A

Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion[1]

SANMAY DAS, Washington University in St. Louis
ALLEN LAVOIE, Washington University in St. Louis
MALIK MAGDON-ISMAIL, Rensselaer Polytechnic Institute

Our reliance on networked, collectively built information is a vulnerability when the quality or reliability of this information is poor. Wikipedia, one such collectively built information source, is often our first stop for information on all kinds of topics; its quality has stood up to many tests, and it prides itself on having a "Neutral Point of View". Enforcement of neutrality is in the hands of comparatively few, powerful administrators. In this paper, we document that a surprisingly large number of editors change their behavior and begin focusing more on a particular controversial topic once they are promoted to administrator status. The conscious and unconscious biases of these few, but powerful, administrators may be shaping the information on many of the most sensitive topics on Wikipedia; some may even be explicitly infiltrating the ranks of administrators in order to promote their own points of view. In addition, we ask whether administrators who change their behavior in this suspicious mannger can be identified in advance. Neither prior history nor vote counts during an administrator's election are useful in doing so, but we find that an alternative measure, which gives more weight to influential voters, can successfully reject these suspicious candidates. This second result has important implications for how we harness collective intelligence: even if wisdom exists in a collective opinion (like a vote), that signal can be lost unless we carefully distinguish the true expert voter from the noisy or manipulative voter.

1. INTRODUCTION

Increasingly, we get information from networked sources that rely on some form of collective intelligence. We turn to information aggregated on the web for everything from product reviews (e.g. Amazon) to travel planning (e.g. TripAdvisor) to basic information on just about any topic (Wikipedia). In the context of the emerging field of computational social science [Lazer et al. 2009], there has been a range of work on the quality of information available through such sources. A particular recent focus has been on trustworthiness, and incentives for subverting these kinds of information aggregation venues. Most of the work on trust has been in the context of recommendation systems covering issues like fake and paid reviews. Wikipedia, which crowdsources the collection of knowledge to millions of editors and is generally regarded as high-quality [Giles 2005], is another major target for manipulation. Thousands of editors are elected as administrators, responsible for conflict resolution and policy enforcement. Administrators have significant social and technical clout which allows them to carry out these functions. Thus, administrators have the ability to significantly influence the readership. Indeed, leaked communications from the political advocacy group CAMERA included plans for electing administrators who could then influence the Israel–Palestine debate ["Candid CAMERA" 2008]. There have also been prominent scandals involving "administrators for hire", who offer to edit for money.[2] Recently, an administrator was banned from English Wikipedia for manipulating the encyclopedia to promote an unaccredited Indian business school [Sloan 2015].[3]

1.1. Administrators and manipulative behavior

To become an administrator, an editor submits a Request for Adminship (RfA). Thereafter, the editor's history on Wikipedia is scrutinized by other editors, and by current administrators. The user must demonstrate good citizenship and the qualities and work

[1]This is an extended version of a paper which appeared at CIKM 2013 [Das et al. 2013].
[2]http://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment/Paid_Editing
[3]https://en.wikipedia.org/wiki/Wikipedia:Arbitration/Requests/Case/Wifione

ethic expected of an administrator. After some time, the editorship votes on whether to promote the candidate or not. After a successful RfA, there is little further oversight as long as the administrator does not blatantly violate Wikipedia policy. The basis for the plan revealed in the CAMERA emails was to exploit this RfA election process. Specifically, their goal was to have members of their group become administrators by displaying edit behavior expected of administrators; then, after successful RfAs, to use their administrator status to influence disputes relating to the Israeli–Palestinian conflict.

While some recent work addresses questions of petty vandalism and the amount of minor janitorial work needed to maintain Wikipedia, there has been no systematic study of targeted manipulation of Wikipedia. We describe the results of such a study in this paper. We propose and validate a measure for quantifying "suspicious" behavior of editors on Wikipedia. Our measure, the *Clustered Controversy* (or CC-) score, captures the focus that an editor has on a particular controversial topic (for example, conflict in the middle east). The measure provides a tool that allows us to not only assess such behavior in isolation, but also to identify patterns that may indicate suspicious *changes* in behavior.

We then use this method to analyze the behavior of editors who successfully become administrators. We find that a higher than expected fraction of successful RfA candidates increase their CC-scores a large amount shortly after election; these admins are exerting significantly more control over controversial topics on Wikipedia, and doing so in a topically clustered way. We do expect them to use their new powers on controversial topics—administrators are expected to intervene in disputes—but in a broad sense, not focusing on topically clustered controversial articles. These administrators may be either trying to help out discussions on a topic in good faith (although even in this case they may unconsciously inject their biases into the pages in question), or they may be infiltrators whose goal was to become administrators primarily to change the conversation on these topics.

1.2. Identifying manipulators prior to election

Is it possible to identify potentially manipulative administrators by their behavior *before* the RfA? We show that two intuitive tests fail to do so. (1) RfAs are accepted or rejected based on the percentage of editors who support a candidate. This vote percentage does not filter out manipulative administrators: if anything, candidates who go on to change their behavior in suspicious ways receive a higher vote percentage. (2) Burke and Kraut [2008] introduced an estimate of the quality of an editor's RfA that is based purely on the behavior of the editor (we refer to this measure as the prior-activity score). The prior-activity score attempts to measure "admin-like" behavior on Wikipedia prior to an RfA, such as participation in maintenance tasks and dispute moderation. The prior-activity score is also unable to filter out manipulative administrators; again, those with higher prior history scores are actually more likely to display suspicious behavior after the RfA.

However, it is possible to reject potentially manipulative candidates by using a measure designed for crowd-sourced spam detection [Ghosh et al. 2011] (we refer to this as the weighted-voter score). This measure gives more weight to more influential voters. Editors with very high weighted-voter scores are unlikely to change their CC-Scores significantly after promotion, whereas those with lower scores are more likely to do so. This indicates that the collective intelligence of the RfA process is capturing something about behavior that is not reflected in the purely quantitative history of the editor's behavior. Actually reading an editor's history of contributions and making an informed decision is valuable. However, this wisdom is lost when computing a simple percentage of support votes for a candidate. Thus, the RfA process already reveals the information needed, but using a simple percentage to aggregate votes is not sufficient. In this case, making informed decisions using crowdsourced opinions requires first learning about the members of the crowd.

2. RELATED WORK

There is a large literature on many different aspects of Wikipedia as a collaborative community. It is now well-established that Wikipedia articles are high quality [Giles 2005] and very popular on the Web [Spoerri 2007]. The dynamics of how articles become high quality and how information grows in collective media like Wikipedia have also garnered some attention [Wilkinson and Huberman 2007; Das and Magdon-Ismail 2010]. While there has not been much work on how Wikipedia itself influences public opinion on particular topics, it is not hard to draw the analogy with search engines like Google, which have the power to direct a huge portion of the focus of public attention to specific pages. Hindman et al. [2003] discuss how this can lead to a few highly ranked sites coming to dominate political discussion on the Web. Subsequent research shows that the combination of what users search for and what Google directs them to may lead to more of a "Googlocracy" than the "Googlearchy" of Hindman et al. [Menczer et al. 2006].

Our work in this paper draws directly on three major streams of literature related to Wikipedia. These are, work on conflict and controversy, automatic vandalism detection, and the process of promotion to adminship status on Wikipedia.

There is a significant body of work characterizing conflict on Wikipedia. Kittur et al. [2007] introduce new tools for studying conflict and coordination costs in Wikipedia. Vuong et al. [2008] characterize controversial pages using both disputes on a page and the relationships between articles and contributors. We use the measures identified by Kittur et al. and Vuong et al. as a starting point for measuring the controversy level associated with a page. This then feeds into our user-level C-Score and CC-Score measures. Our results on the blocked users dataset serve as corroborating evidence for the usefulness of these previously identified measures. Conflict on Wikipedia is traditionally resolved by appealing to outside sources. However, Lopes and Carriço [2008] find that accessibility issues significantly impede this process. Welser et al. [2011] identify social roles within Wikipedia: substantive experts, vandal fighters, social networkers, and technical editors

Automatic vandalism detection has been a topic of interest from both the engineering perspective (many bots on Wikipedia automatically find and revert vandalism), as well as from a scientific perspective. Potthast et al. [2008] use a small number of features in a logistic regression model to detect vandalism. Smets et al. [2008] report that existing bots, while useful, are "far from optimal", and report on the results of a machine learning approach for attempting to identify vandalism. They conclude that this is a very difficult problem to solve without incorporating semantic information. While we touch on vandalism in dealing with blocked users, we are focused on "POV pushing" by extremely active users who are unlikely to engage in petty vandalism, which is the focus of most work on automated vandalism detection.

Wikipedia administrator selection is an independently interesting social process. Burke and Kraut study this process in detail and build a model for which candidates will be successful once they choose to stand for promotion and go through the Request for Adminship (RfA) process [Burke and Kraut 2008]. The dataset of users who stand for promotion is useful because it allows us to compare both previous and later behavior of users who were successful and became admins and those who did not.

3. DATA AND METHODOLOGY

We begin by discussing our methodology in computing a "simple" Controversy Score for each user, and then describe how we can compute a Clustered Controversy Score to find editors who focus on articles related to a single, controversial topic. All data is from the entire history of English Wikipedia as of February 2012.

