits	474.4	755.9		5.2%
WikiProject edits	144.0	569.1		-1.5%
User interaction (α = 0.74)			7.8% *	
Article talk edits	415.2	775.4		5.0%**
User talk edits	786.6	1169.9		0.9%
User edits	219.0	296.7		-1.5%
Wikipedia talk edits	87.6	179.5		1.6%
Arbitration /mediation/wikiquette edits	9.8	47.1		-6.0%***
Newcomer welcomes	76.9	321.1		-2.2%
"Please" in summary	31.7	83.8		0.7%
"Thanks" in summary	21.8	39.3		7.7%***
Helping with chores (α = 0.59)			3.6%	
"Revert" in summary	257.6	563.2		2.2%
Vandal fighting (AIV)	26.5	108.7		-3.0%+
Requests for protection	3.7	12.2		-1.0%
"(N)pov" in summary	26.7	46.9		2.5%
Administrator attention/noticeboard	18.7	57.9		-3.3%+
Minor edits (%)	25.5%	23.0%		3.0%*
Observing consensus (α = 0.43)			3.3%	
X for deletion/review	252.2	513.6		2.0%
Other RfAs	41.7	99.1		-1.5%
Village pump	9.5	34.6		-0.6%
Edit summaries (α = 0.02)			25.1% ***	
Summarized (%)	80%	20%		25.3%***
Avg. summary length (log2chars)	5.0	0.8		0.9%

*** p < .001 ** p < .01 * p < .05 + p < 0.1

Table 1. Descriptive statistics and probit regression on the likelihood of promotion to administrator status in Wikipedia during 2006-7. Model 1 presents predictive variables at the category level. Model 2 presents the individual behavioral factors.

All variables have been standardized, so the rightmost column indicates the change in probability of success when a continuous variable (such as # of article edits) is increased by one standard deviation. All variables are edit counts unless otherwise noted.

User interaction

Some forms of user interaction have significant positive impact on a candidate's likelihood of promotion, while others are surprisingly harmful. Article talk pages are mechanisms for coordination and dispute resolution [21], so it is not surprising that future administrators participate more heavily there than do unsuccessful nominees. Editing user pages did not improve the model, perhaps because the norm in Wikipedia is to edit only one's own user page except rarely to add rewards, known as barnstars, to others' pages.

Somewhat unexpectedly, user talk page edits do not affect likelihood of becoming an administrator. This is perhaps because the norm is to hold disagreements over content on article talk pages, moving to user talk pages when the disagreement covers multiple articles or a user's overall behavior. Thus, user talk edits are likely to be mixed, and may have higher interpersonal conflict. This is supported by the finding that posts to Arbitration or Mediation Committee pages, or to Wikiquette notices, all of which are venues for dispute resolution, decrease the likelihood of success. Though the Guide to RfAs indicates a desire to promote editors who handle conflict appropriately, the more a nominee has participated in the appropriate forums, the less likely he is to be promoted. Although the current model does not take into account the content of these nominee's posts to the dispute resolution committees, it suggests the need for more thorough analysis of the language and coalition formation at these venues.

Politeness helps modestly; though it was rare, every 39 additional edit summaries with the word "thanks" in them increases the likelihood of success by 7.7%. Saying "please" and welcoming newcomers had no effect.

Helping with chores

The category for helping with chores represents one of the greatest disconnects between the stated criteria and those that actually predict promotion. The Guide to RfA states that many voters are looking for editors already engaged in administrator-like activities and editors demonstrating a need for administrator tools for blocking vandals and deleting pages. However, participating in typical chores such as reverting vandalism, alerting administrators to pages needing protection, and removing biased language do not increase a candidate's likelihood of promotion. In fact, requesting intervention against persistent vandals or posting to the Administrators' Noticeboard tend to harm promotion ($p=.14$ and $p=.06$, respectively). Much like the effect of posts to the Arbitration Committee discussed above, editors who escalate problems with other users to formal venues like the Administrators' Noticeboard are less likely to be promoted.

One type of chore—making minor edits, such as repairing formatting or spelling errors—does, however, slightly improve the likelihood of promotion. This measure may be suspect because editors themselves describe their edits as minor.

Observing consensus

Contrary to our expectations, none of the metrics of consensus building were predictive of promotion. Participating in other users' RfAs, discussing policy at the Village Pump, and voting for content to be deleted or reviewed did not increase a candidate's likelihood of promotion. However, these are simplistic measures of consensus-building, and a more in-depth analysis of the language and participation quality is warranted.