We introduce a simple measure that captures the proportion of attention an editor focuses on contentious topics. We call this the Controversy Score (C-Score). Using the C-Score,

we confirm that administrators participate in controversial topics significantly more than they did as editors prior to their RfA. This is not surprising, because one of the major roles of an administrator is conflict resolution, and it is needless to say that conflicts will arise disproportionately in contentious topics. Thus, controversy per se is not indicative of a manipulative editor. This motivates a more refined behavioral measure, our Clustered Controversy Score (CC-Score).

3.1. Controversy Score

We define the C-Score for a user as an edit-proportion-weighted average of the level of controversy of each page. The controversy of a page follows the article-level conflict model of Kittur et al. [2007]: we train a regression model to predict the number of revisions to an article which include the "{{controversial}}" tag (CRC, or Controversial Revision Count). Intuitively, articles which spend more revisions marked as being controversial are more likely to actually be controversial, since this implies at least implicit agreement with the designation by more authors. Since Kittur et al. study a 2006 Wikipedia dataset, we perform some additional validation on our newer data. As in Kittur et al., we only train on articles which are controversial in the latest revision available in our dataset. This leaves 1640 articles, of which we train on a randomly selected 1000 and test on 640. We use the same features: revision counts, page length, unique editors, links, anonymous edits, administrator edits, minor edits, reverts, and combinations of these involving the talk pages, article, or both. This yields an R^2 of 0.79 on our test set, somewhat lower than Kittur et al. report from 2006. We use this predicted CRC to measure controversy for each Wikipedia article, computed using the regression model. To normalize the page-level score, we divide by the predicted CRC of the most controversial page (the page for Wikipedia itself). This yields a score between 0 and 1 for each page which we would expect to correlate well with expert judgments of controversy based on the comparisons to such judgments performed in Kittur et al. [2007].

Let p_k be the fraction of a user's edits on page k. The controversy score for a user is then an edit-weighted average of the page-level controversy scores:

$$\text{CScore} = \sum_k p_k c_k \tag{1}$$

We would expect this measure to be effective at finding users who edit controversial pages. However, as mentioned above, many Wikipedia users dedicate at least part of their time to removing blatant vandalism, which occurs disproportionately on controversial pages. Thus we turn to a measure that combines topical clustering with controversy.

3.2. Clustered Controversy Score

While all administrators deal with controversial topics on a regular basis, they are supposed to do so in a neutral way. A sudden sharpening of focus may indicate an undisclosed interest; and especially if that topic is controversial, the behavior change is suspicious.

In order to measure topical concentration, we could define topics globally, but this is both expensive and sensitive to parameter changes: what is the correct granularity for a topic? Instead, we focus on a local measure of topical concentration. Given a similarity metric between articles, we can measure the extent to which a user's edits are clustered. We extend a clustering measure originally developed for gene networks [Kalna and Higham 2007] to quantify how coherent an administrator's controversial edits are.

3.2.1. Page similarity. There are many approaches to comparing text documents based on word frequencies. We first model articles as belonging to a relatively small set of topics, then base comparisons on those topics. To find the topics associated with each article, we train a topic model—Latent Dirichlet Allocation (LDA) [Blei et al. 2003]—on the text of Wikipedia

pages. We use a procedure similar to Griffiths and Steyvers [2004]. We model articles as containing a mix of 1000 topics, which allows fine-grained comparisons while avoiding the curse of dimensionality inherent in comparisons with orders of magnitude more features. LDA finds a distribution over these topics for each article, effectively clustering them. We compare the resulting topic distributions using cosine similarity.[4] Thus we make abstract comparisons between articles based on topics rather than concrete words or structural features. Alternative methodologies are explored in Section 5.

3.2.2. Computing the CC-Score. Consider a set of edits from a user. Let N be the number of unique pages in this set and w_{ij} be the similarity score between pages i and j. We start with a generalization of the clustering coefficient to graphs with edges between 0 and 1 [Kalna and Higham 2007]. Let p_k be the proportion of a user's edits on page k, and c_k be some measure of controversy. For a page k, define the impact of that page as:

$$\iota(k) = c_k p_k \tag{2}$$

Then the clustering score of a page is:

$$\text{clust}(k) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \iota(i)\iota(j)w_{ki}w_{kj}w_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} \iota(i)\iota(j)w_{ki}w_{kj}} \tag{3}$$

$\text{clust}(k)$ is a weighted average of the connection strengths between neighbors of k. It is higher when the controversial, highly edited, and well connected neighbors of k are themselves similar[5]—that is, when a page is connected to a coherent and controversial topic which the user edits frequently. Note that $\text{clust}(k)$ depends heavily on the user's local edit graph, and is not a proper function of the page k. Finally, we combine the page-level clustering scores into a user-level score:

$$\text{CCScore} = \sum_{k=1}^{N} \iota(k)\text{clust}(k) \tag{4}$$

If $c_k, p_k \in [0,1]$, then $\text{CCScore} \in [0,1]$.

There is no reason that c_k must be a measure of controversy. Instead, it can measure any property of a page which is of interest. For example, a c_k measuring how much a page relates to global warming would yield a ranking of editors based on the extent to which their edits concentrate coherently on global warming. The CC-Score is a general tool for ranking single-topic contributors. We also compute a raw Clustering Score where each page has $c_k = 1$ in (4)—this yields a measure of topical clustering independent of any properties of the particular pages.

We choose a measure that combines clustering and controversy page-wise rather than user-wise so that we do not end up with editors who are very topically focused on uncontroversial pages, but also spend a significant fraction of their time combating vandalism across a spectrum of topics. We also note that the only Wikipedia-specific contributions to the CC-Score are encapsulated in the computation of c_k and w_{ij}. The same quantities can be computed for a wide variety of collaborative networks. Consider email messages: w_{ij} between two threads could be based on message text, and c_k based on the length of the

[4] Alternatively, since we are comparing distributions, we could employ Jensen-Shannon Divergence. We ran a subset of our experiments using different similarity metrics as a robustness check, and did not observe any qualitative changes in results.

[5] Including the controversy and edit fraction of connected nodes, as we do through a page's impact $\iota(\cdot)$, deviates from a traditional clustering coefficient. The edit fraction avoids focusing disproportionately on connections to lightly edited pages. Similarly, we are more interested in connections to a user's other controversial edits.

thread as a measure of controversy. These quantities can be entirely language independent, for example replacing text with a contributor-based similarity model [Li et al. 2011].

3.3. The RfA process

Standing for promotion to adminship on Wikipedia is an involved process. An editor who stands for, or is nominated for, adminship must undergo a week of public scrutiny which allows the community to build consensus about whether or not the candidate should be promoted. A special page is set up on which the candidate makes a nomination statement about why she or he should be promoted, based on detailed evidence from their history of contributions to Wikipedia. Other users can then weigh in and comment on the case, and typically a large volume of support (above 75% of commenters) as well as solid supporting statements from other editors are necessary for high-level Wikipedia "bureaucrats" to approve the application. Burke and Kraut [2008] provide many further details on this process. Wikipedia policies call for nominees to demonstrate a strong edit history, varied experience, adherence to Wikipedia policies on points of view and consensus, as well as demonstration of willingness to help with tasks that admins are expected to do, like building consensus. Burke and Kraut note that the actual value of some of these may be mixed: participating in seemingly controversial tasks like fighting vandalism or requesting admin intervention on a page before becoming an admin actually seems to hurt the chances of success.

Overall, the Wikipedia community devotes significant effort to the RfA process, and there is a lot of human attention focused on making sure that those who become admins are worthy of the community's trust.

3.4. Scoring RfAs

There is a significant amount of information associated with the RfA process aside from the binary determination of whether a user should be an administrator or not. We can use this information to determine what, if anything, the RfA process reveals about the future behavior of an administrator. We use two proxies for RfA quality: behavioral features of a candidate which predict RfA success, and the votes and voting history of users who participate in the RfA. We can compare these measures to simply using the percentage of support votes a candidate receives during an RfA.

3.4.1. Prior activity. We implement the model of Burke and Kraut [2008], which uses overall activity and participation in admin-like activities to model the administrator selection process and predict which RfAs will be successful. They perform a probit regression with success in the RfA as the dependent variable and features that encode characteristics including "strong edit history," "varied experience," "user interaction," "helping with chores," "observing consensus," and providing "edit summaries" as the independent variables. We perform the same regression and use the estimated probability p_i that editor i's RfA will be successful. This proxy for RfA success, which does not take votes or voters into account, still predicts success well, with an AUC of 0.82.

Table I shows the results of the RfA-success-predicting probit regression, based on the results of Burke and Kraut [2008]. Our regression is over a longer period of time, so we have added the RfA date as a feature to accommodate changes in the process (it has become significantly harder to become an administrator). We use a standard probit regression, omitting some features used by Burke and Kraut which had very little effect in their regression. To test performance, we held out a randomly selected 5% of the RfAs (yielding the 0.82 AUC figure referenced above).

3.4.2. Voter model. Wikipedia typically eschews decisive voting in favor of consensus building. Many Wikipedians would claim that a simple vote percentage is close to meaningless, or at least that it is not sufficient for a high quality RfA (although we document below that

Feature	Mean	Std.	Change in prob.	
Attempt number	1.2	0.6	-7.1%	***
Articles edited	1902	4060	7.4%	***
Months since first edit	16.0	12.8	4.1%	***
Date of rfa (months since 2000)	88.6	18.4	-11.2%	***
Namespaces edited	10.5	3.2	0.1%	
Wikipedia policy edits	738	1202	2.2%	*
Article talk edits	540	1287	0.9%	
User talk edits	1124	3071	-5.2%	***
Wikipedia talk edits	113.0	247.0	1.6%	*
Arbitration edits	49.6	185.1	-1.3%	
"Thanks" in edit summary	24.8	44.5	3.9%	***
Reverts (from edit summary)	914.7	3583.8	1.4%	
Vandal reporting (AIV)	49.6	171.3	-2.0%	**
Requests for protection	34.0	160.4	-0.4%	
"Npov" in edit summary	27.6	51.0	0.4%	
Administrator attention (ANI)	124.2	342.9	7.1%	***
Minor edits (%)	27%	23%	2.6%	***
Articles for deletion (AfD)	326.1	1155.0	0.5%	
Other RfAs	93.7	245.6	-2.5%	***
Ideas (village pump)	25.1	91.4	-1.6%	
Edits summarized (%)	80%	20%	6.6%	***

Table I: Features for the probit regression predicting the probability of a successful RfA, with the mean and standard deviation of feature values, the effect of moving up one standard deviation in the given feature (starting with a vector of mean feature values), and the result of a significance test for the feature weight (*** $p = 0.001$, ** $p = 0.01$, * $p = 0.05$).

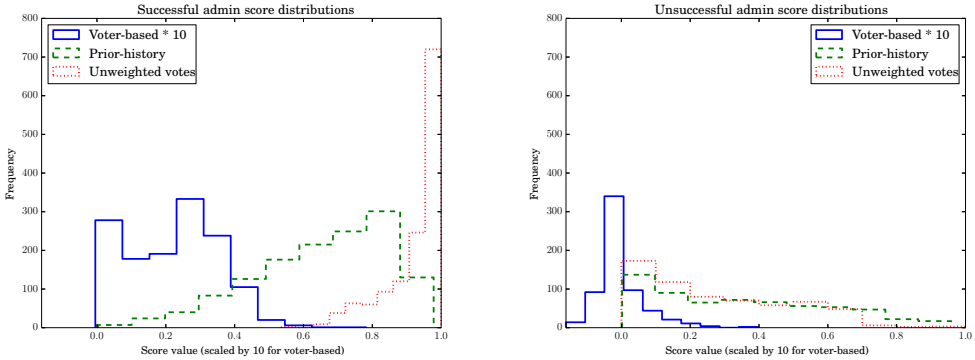

Fig. 1: Distributions of the three RfA or pre-RfA scores for admin candidates. Successful candidates are shown on the left, unsuccessful on the right. The weighted-voter score is multiplied by a factor of 10 to show detail.

it is the most predictive measure of success). We can attempt to improve upon the simple vote percentage by inferring the quality of voters.

We adapt a technique of Ghosh et al. [2011] for aggregating noisy votes in abuse detection for user-generated content. On websites where many users rate some content, how does one differentiate between bad content and a bad rater? The basic idea behind Ghosh et al.'s

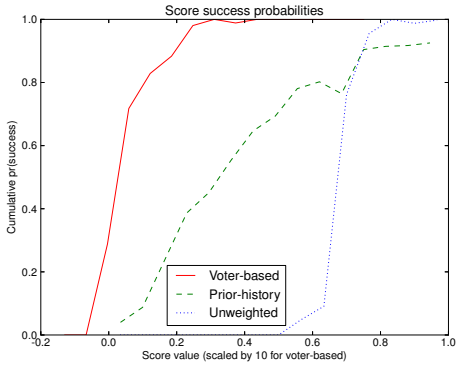

Fig. 2: Probability of a successful RfA as a function of the weighted-voter score, the prior-activity score, and the unweighted vote fraction. The weighted-voter score is multiplied by 10 to show detail.

technique is to discover probabilities with which each rater provides a correct rating of some content; these probabilities serve as a measure of user quality. They show that if you know the identity of a single agent who provides a correct rating with probability greater than chance, it is possible to achieve good performance.

For RfAs, we use the outcome of the RfA as our signal of "$50\% + \epsilon$" correctness (assuming only that the judgments of the bureaucrats who make the final decision are not pathologically incorrect). The algorithm implicitly determines the "trustworthiness" of each voter and aggregates weighted votes into an explicit score for each RfA. We use this score directly in our analyses.

In contrast to the activity-based score, the weighted-voter score depends only on the RfA process itself. The procedure is a straightforward application of the algorithm of Ghosh et al. [2011]. We first construct a matrix U, with each element u_{ji} being the rating of RfA j by user i: 0 if i did not vote on RfA j or cast a neutral vote, 1 if i cast a positive vote, and -1 if i cast a negative vote. As in Ghosh et al., columns of U are then vectors of ratings by a given user. Under their model, each user has some probability of correctly marking an item (in our case an RfA), and these probabilistic markings can be aggregated by taking the top eigenvector of UU^T (without first knowing each user's probability). The top eigenvector of UU^T then represents *two* possible consensus estimates under the probabilistic rating model, exactly opposite, of the quality of each RfA. The ambiguity arises because we have never told the model which users are "right", but merely which users are in agreement. To disambiguate, we select the consensus estimate that is closer to the true RfA outcomes (i.e. decisions by Wikipedia "bureaucrats", who formally add administrator status after judging an RfA to be successful). Note that this is only a single bit of information, essentially assuming that the majority is not pathologically incorrect in its judgments (formally that greater than 50% of RfAs are judged "correctly").

This procedure has the effect of weighting some users more highly, judging them to give "correct" ratings to RfAs more often. As we only run the procedure once on all of the RfA votes in our dataset, we use some information about the voting behavior of RfA participants chronologically after an RfA in question, and so the procedure as we implement it is strictly post hoc. However, one could easily "score" an RfA in real time by using only votes cast in it and previous RfAs. The "future information" given to the algorithm in our implementation is unrelated to administrator behavior changes, and so will not qualitatively affect our results.

3.4.3. Comparing the models. We first note in practice, the simple support percentage effectively determines the outcome of an RfA (AUC 0.998, with a *de facto* threshold at 69%). The weighted-voter model achieves an AUC of 0.94 (editors with scores below zero are exceedingly unlikely to succeed, while those with scores above 0.02 almost always do), while the prior-activity model achieve an AUC of 0.82. Figure 1 shows the distributions of all three scores for successful and unsuccessful candidates. Figure 2 compares the distribution of success probabilities associated with the weighted-voter score with that of the prior-activity score and raw vote fraction. While the raw vote percentage is more discriminative than the weighted-voter score, we show later that unweighted votes behave more like the prior-activity score in terms of after-election administrator behavior (i.e. they select for a similar type of administrator).

These scores allow us to divide administrators into two broad clusters—the ones who receive a ringing endorsement from a given score, and those whose cases were more contentious. We can use these clusterings to differentiate the behavior of these two groups, and to compare the scores themselves. In particular, the contentious cases provide us a useful division into treatment and control groups – since many editors with borderline weighted-voter and prior activity scores do not make the cut, we can compare the behavior of two populations who were equally likely to be successful based on those scores, but some of whom happened to make it and some who didn't. We will use this to analyze the effect that becoming an admin plays on editors.

4. RESULTS

In this section, we first establish the validity of our metrics by examining whether they provide discriminatory power in identifying manipulative users. In order to do so, we need an independent measure of manipulation, so we focus on users that were blocked from editing on Wikipedia, and compare them with a similar set who were not blocked. We then move on to using the metrics to identify suspicious behavior in the population of admins. A reasonable hypothesis, suggested by the CAMERA messages discussed in Section 1, is that people who wish to seriously push their points of view on Wikipedia may try to become admins by editing innocuously, and then changing their behavior once they become admins. We test this hypothesis for the population of administrators by comparing the distribution of behavior changes among administrators with those of similar groups who did not become administrators.

4.1. Validation: Identifying manipulative users

We first validate the C- score and CC-Score by showing that they can find editors who are pushing their point of view. We use data on users blocked from editing Wikipedia in order to do so. Users can be blocked from Wikipedia for a variety of reasons. Reasons for blocks include blatant vandalism (erasing the content of a page), editorial disputes (repeatedly reverting another user's edits), threats, and more. Many blocks are of new or anonymous editors for blatant vandalism; we are not interested in these blocks.

We are interested in blocks stemming from content disputes. While editors are not directly blocked for contributing to controversial articles, controversy on Wikipedia is often accompanied by "edit warring", where two or more editors with mutually exclusive goals repeatedly make changes to a page (e.g., one editor thinks the article on Sean Hannity should be low priority for WikiProject Conservatism, and another thinks it should be high priority).

We examine a set of users who were active between January 2005 and February 2012. For blocked users, we use 180 days of data directly before their first block. For the users who were never blocked, the 180 days ends on one of their edits chosen randomly. To filter out new or infrequent editors, we only consider users with more than 500 edits. By examining only active users, we eliminate most petty reasons for blocks: users who have made significant

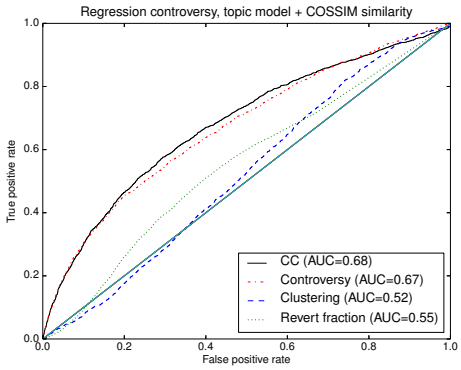

Fig. 3: ROC curve for CC, Controversy, and Clustering Scores when differentiating between blocked and not-blocked users, based on 180 days of data. As a baseline, the fraction of a user's edits during this period which were reverts is also included. The CC and Controversy Scores effectively discriminate between these classes, whereas the Clustering Score alone does not; there is no significant difference between the CC and Controversy Score curves. The curve indicates the true positive (TPR) at a given false positive rate (FPR) at different thresholds, when classifying each user as either blocked or not blocked. Area under the ROC curve (AUC) indicates how discriminative the scores are, and is the probability that a random blocked user is ranked higher by the given score than a random non-blocked user.

legitimate contributions are unlikely to start blatantly vandalizing pages. Finally, we only examine users who were blocked for engaging in point of view pushing: edit warring, 3 revert rule violations, sock puppets (creating another account in order to manipulate), and violations involving biographies of living persons. This leaves 2249 manipulative blocked users out of 4744 blocked users with at least 500 edits. There are 330720 total registered users who were blocked at least once in the dataset.