Edit summaries

Finally, consistent with the stated ideals in the Guide to RfAs, editors who frequently summarize their edits and leave coordination notes for future editors are more likely to be promoted. As one editor on the Guide to RfA talk page notes:

> "*You should use edit summaries most of the time because it explains what you've done to (1) the Recent Changes patrollers, who will be saved some trouble if you explain why your edit is legitimate, and (2) past and future editors, who will find your edits in the page history. It shows that you're considerate of other people in the Wikipedia community.*"

DISCUSSION

The model identifies behavioral criteria correlated with successful promotion to administrator status in Wikipedia, and in particular, highlights both similarities and differences between the community's stated promotion criteria and those actually correlated with RfA success. Extensive and diverse experience in Wikipedia, as well as article-level coordination on talk pages and edit summaries are good predictors of promotion, in line with criteria on the Guide to RfA. However, despite the community's strong claim that administrator status is not a trophy[6] and is granted to editors demonstrating a willingness to help out with janitorial chores, editors who do help with chores are not more likely to be promoted. Recently Wikipedians have discussed giving administrator-level tools for rolling back vandalism to non-administrators[7], suggesting a possible shift in the role of administrator away from janitorial duties.

Furthermore, editors who elevate problems with vandals to appropriate forums such as the Administrators' Noticeboard or those who seek conflict resolution at the Arbitration or Mediation Committees are in fact less likely

[6] http://en.wikipedia.org/wiki/Wikipedia:ANOT#TROPHY

[7] http://en.wikipedia.org/wiki/Wikipedia:NAR

to be promoted. Certainly, some editors who post to these venues are inexperienced and may have missed opportunities to resolve disputes more informally and discreetly. Moreover, the model does not distinguish between the person bringing the complaint to the committee and the respondent. Participation in these conflict-resolution forums may require extra diplomacy or other rhetorical strategies. Informal review of the discussions on these pages, as well as nominees' answers to the RfA question about handling conflict indicate that while all candidates downplay past disputes, successful candidates tend to deflect personal attacks, reflect on their own behavior and that of others, use Wikipedia jargon, and cite relevant policies and evidence. As one successful administrator describes his experience (underlines indicate links to evidence):

> *"My editing has not been particularly contentious, and years of dealing with troublemakers on my own websites has given me a pretty thick skin, so I am not the kind of person to get in heated arguments. I know some people on RfA see conflict resolution as an important test before adminship, so how about the argument over repeated copyvios at Talk:Bloodsport (film) (where I am called a poopoo head (heh)) . . . The page was eventually vprotected after I asked for help."*

This candidate successfully diffuses conflict with humor and drops Wikipedia jargon ("copyvios" for copyright violations and "vprotected" for a form of protection from vandalism), and links to the record of the conflict. Unsuccessful nominees tend to use less diplomatic language, reflection, and evidence:

> *"Not many users have annoyed me, but {name omitted} can sometimes annoy me, by reverting my edits, but also to try and get me blocked for things I haven't done."*

Another unsuccessful nominee said:

> *"Of all my contributions and interactions with other editors, I have only had one editing conflict (as opposed to civilized debate). I started out logical and as calm as can be, but through the course of the debate, other editors began with uncivil behavior. I just brushed it off, but more incivility and personal jabs followed."*

Overall, the quantitative results of this model indicate the need for more qualitative analysis of conflict discussions.

Applications

This model can be applied in three ways: as a dashboard of relevant statistics for RfA voters, as a self-evaluation tool for editors considering becoming administrators, or as a tool that automatically searches all editors' histories and picks out likely future administrators.

As a self-evaluation tool or voter dashboard, this model would allow editors or voters to size up an editor compared to previous RfA nominees, indicating areas where the editor needs improvement, or highlighting the editor's varied experience. However, this introduces the potential for editors to game the system, racking up minor edits and saying "thanks" in every edit summary as a way to increase their relative "score." A dashboard could promote an extreme version of "editcountis." Yet, as the Guide to RfA states, "*The reality is that adminship is oriented to communal trust and confidence, not percentages and numbers."* The current model does not take quality of contribution or trustworthiness into account, both criteria that require more thorough human review of a candidate's history. A dashboard would not replace human discussion, it would simply augment it, and the model itself could be improved by observing how voters refer to and use the statistics. We are currently complementing this quantitative model by performing a qualitative analysis of the RfA discussions to determine which criteria voters cite, and how candidates successfully respond to critical questions.

The model can also be applied as an "AdminFinderBot"—a user account for a computer program that runs the model automatically— to search all editors' histories and identify those with behavior similar to editors promoted to administrator status. Kittur and colleagues found that while administrators once accounted for nearly 60% of editing activity, their influence has declined to approximately 10% due to an influx of new editors [13]. Yet administrators are working harder than ever: while their edits per month have steadily increased, backlogs of work requiring administrator privileges continue to grow[8], suggesting a need for additional editors to become administrators. This model could help identify strong candidates.