Figure 3 shows the performance of the CC, Controversy, and Clustering Scores when discriminating between the blocked users and users who were never blocked. Both the CC- and C-Scores show significant discriminative power, while Clustering alone is no better than guessing. As a baseline, we include the percentage of a user's edits which were reverts during the 180 day period used to compute the other metrics. Surprisingly, this revert fraction is barely more predictive than the Clustering Score. Account creation date was a somewhat better predictor, with an AUC of 0.59. A single model trained on these features (CC-Score, revert fraction, account creation date) had no better generalization performance than the CC-Score itself.

The performance of the CC- and C-Scores on the blocked users data set validates both measures for detecting users who make controversial contributions to Wikipedia. Many blocks in this data set involve violations of Wikipedia's "3 Revert Rule", limiting the number of contributions which an editor can revert on a single page during any 24 hour period, which implies that editors are not only making controversial changes but are vigorously defending them. This rule is not automatically enforced and does not apply to blatant vandalism; instead, another user must post a complaint which is then reviewed by an administrator. The discriminative power of the CC- and C-Scores in detecting this and other types of point of view pushing provides strong evidence that these scores are correctly detecting controversial editors.

4.2. High-scoring administrators insert more politically charged phrases

Finding manipulative users in the general population is a useful but somewhat indirect measure of whether administrators with high CC-Scores manipulate the encyclopedia at a higher rate than do administrators with lower CC-Scores. To address exactly this question in a direct way, we now turn to an analysis of the contributions that administrators themselves make to the Wikipedia articles they edit. We base this analysis on a single topic, U.S. politics, which has a relatively large set of natural language tools and corpora. The CC-Score is useful in part because it is topic agnostic, but we concentrate on a single topic here for validation. If those who score high on the CC-Score measure are more likely to insert politically charged phrases in the context of US politics, they are also more likely to do so in other controversial arenas.

We first collected 14145 revisions sampled from those of the top 20% of administrators by post-election CC-Score (randomly sampling 50 revisions per administrator), and an additional 14094 revisions in the same way from the bottom 20% by CC-Score. Using political bigrams and trigrams identified by Gentzkow and Shapiro [2010] as being indicative of partisanship in the U.S. Congressional Record, we count the number of revisions in each group of 14000 which have *added* one of these key phrases to an article.

As expected, the overall rate of administrators adding biased U.S. political phrases to articles is quite low (keep in mind that we did not filter for revisions relevant to politics or the U.S.). Among administrators with the lowest CC-Scores, it is 29 in 14094, while those with high CC-Scores added political phrases in 54 of 14145 revisions. The difference is statistically significant, with Fisher's exact test yielding $p = 0.008$. The result is nearly identical if we look at the number of administrators who have added a partisan phrase even once in the random sample of 50 of their edits. 46 out of 283 high-scoring administrators did so, but only 27 out of 283 low-scoring administrators ($p = 0.017$).

This analysis is not simply finding administrators who are interested in or mediating political articles, but rather those who insert phrases into articles which can be identified as either Democratic or Republican talking points. With regard to U.S. politics, therefore, the CC-Score does find manipulative behavior among administrators, with high-CC administrators adding biased phrases at nearly twice the rate of their low-CC counterparts.

4.3. Administrator behavior changes: Case studies

We have established that the CC- and C-Scores are indicative of manipulative behavior. However, an increase in controversy is expected among administrators. Even so, anecdotes such as those in Table II, which details the editing behavior of two admins with very large changes in CC-score immediately after promotion, indicate that suspicious behavior changes do exist, and that the CC-Score may be useful in finding them.

Another example of interest is the Wikipedia user Wifione, discussed in Section 1, an administrator who was banned from editing the encyclopedia for promoting the Indian Institute of Planning and Management (IIPM) and denigrating competitors [Sloan 2015]. Figure 4 shows the CC-Score of this user over time, from a period of intense IIPM editing early on, through a relatively restrained period directly before Wifione ran for administrator status (the consensus at the time seeming to be that Wifione had changed behavior for good), then a second period of questionable edits as an administrator, followed by inactivity and finally the ban. This example again highlights the value of the CC-Score for quantifying focused controversial editing.

4.4. Administrator behavior changes: Population level analysis

We now turn to analyzing the behavior of administrators at the population level, to identify whether there are serious issues with administrator manipulation beyond a few "bad apples." Figure 5 gives an overview of the (human-labeled) focus areas of administrators with very

Admin 1

Before RfA		After RfA	
Article	cc%	Article	cc%
Search engine optimization	48.7%	Homeopathy	73.8%
Web 2.0	14.7%	Waterboarding	22.1%
Kiev	12.3%	World Trade Center controlled demolition conspiracy theories	1.6%
Zango (company)	2.5%		
Wi-Fi	2.1%		
Vanessa Fox	2.1%	Electronic voice phenomenon	0.4%
Scientology	1.6%	Web 2.0	0.4%
Gamma-ray burst	0.8%	SS Edmund Fitzgerald	0.3%
Search engine submission	0.8%	Collapse of the World Trade Center	0.2%
Animal testing	0.8%		
		Naked short selling	0.2%
		Joe Lieberman	0.2%

Admin 2

Before RfA		After RfA	
Article	cc%	Article	cc%
Wikipedia	10.9%	Abortion	84.0%
Boolean algebra (structure)	9.3%	Support for the legalization of abortion	1.1%
The Beatles	5.5%		
Association football	3.3%	Safe sex	1.1%
Philosophy	3.0%	Condom	0.8%
Irony	2.7%	Hippie	0.7%
Lysergic acid diethylamide	1.9%	Fox News Channel	0.7%
Hippie	1.3%	Planned Parenthood	0.6%
Bill O'Reilly (political commentator)	1.3%	The Beatles	0.5%
		Masturbation	0.5%
Iraq War	1.2%	Lysergic acid diethylamide	0.4%

Table II: Two suspicious examples of large behavior changes 180 days before and after a successful RfA, with the percent contribution of that page to the user's CC-Score, selected from the top 5 largest log CC-Score changes among successful RfAs.

high and very low CC-Scores. It shows that those with high CC scores tend to focus on topics that we would intuitively view as more controversial. With this as background, we turn to statistical tests that can help tease apart the question of whether administrators change their behavior more than one would expect.

Our analysis focuses on three groups of Wikipedia users: (1) those who actually become administrators, (2) those who try unsuccessfully to become administrators, and (3) those who never make the attempt. The first two groups have self-selected to stand for promotion, either nominating themselves or accepting the nomination of another user. It is reasonable to assume that this group is not representative of the general population of Wikipedia users. Indeed, both successful and unsuccessful users who stand for promotion have significantly higher CC-Scores before their RfAs than a sample of those who never attempt to become administrators (p-value < 0.001). This may be due to "campaigning" by participating in admin-like activities, or could instead represent a tendency of more focused or controversial editors to want to participate in administration.

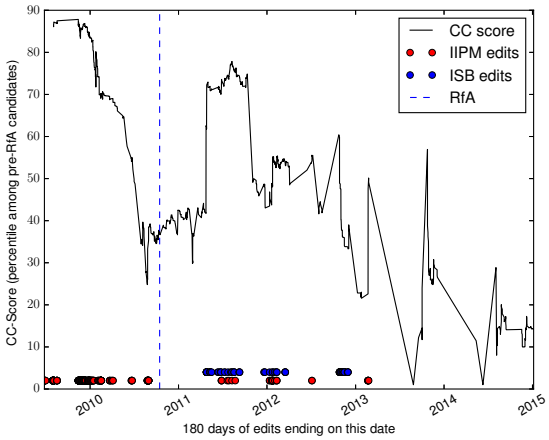

Fig. 4: A plot of the CC-Score (presented as a percentile among all candidates) of one Wikipedia user, Wifione, over time. After joining Wikipedia in 2009, Wifione began heavily editing articles related to the Indian Institute of Planning and Management (IIPM), but significantly reduced this type of editing before making a successful request for administrator status (RfA). After becoming an administrator, Wifione waited about eight months before again editing articles about IIPM and several of its competitors, including the Indian School of Business (ISB). Although relatively inactive after 2012, allegations of improper commercially-motivated editing (supporting IIPM and denigrating competitors) led English Wikipedia's Arbitration Committee to ban Wifione in February 2015.