Following the lessons learned by Cosley and colleagues' SuggestBot [4], which matched pages needing work with editors who had similar interests, a kind of "AdminFinderBot" would need to follow Wikipedia norms and work with the bot approval committee to be most effective and accepted. As a very lightweight process, it already meets one of the main criteria for Wikipedia bots: it would not be a server hog. It would also need to respect the privacy of editors: As the discussion on the archival page of non-administrators with high edit counts shows[9], some highly contributing members do not want to become administrators, so their preferences will need to be considered in the implementation.

---

[8] http://en.wikipedia.org/wiki/Category:Administrative_backlog

[9] http://en.wikipedia.org/w/index.php?title=Wikipedia:List_of_non-admins_with_high_edit_counts&oldid=212771572

Models similar to this one can be used to analyze promotion decisions in other online communities, as well. In open source software projects, for example, Ducheneaut finds that being promoted to "committer" status requires learning both norms and politics [6]. Providing statistics such as the known defect density of a contributor's code, his number of messages to the project listserv, who he collaborates with, and his number contributions to other projects may assist project managers in deciding whether to grant committer status to a contributor, or allow that contributor to evaluate himself compared to others in the project. Health care forum administrators seeking moderators, or World of Warcraft guilds seeking new leaders could use similar tools to search for current participants with behaviors similar to current leaders. While statistically-based dashboards may not be good replacements for deliberate human judgment, decision-making research shows that people are poor at making these kinds of decisions, and that statistical models often perform better [5,9]. Thus, statistics may aid in informing promotion discussions.

Although we have emphasized the practical use of the models predicting status change in Wikipedia, the findings also contribute to the larger literature on policy capture, which often finds moderate differences between subjective estimates of policy importance and objective captured policy [19,23]. However, most studies of policy capture have employed artificial rating situations that some researchers argue are very different from actual evaluation settings and make generalization to real-world situations difficult [10]. Here we demonstrate the benefits that archival online communities such as Wikipedia present as a way to study policy capture in actual promotion settings.

Limitations and Future Work

An important limitation of the current model is that it uses easy-to-measure behavioral data as proxies for more abstract concepts. In particular, we do not capture well what Wikipedians mean when they refer to "user interaction" or "varied experience." And, as the Guide to RfA states, Wikipedians themselves may have very different personal definitions of these criteria. An alternate explanation for the discrepancies between voters' stated criteria and the significant criteria in the model is that simplistic inputs led to a model that is incorrect. This could account for cases, such as the failure to find that observing consensus leads to promotion, where seemingly important variables show no effect. However, the model also reveals unexpected effects, such as the negative impact of arbitration attempts or helping with chores. We are currently analyzing the rationale stated by RfA voters and the concrete behaviors they cite in their promotion decisions [3]. This qualitative research illuminating the particular strategies and metrics RfA voters use will fill in many of the gaps left by the present quantitative model.

The model performs moderately well using easy-to-measure behavioral data, but it could potentially be improved with better machine-learning models employing natural language processing or information retrieval. Furthermore, the model does not take the quality of contribution into account. We plan to improve the model by examining measures of length, persistence, and pageviews of edits, which are already being used in more computationally complex models of existing administrator behavior [13] and the impact of edits [18].

Eager editors seeking to "level up" to administrator status should note that the findings presented here are correlational, rather than causal. Editing thousands of articles and saying "thanks" does not automatically lead to promotion; rather, these behaviors may be correlated with other underlying behaviors desirable in Wikipedia administrators, such as responsibility and courtesy. Furthermore, voters' standards change over time. These changes are evidenced in the reduction in successful RfAs between the encyclopedia's first five years and the two most recent ones and the trend in recent RfA dialogs to ask candidates to respond to hypothetical scenarios, in which they describe how they would deal with problematic cases[10]. Thus, accurate models are likely to change with time and should weight recent behavior more heavily to adapt to changing standards.

Although this research has shown that judges pay attention to candidates' job-relevant behavior and especially behavior that suggests the candidate has extensive and varied experience, it is silent about whether other factors identified in the organizational literature [17]—social networks, irrelevant attributes, or strategic self-presentation. Future research in Wikipedia using techniques like those in the current *paper* can be used to test theories in organizational behavior about criteria for promotion.