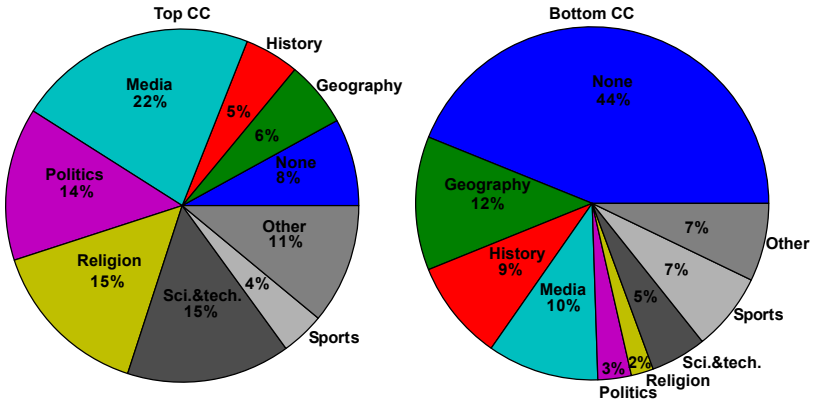

Fig. 5: Blind human evaluation of the general category of edits (if any) for administrators directly after their RfA. The 100 highest and 100 lowest scoring administrators according to a previous version of the CC-Score are shown (using metadata page comparisons and a slightly different controversy measure). The charts illustrate the behaviors which the CC-Score selects for in administrators: controversial edits on a focused topic.

We do not, however, find significant differences between the pre-RfA behavior of successful and unsuccessful candidates, as measured by the CC-Score. A t-test[6] comparing the

[6]Unless otherwise specified, we compute statistics using the log of the Clustering, C- and CC-Scores, as these log-transformed random variables are approximately normally distributed.

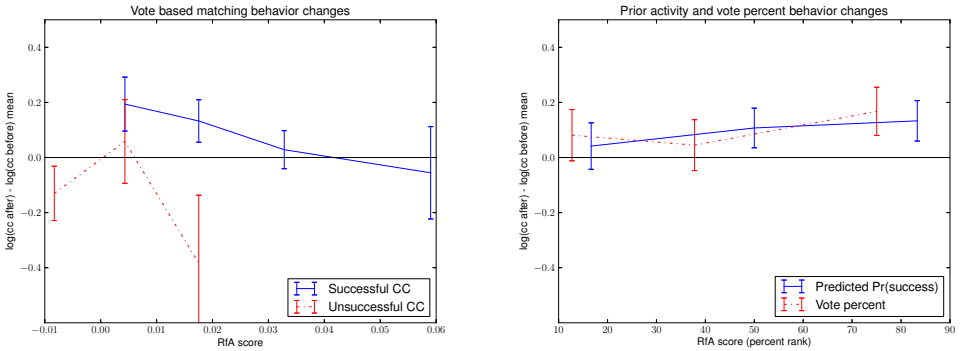

Fig. 6: The vote-based score of a Request for Adminship (RfA) (left) discriminates between administrators who change their behavior significantly and those who do not; a small group with low vote-based scores skew the average for successful administrators. The activity-based score (right) does not filter out administrators who change their behavior; if anything, higher scoring administrators are more likely to change their behavior. Raw vote percentage performs similarly.

expected values of the CC-Score for successful and unsuccessful candidates is inconclusive (p-value 0.87), meaning that we cannot reject the null hypothesis that these distributions have an identical mean. Neither does a KS-test find any statistically significant difference between the two distributions (p-value 0.06). Successful and unsuccessful candidates show nearly identical behavior before their RfAs, but how do they behave after either becoming an administrator or failing to do so? We now examine the effects of the outcome of the RfA process on these two groups, focusing on the changes in behavior between the pre- and post-RfA periods. Group 3 above (users who have never participated in an RfA) serve as a baseline for what constitutes typical behavior changes over time.

4.4.1. More suspicious behavior changes than expected among those who succeed in becoming admins. To summarize our statistical result: **the distribution of CC-Score changes among those who successfully become admins has a fatter tail in the positive direction than we would expect**.

Administrators are expected to engage in controversial topics. Therefore, we would expect editors to show an increase in their C-Score after promotion to administrator status, and indeed we do see this pattern. However, we also see a tightening of focus on controversial topics in a small group of successful administrators, measured by an increase in their CC-Scores. Users who never attempt to become administrators decrease their CC-Scores over time on average (95% confidence interval on the mean change in log CC score 180 days before and after a randomly chosen edit [-0.046, -0.015]). Intuitively, this corresponds to a broadening of interests: users who stick around tend to find new topics to contribute to (there is a corresponding decrease in clustering, but no decrease in controversy). In contrast, administrators as a group significantly increase their CC-Scores after election (95% confidence interval [0.05, 0.14]). How big is the problem? We find 119 successful administrators with changes that are above the 95$^{\text{th}}$ percentile of the distribution of changes in CC-Scores of Group 3 users (those who never tried to become administrators), while we would expect 67.5 due to random chance.

Administrators show significant increases in controversy, clustering, and CC-Score: they tighten their topical focus in an absolute sense, and do so on controversial topics. It is worth noting that administrators as a whole simultaneously *decrease* their clustering scores: while

they may edit on specific controversial topics, they are actually less focused than they were before becoming administrators.

4.4.2. Unsuccessful candidates are not suspicious. Our statistical result here is as follows: when comparing a matched sample of successful and unsuccessful candidates for promotion to admin status, **the change towards focusing on more controversial topics only occurs among those who actually become administrators**.

We break the successful candidates into two groups, and look at the group that was "just above threshold" in terms of their weighted-voter scores. This group has scores in the range where they could have been either successful or unsuccessful in their RfAs; we also examine the population of unsuccessful candidates that scored equally highly on the weighted-voter measure. The idea here, as in propensity score matching in general, is that the *only* differences in the two populations should be in whether they succeeded or not – they are not intrinsically different groups of people (ensured by leaving out the very-high scoring successful candidates and the very-low scoring unsuccessful candidates). Therefore, any differences in behavior can be attributed to something having to do with the actual effects of being an administrator, rather than an endogenous variable which made those people more likely to succeed in the first place. In our case, the matched group of unsuccessful candidates does not demonstrate an increase in the CC-score similar to that shown by the successful candidates (Figure 6, left). Many of the unsuccessful candidates actually decrease their scores, behavior typical of users who never attempt to become administrators. Therefore, we conclude that the change in behavior among successful admins who were "just above threshold" is not something that can be attributed to intrinsic features of the people, but is directly linked to the fact that they were actually successful in becoming admins. There would likely not exist the fat tail discussed above among this group of people if they had failed in their RfAs.

These conclusions are subject to the limitations on causal inferences inherent to a purely observational study. Nonetheless, propensity score matching is a standard methodology for estimating causal effects when experimentation is not feasible (see for example Aral et al. [2009]).

4.4.3. Suspicious behavior changes are predictable at RfA time, but only with the help of expert human judgment. To summarize in advance of presenting the detailed results: **successful administrators with high weighted-voter scores are much less likely to exhibit large changes in their CC scores than those with moderate weighted-voter scores**. The same is not true of simpler measures like raw vote count or the prior-activity model.

First, the weighted-voter results. We divide administrators into groups on the basis of their weighted-voter scores, and find that the C-Score rises significantly after election for each group (Figure 7). This is expected: administrators mediate disputes and deal with vandals, both of which target controversial pages disproportionately. In contrast, the behavior of the CC-Score is quite different when we examine it from the perspective of this grouping. There are distinct population-level behaviors among two clusters: administrators with moderately high weighted-voter scores show a statistically significant increase in their CC-Score after a successful RfA, whereas administrators with very high weighted-voter scores show no such increase (Figure 7).

For example, consider editors who succeed in their RfAs with a weighted-voter score below 0.025. Our data has 708 such cases, and a 95% confidence interval on the mean of the log ratio of the CC-Score is $[0.13, 0.27]$. Moreover, the distribution of behavior changes in this group is skewed toward large increases in topically focused controversial editing (skewness 0.24, p-value 0.01). Conversely, the 642 administrators with scores above 0.025 show neither statistically significant mean nor skewness in the same log ratio of CC-Scores. For comparison, this same high-scoring group shows both a significant average increase in

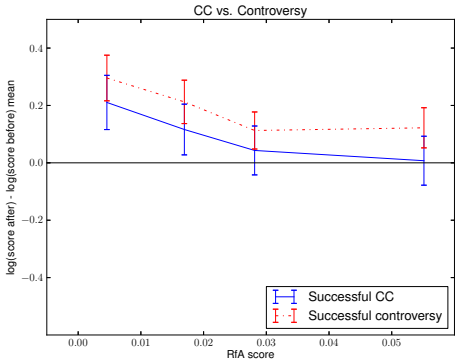

Fig. 7: Behavior changes upon becoming an administrator, measured by the CC- and C-Scores for 180 days of edits before and after a successful Request for Adminship (RfA). The x axis is the vote-based RfA score, with a higher score implying a stronger consensus. The Controversy Score increases on average for both low and high scoring administrators, while only low scoring administrators increase their CC-Score.

C-Score (95% confidence interval $[0.07, 0.17]$) and significant skewness in the distribution of the C-Score (skewness 0.65, p-value 4×10^{-10}).

One reasonable explanation might be that high scoring administrators have higher CC-Scores to begin with (pre-RfA), and that the low scoring administrators are simply "catching up". This is not the case: as with successful and unsuccessful candidates, the pre-RfA behavior of high and low scoring administrators is identical. Comparing the pre-RfA distributions of CC-Scores in these two groups (again using 0.025 as a splitting point), neither a t-test (p-value 0.50) nor a KS-test (p-value 0.51) finds a significant difference.

The conclusion is that administrators who are "just above threshold" by the weighted-voter score exhibit significantly different behavior as a group than administrators who were clearly well above the threshold. These just-above-threshold administrators are more likely to change their behavior significantly in the direction of pursuing more controversial topics.

Now, let us turn to simpler measures. We analyze the CC-Score changes of administrators using two other measures: the prior-activity model, and an unweighted voter model that simply looks at the proportion of positive votes on an editor's RfA. We find that neither of these measures is discriminative in the same way that the weighted-voter model is (Figure 6, right). When we group by the prior-activity score, there is no clear trend in CC-Score changes. If anything, the most likely candidates by this measure show the most suspicious behavior changes. Grouping by the unweighted vote count reveals no clear trend either. Quantitatively, there is a statistically significant negative correlation between the weighted weighted-voter score and changes in the CC-Score (lower scorers change behavior more), where we find no such relationship when considering the unweighted or prior-activity scores (there is a small positive correlation, but it is not statistically significant).

Our results show that the RfA process has significant discriminative potential in filtering out users who will change behavior upon becoming an administrator. Some members of the "just above threshold" group (using the weighted-voter score) may be misrepresenting themselves in order to become administrators, at which point they change their behavior significantly. Clearly, the RfA process has the potential to separate truly excellent administrators from this group, because those who score very highly on the weighted-voter measure do not change their behavior significantly.

Taken together, these results have important implications: the human element of the RfA process, in particular the votes and opinions of more informed and reliable humans, reveal extra information and are useful for keeping out those who may have nefarious intent, even if they misrepresent themselves as non-controversial editors beforehand. As a corollary, those with nefarious intent are quite good at concealing this intent in terms of various quantitative metrics, and may be using "less respected" voters in order to boost their scores when they stand for election to administrator status.

5. ALTERNATIVE SIMILARITY AND CONTROVERSY

The CC score relies on two main components: page controversy and page similarity. We have defined the score in Section 3 in terms of one particular choice of each. How sensitive are our results to these specific choices? In this section we explore several sets of features for assessing similarity, along with different ways of quantifying the similarities and differences between feature vectors.

5.1. Features: Topic modeling and metadata page features

How similar are two pages? This is an ill-defined question, with many possible answers. The text of a page, its links to other pages, the categories it is in, and the users who edit it are all informative in different ways about similarities. We consider a textual similarity that uses topic models, which allows for more abstract comparisons than word-level features would provide, and also consider another approach that makes use of the page metadata: links to other pages, the categories it is in, and the users who have edited it.

5.1.1. Topic modeling. After removing stop words and words which appear in only one document, we are left with 41180 terms. We then fit LDA using 1000 topics, with $\alpha = 0.05$ and $\beta = 0.1$ (symmetric parameters for the Dirichlet priors on topic and word distributions respectively) as suggested by Griffiths and Steyvers [2004]. For approximate inference on the model parameters, we use PLDA [Liu et al. 2011] to perform parallel Gibbs sampling. We use 100 iterations across 64 processes, which is roughly equivalent to 6400 sequential Gibbs sampling iterations (given an approximately linear speedup [Liu et al. 2011]). The log-likelihood converges well before this point.

Having computed the raw feature vectors r described above, we then compute a TF-IDF weighting in order to emphasize more specific similarities between pages. We use a standard formulation with log-transforms of both term frequency and inverse document frequency:

$$v_f^{(i)} = \left(1 + \ln r_f^{(i)}\right) \ln \frac{D}{d_f} \tag{5}$$

Where $d_f = \sum_j I\left(r_f^{(j)} > 0\right)$ is the number of documents having feature f and D is the total number of documents.

5.1.2. Metadata features. For a page of interest i, we have a binary vector indicating if there is a link to another page j (either incoming or outgoing). Likewise we have for each page a binary vector representing category membership, and finally a vector indicating how many times any given user has edited the page. We concatenate these vectors into a single feature vector representing the page. Since the meta-data features already cover various aspects we might want in an abstract comparison, we simply use a an inverse document frequency weighting rather than performing further processing.

5.2. Similariity measures: Cosine Similarity and Jensen-Shannon Divergence

Given the choice of one of the two sets of features described above, the next question is how we should translate vectors into a single number representing the similarity between two pages. Let v denote positive real-valued document vectors, and u denote vectors which

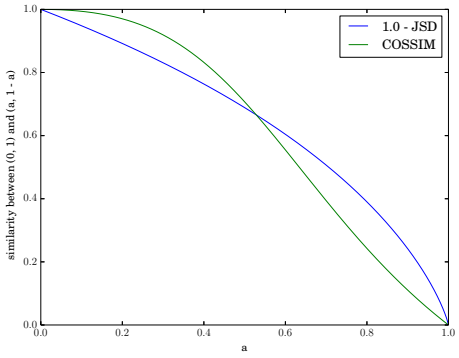

Fig. 8: Cossine similarity and Jensen-Shannon divergence when computing the similarity between a fixed discrete distribution $(0, 1)$ and a family of distributions $(a, 1-a)$ parameterized by a. Cossine similarity often takes more extreme values: closer to one than JSD when distributions are similar and closer to zero than JSD when distributions are dissimilar.

must be valid probability distributions. Cosine similarity, a common choice for general vector similarity, is defined as:

$$\text{COSSIM}_{ij} = \frac{v^{(i)} \cdot v^{(j)}}{||v^{(i)}||||v^{(j)}||} \tag{6}$$

An information-theoretic alternative to cosine similarity is the Jensen-Shannon Divergence (JSD), a measure commonly used to assess similarity between probability distributions. JSD is a symmetrized and bounded score derived from KL-divergence:

$$\text{JSD}_{ij} = \frac{D_{\text{KL}}(u^{(i)}||M_{ij}) + D_{\text{KL}}(u^{(j)}||M_{ij})}{2} \tag{7}$$

$$M_{ij} = \frac{u^{(i)} + u^{(j)}}{2}$$

$$D_{\text{KL}}(r||q) = \sum_k r_k \log_2 \frac{r_k}{q_k}$$

Since all components of $v^{(i)}$ are positive, it is also possible to use JSD to compare TF-IDF vectors by setting $u^{(i)} = v^{(i)} / \sum_k v_k^{(i)}$ (interpreting the vectors as probability distributions over sets of objects). Both COSSIM and JSD are bounded between 0 and 1 (since all of our vectors are positive, cosine similarity is non-negative). Since JSD measures divergence rather than similarity, we set edge weights when computing the CC and clustering scores to $w_{ij} = 1 - \text{JSD}_{ij}$.

Figure 8 compares COSSIM and JSD in the simple case of two-outcome distributions. Cosine similarity takes more extreme values in this case, a pattern that we also see when computing the CC and clustering scores with both similarities: Cosine similarity tends to emphasize clustering over controversy.

5.3. Controversy measures: Regression-based controversy and evenly weighted indicators

In addition to similarity, the other important component of the scores we use is controversy, another concept that does not have a single objective measure. One method from prior work,

described in Section 3, is based on user tagging of controversies. Not every controversy is tagged, and so the method attempts to determine for every page how many revisions would have been tagged as controversial, using various features of a discussion to facilitate the learning problem. The weights on these features are learned using regression.

How dependent are our results on this methodology? A sensitivity analysis for the controversy score has two primary concerns. First, are our results sensitive to the particular weighting of controversy-relevant features that led to the page-level controversy score? The second concern is the distribution of controversy scores. Nearly any linear weighting of page-level controversy features (edits, protections, etc.) produces a distribution with exceptionally few very controversial pages, with most having negligible scores, but this does not necessarily mean that a page-level controversy score should mimic that distribution.

With this in mind, we compare the results with those obtained when we simply weight a small set of controversy indicators evenly. Under this alternate methodology, the controversy of a page (loosely following the article-level conflict model of Kittur et al. [2007]) is based on the number of revisions to an article's talk page, the fraction of minor edits on an article's talk page, mentions of "POV" in edit comments, and the number of times a page is "protected", where editing by new or anonymous users is limited. To address the distribution question, we employ these evenly-weighted features in two ways: with a simple non-linear transformation, and with only a linear transformation (as in Section 3, but with a different feature weighting).

For the non-linear transformation, we scale and shift each of the four quantities above such that their 5th and 95th percentiles are equal, then take the mean. Next, we transform this number such that the lowest values are at -5 and 1% of articles have scores above 0. Finally, the scores are transformed using the logistic function $1/(1+e^{-t})$. This produces a controversy score $c_k \in [0,1]$ for each page.

The particular weighting has a minor effect on which pages are designated as very controversial: highly controversial pages by one weighting tend to be controversial by others as well. For example, the average percentile of the controversy score for articles with six mentions of "POV" in edit comments is above 99, while a page with six mentions of "POV" but no protections or talk page edits is only in the 97^{th} percentile. This is an intuitive phenomenon: pages where content is repeatedly disputed ("POV" in edit comments) but none of the editors discuss the dispute (talk page edits) are very rare. Likewise for articles with three protections, or articles with 75 talk page edits, despite neither of these factors alone being sufficient for a 99^{th} percentile controversy score.