Finally, the model is based only on nominated candidates for promotion, rather than the general editor population of Wikipedia. We intend to extend the model to allow prediction across all editors, not just those singled out (or self-nominated) for promotion. Furthermore, the model only describes nominees who become administrators, not necessarily those who become good administrators. To predict good administrators, one needs to answer additional questions. How should good administrators be measured? Is it sticking around and maintaining high levels of activity? "Taking up the mop" and diligently clearing out administrative backlogs? Eventually becoming a bureaucrat? Our next step is to determine if administrators change their behavior after their RfA, to determine what happens after a community promotes a member to a

[10]http://en.wikipedia.org/wiki/User:Filll/AGF_Challenge_2_Directions

managerial role, and to measure the quality of that manager's future behavior.

ACKNOWLEDGMENTS

We would like to thank Niki Kittur and Ben Collier for feedback on the model creation, and Tom Murphy VII for insight into Wikipedia adminship. This work is supported by NSF grants IIS IIS-0325049 and IIS-0729286 and an NSF Graduate Research Fellowship.

REFERENCES

1. Bryant, S. L., Forte, A., & Bruckman, A. (2005). Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. GROUP 2005, 1-10.
2. Butler, B., Joyce, E., and Pike, J. Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. Proc. CHI 2008, ACM Press (2008).
3. Collier, B., Burke, M., Kittur, N. and Kraut, R.E., Retrospective versus Prospective Evidence for Promotion: The Case of Wikipedia. In *2008 annual meeting of the Academy of Management*, (Anaheim, California, 2008), Academy of Management.
4. Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. SuggestBot: Using intelligent task routing to help people find work in Wikipedia. Proc. IUI 2007, ACM Press (2007), 32-41.
5. Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. American Psychologist, 34(7), 571–582.
6. Ducheneaut, N. Socialization in an open source software community: A socio-technical analysis. Computer Supported Cooperative Work 14 (2006), 323-368.
7. Forte, A., and Bruckman, A. Scaling consensus: Increasing decentralization in Wikipedia governance. Proc. HICSS 2008.
8. Giles, G. (2005). Internet Encyclopedias Go Head to Head. Nature, 438(7070), 900-901.
9. Hastie, R., & Dawes, R. M. (2001). Rational choice in an uncertain world: The psychology of judgment and decision making. Thousand Oaks, CA: Sage Publications.
10. Hobson, C. J., & Gibson, F. W. (1983). Policy Capturing as an Approach to Understanding and Improving Performance Appraisal: A Review of the Literature. The Academy of Management Review, 8(4), 640-649.
11. Hocking, R. (2003). Collinearity in multiple regression. Chapter 5 of Methods and Applications of Linear Models: Regression and the Analysis of Variance. Hoboken, NJ: Wiley-Interscience..
12. Karau, S., and Williams, K. Social loafing: A meta-analytic review and theoretical integration. Journal of Personality and Social Psychology, 65 (1993), 681-706.
13. Kittur, A., Chi, E., Pendleton, B., Suh, B., and Mytkowicz, T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. Proc CHI 2007, ACM Press (2007).
14. Kittur, A., Suh, B., Pendleton, B. A., Chi., E. (2007). He says, she says: Conflict and coordination in Wikipedia. Proc CHI 2007, ACM Press (2007), 453-462.
15. Locke, E.A. and Latham, G.P. Building a practically useful theory of goal setting and task motivation: A 35 year odyssey. American Psychologist, 57 (9). 705-717.
16. Meehl, P.E. Clinical vs. statistical prediction: A theoretical analysis and review of the literature. Menneapolis, University of Minnesota Press, 1954.
17. Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. Personnel Psychology, 58(2), 367-409.
18. Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen, L., and Riedl, J. 2007. Creating, destroying, and restoring value in Wikipedia. Proc GROUP 2007, ACM Press (2007), 259-268.
19. Stumpf, S. A., & London, M. (1981). Capturing rater policies in evaluating candidates for promotion. The Academy of Management Journal, 24(4), 752-766.
20. Viegas, F., Wattenberg, M., and Dave., K. Studying cooperation and conflict between authors with history flow visualizations. Proc CHI 2004, ACM Press (2004), 575-582.
21. Viegas, F., Wattenberg, M., Kriss, J., and van Ham, F. Talk before your type: Coordination in Wikipedia. Proc HICSS 2007, 575-582.
22. Vroom, V., Porter, L., and Lawler, E. Expectancy Theories. in Miner, J.B. ed. Organizational Behavior 1: Essential Theories of Motivation and Leadership, ME Sharpe, Armonk NY, 2005, 94-113.
23. Zedeck, S., & Kafry, D. (1977). Capturing Rater Policies for Processing Evaluation Data. Organizational Behavior and Human Performance, 18(2), 269-294.