While the weighting makes little difference, the logistic transformation is quite impactful when considering behavior changes. Our results on detecting blocked users depend on the **rank** of a page's controversy score among other pages, and so are insensitive to monotonic transformations. However, the suspicious administrator behavior changes we have identified are from low- and medium-controversy pages to exceptionally high controversy pages (e.g. abortion, homeopathy), and this distinction can get lost if too many pages are grouped together at the high end of the page-level controversy score. For this reason, we adopt a version of the evenly-weighted controversy score which is simply scaled and shifted to be between 0 and 1 (referred to as the linearly-transformed evenly-weighted controversy score).

5.4. Analysis under changes in similarity and controversy

Our goal is a sensitivity analysis: how much do our conclusions about the behavior of administrators depend on the specific (reasonable) choices of similarity and controversy measure? We reiterate each of our main findings when using topic modeling with cosine similarity and the regression-based controversy metric, then examine how the claims hold up under alternative methodology.

To summarize these methodologies, we have choices between

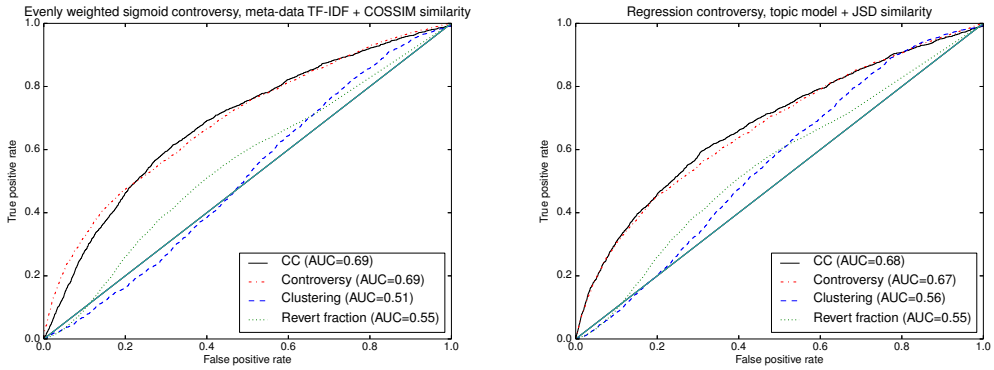

Fig. 9: Orthogonal measures of controversy and similarity nonetheless produce consistent results when differentiating manipulative blocked users from users who were never blocked.

(1) Topic modeling and metadata for page-level similarity features
(2) Jensen-Shannon divergence and cosine similarity for measuring similarity
(3) Regression-based controversy measurement and an even weighting, with or without a sigmoid transformation.

Any of the twelve combinations leads to its own version of Controversy, CC, and Clustering scores.

5.4.1. Finding manipulative users. Controversy and the CC-Score, as defined in Section 3, differentiate users who are blocked for manipulative behavior from those who are never blocked (see Section 4.1). How is this ability influenced by the choice of controversy score and the weighting of controversy within the CC-Score implied by different measures of similarity between pages?

We see little difference in predictivity between the different methodologies on this task. Figure 9 shows one example, with similar predictivity among two orthogonal methodologies. This indicates that there are consistent quantities underlying our concepts of similarity and controversy. We see a similar pattern across the other methodologies, with an AUC of just below 0.7 for the CC and Controversy Scores, and performance by Clustering around that of the fraction of a user's edits which are reverts, neither being much greater than random guessing.

The sigmoid transformation of the evenly weighted controversy scores does not significantly impact the manipulative user results, with both the Controversy and CC-Scores just below 0.7 with or without it. Since we are taking a weighted average of page-level controversy scores, this is not directly implied by the use of a monotonic transformation in a ranking task, but is nonetheless intuitive. The transformation does, however, impact our results on administrator behavior changes, described in the next sections.

5.4.2. Administrator behavior changes. We find in Section 4, using the regression-based controversy and topic modeling with cosine similarity described in Section 3, that users who actually become administrators change behavior in ways that users who unsuccessfully attempt to become administrators do not, even when they receive similar levels of support during the RfA process.

Our main results are qualitatively invariant to different similarity measures and different weightings of controversy features. In order to show this, we can compare the tails of the CC-score change distribution for various combinations of similarity measures and feature

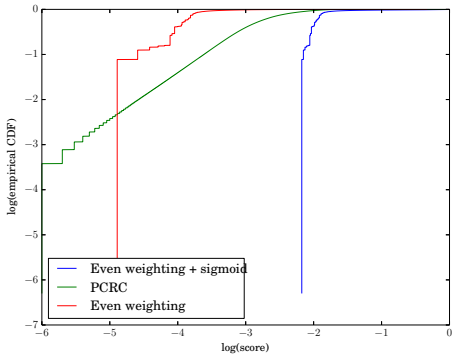

Fig. 10: The CDFs of two different controversy scoring methodologies, an even weighting of four features and a regression-based measure, along with a sigmoid transform of the even weighting. The linearly-transformed scores assign high values to relatively few pages, with most pages getting very low controversy scores.

weightings for the controversy score. For example, there are 146 successful administrators with CC-Score behavior changes above the 95^{th} percentile of changes for non-candidates (67 expected) according to the linearly-transformed evenly-weighted controversy score with topic modeling and cosine similarity for computing edge weights between pages, versus 119 using regression-based controversy. Under regression-based controversy, there are 129 above this threshold when using a TF-IDF weighting with Jensen-Shannon Divergence. In general there is a group of successful administrators who increase their CC-Scores post-RfA, while users who never attempt to become administrators decrease their CC-Scores over time. This pattern holds for the other ways of measuring similarity.

Similarity is an important aspect to consider. One natural question is, given that the results are invariant to quite different measures of similarity (metadata vs. natural language), whether similarity is contributing anything to the analysis, or if instead it is driven by controversy alone. To test this, we computed random edge weights for every pair of pages (uniform between 0 and 1) (essentially making the CC-Score a noisy version of the Controversy Score). Under this new score, users who never attempt to become administrators neither increase nor decrease their CC-Scores over time (95% confidence interval $[-0.013, 0.012]$) rather than decreasing them, and even those administrators who have the highest weighted-voter scores increase their CC-Scores significantly. Thus, it is in fact important to account for page similarity in the analysis.

Controversy scaling matters. While the results are qualitatively invariant across several natural ways of measuring similarity between pages, the same is not true for all of the controversy measures we tested. Results are consistent between different weightings of page features (i.e. regression-based and evenly-weighted controversy), but the sigmoid transformation leads to a CC-Score where administrators appear to be changing behavior very little. In fact, we see *fewer* successful administrators above the 95^{th} percentile of non-RfA behavior changes than we would expect if the distributions were identical. Administrators still have higher mean CC-Score changes, but the variance of their score changes is much smaller, and consequently there are no outlying changes.

This is due to compression at the high-end of the controversy score. Figure 10 shows the CDFs of the three scores: the two "natural" distributions, and the distribution of sigmoid-transformed scores. Regression-based controversy (i.e. the predicted controversial revision

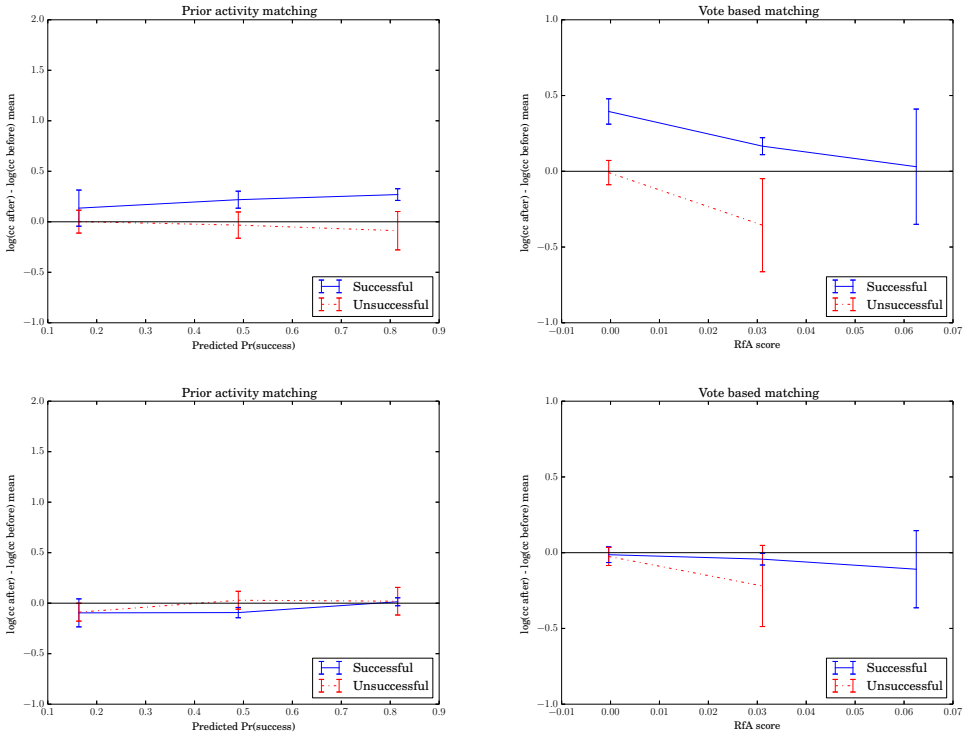

Fig. 11: Evenly weighted controversy results (top, shown here with topic model page features and cossine similarity) echo our earlier administrator behavior change findings using the regression-based controversy score. The evenly-weighted score with a sigmoid transformation (bottom), which marks significantly more pages as having high controversy, does not distinguish between successful and unsuccessful administrators. The lack of differentiation at the high end of the sigmoid-transformed controversy score "hides" behavior changes from somewhat controversial topics to very controversial topics. Plots show the CC-Score changes of matched groups of successful and unsuccessful candidates for administrator status, matched according to success-predicting editor characteristics (left) and the weighted voter model (right). Error bars show 95% confidence intervals.

count or pCRC), and to an even greater extent the linearly-transformed evenly weighted features, assign very low controversy scores to the vast majority of pages, reserving higher scores for a very small minority. Thus administrators do not change behavior by moving from obscure topics to somewhat controversial topics (which would be picked up by the sigmoid-transformed score), but some do change behavior by moving from topics of middling controversy to Wikipedia's most contentious issues.

Figure 11 illustrates the replication of our matched sample results with the evenly weighted linearly-transformed controversy score (top two plots) and the sigmoid-transformed version (bottom two plots). The top left plot shows that, as with regression-based controversy, candidates who are similarly situated before their RfA show quite different behavior after. As the only difference between these groups is the new social and technical position afforded one but not the other, users seem to change behavior as a result of becoming administrators. The sigmoid-transformed controversy score masks these changes.

5.4.3. Predicting behavior changes from RfAs. When using cosine similarity and topic modeling, we show in Section 3 that it is possible to find candidates who do not change behavior in suspicious ways upon becoming administrators, but that predictors of RfA success and simple vote aggregation are not sufficient. A more sophisticated vote aggregation method that reweights the voters does find such candidates.

The plots in Figure 11 show behavior changes arranged by the two RfA scoring methods we considered. On the left is the predicted probability of success based on visible features of an editor, e.g. the number of edits or time spent as an editor pre-RfA. On the right is the weighted-voter score, which favors voters who adhered to what is inferred to be the "correct" outcome in other votes. The evenly-weighted controversy score parallels our results with the pCRC: surface-level features of an editor do not discriminate between those who do and do not go on to change behavior post-RfA. However, there is information in the RfA process. Using the weighted-voter score, Figure 11 right, we see the upper half of successful candidates in terms of the weighted-voter score increases their CC-Score significantly less than the lower half (various p-values, but consistently less than 0.01). Depending on the choice of controversy and similarity measure, this higher-scoring half is nonetheless occasionally increasing their CC-Score.

The bottom half of Figure 11 shows the equivalent plots for the sigmoid-transformed controversy measure. As before, this score loses the differentiation between behavior changes of successful and unsuccessful candidates. Despite this, we do see the upper half of successful candidates according to the weighted-voter score changing behavior less than their lower-scoring (but still successful) counterparts.

6. DISCUSSION

Is the crowd really wise, and can we depend on it for reliable information? This question has become increasingly important in an era where it is easy to both find and contribute new information. For example, there has been significant research on judging the correctness of prediction markets as predictors of future events [Wolfers and Zitzewitz 2004], and on understanding the incentive-compatibility properties of these markets when used for different purposes (for example, when a stakeholder makes decisions based on the outcomes of contingent markets [Hanson 2002]). Researchers have also focused attention on websites that rely heavily on consumer ratings, ranging from Amazon to TripAdvisor and Yelp. A Scientific American story from 2010 says "The philosophy behind this so-called crowdsourcing strategy holds that the truest and most accurate evaluations will come from aggregating the opinions of a large and diverse group of people. Yet a closer look reveals that the wisdom of crowds may neither be wise nor necessarily made by a crowd. Its judgments are inaccurate at best, fraudulent at worst" [Moyer 2010]. That story focuses on the biases that may effect online rating systems, including selection effects, timing issues, and deliberate manipulation. There has been academic research both on uncovering the types of bias and manipulation that may impact recommender systems as well as on designing robust recommender systems [Resnick and Sami 2007].

Online encyclopedias like Wikipedia raise a related but different set of challenges. It is harder to quantify manipulation, since the actions taken by participants span a much broader range of possibilities. Further, individual users can have outsize effects on the content of an article. In this paper, we take the first steps towards putting the study of manipulation of online content-aggregation systems like Wikipedia on a sound analytical footing. We describe a methodology for computing a score based on a user's editing history that measures how focused they are on a controversial topical theme. We can use changes in this measure to detect suspicious behavior, particularly around the time of promotion to administrator status.

In doing so, we discover several interesting facts about the Wikipedia ecosystem. There is evidence for the existence of manipulation. This could be intentional manipulation, with

someone trying to infiltrate the admin cadre, or it could be largely in good faith, but nevertheless worth monitoring because of the potential for a good-faith administrator's intrinsic or unconscious biases to become the dominant factor in the viewpoint reflected on a page. On the positive side, we find that the election process already reveals the information necessary to filter out potential manipulators. Some particularly good voters are the ones who are doing a good job of filtering out potential manipulators in the promotion process: neither quantitative measures of prior behavior, nor simple vote counts are as discriminative in identifying potential manipulators as is a measure that takes into account how influential different voters who participate in a particular editor's promotion decision are.

ACKNOWLEDGMENTS

This research was supported in part by an NSF CAREER Award (grants 1303350 and 1414452) to Das. Magdon-Ismail was sponsored in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

Sinan Aral, Lev Muchnik, and Arun Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* 106, 51 (2009), 21544–21549.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

Moira Burke and Robert Kraut. 2008. Mopping up: Modeling Wikipedia promotion decisions. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. 27–36.

2008. Candid CAMERA. *Harper's Magazine* (July 2008).

Sanmay Das, Allen Lavoie, and Malik Magdon-Ismail. 2013. Manipulation Among the Arbiters of Collective Intelligence: How Wikipedia Administrators Mold Public Opinion. In *Proceedings of the Twenty-Second ACM Conference of Information and Knowledge Management (CIKM '13)*. 1097–1106.

Sanmay Das and Malik Magdon-Ismail. 2010. Collective Wisdom: Information Growth in Wikis and Blogs. In *Proceedings of the ACM Conference on Electronic Commerce*. 231–240.

Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78, 1 (2010), 35–71.

Arpita Ghosh, Satyen Kale, and Preston McAfee. 2011. Who moderates the moderators? Crowdsourcing abuse detection in user-generated content. In *Proceedings of the ACM Conference on Electronic Commerce*. ACM, New York, NY, USA, 167–176. DOI:http://dx.doi.org/10.1145/1993574.1993599

Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature* 438, 7070 (December 2005), 900–901.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101 (6 Apr 2004), 5228–5235. DOI:http://dx.doi.org/10.1073/pnas.0307752101

R. Hanson. 2002. Decision markets. *Entrepreneurial Economics: Bright Ideas from the Dismal Science* (2002), 79.

M. Hindman, K. Tsioutsiouliklis, and J.A. Johnson. 2003. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *Annual Meeting of the Midwest Political*

Science Association, Vol. 4. 1–33.

Gabriela Kalna and Desmond J. Higham. 2007. A clustering coefficient for weighted networks, with application to gene expression data. *AI Communications* 20 (Dec 2007), 263–271. Issue 4. http://dl.acm.org/citation.cfm?id=1365534.1365536

Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: Conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

D. Lazer, A.S. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. 2009. Life in the network: the coming age of computational social science. *Science* 323, 5915 (2009), 721.

Chenliang Li, Anwitaman Datta, and Aixin Sun. 2011. Mining latent relations in peer-production environments: A case study with Wikipedia article similarity and controversy. *Social Network Analysis and Mining* (2011), 1–14.

Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, and Maosong Sun. 2011. PLDA+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing. *ACM Transactions on Intelligent Systems and Technology, special issue on Large Scale Machine Learning* (2011).

Rui Lopes and Luis Carriço. 2008. On the credibility of wikipedia: an accessibility perspective. In *Proceedings of the 2nd ACM workshop on Information credibility on the web (WICOW '08)*. ACM, New York, NY, USA, 27–34. http://doi.acm.org/10.1145/1458527.1458536

F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. 2006. Googlearchy or Googlocracy? *IEEE Spectrum Online* (2006).

M. Moyer. 2010. Manipulation of the Crowd. *Scientific American Magazine* 303, 1 (2010), 26–28.

Martin Potthast, Benno Stein, and Robert Gerling. 2008. Automatic vandalism detection in Wikipedia. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval (ECIR'08)*. Springer, Berlin, Heidelberg, 663–668. http://dl.acm.org/citation.cfm?id=1793274.1793363

P. Resnick and R. Sami. 2007. The influence limiter: Provably manipulation-resistant recommender systems. In *Proceedings of the ACM Conference on Recommender Systems*. ACM, 25–32.

Alastair Sloan. 2015. Manipulating Wikipedia to Promote a Bogus Business School. *Newsweek* (March 24 2015).

Koen Smets, Bart Goethals, and Brigitte Verdonk. 2008. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence*.

Anselm Spoerri. 2007. What is Popular on Wikipedia and Why? *First Monday* 12, 4 (April 2007).

Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, and Hady Wirawan Lauw. 2008. On ranking controversies in Wikipedia: Models and evaluation. In *Proceedings of the International Conference on Web Search and Web Data Mining*. 171–182.

Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. 2011. Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference (iConference '11)*. ACM, New York, NY, USA, 122–129. DOI:http://dx.doi.org/10.1145/1940761.1940778

Dennis M. Wilkinson and Bernardo A. Huberman. 2007. Assessing the Value of Coooperation in Wikipedia. *First Monday* 12, 4 (Feb 2007).

Justin Wolfers and Eric Zitzewitz. 2004. Prediction Markets. *Journal of Economic Perspectives* 18, 2 (2004), 107–126.