Manipulation Among the Arbiters of Collective Intelligence: How Wikipedia Administrators Mold Public Opinion

Sanmay Das*
Virginia Tech
sanmay@cs.vt.edu

Allen Lavoie*
Virginia Tech
allenbl@cs.vt.edu

Malik Magdon-Ismail
Rensselaer Polytechnic Institute
magdon@cs.rpi.edu

ABSTRACT

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. We find a surprisingly large number of editors who 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. Neither prior history nor vote counts during an administrator's election can identify those editors most likely to change their behavior in this suspicious manner. We find that an alternative measure, which gives more weight to influential voters, can successfully reject these suspicious candidates. This 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.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]; I.5.4 [**Pattern Recognition**]: Applications; K.4.1 [**Computers and Society**]: Public Policy Issues

Keywords

Social networks; Wikipedia; manipulation

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 [13], 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. Another major target for manipulation could be Wikipedia, which crowdsources the collection of knowledge to millions of editors, and is generally regarded as high-quality [6]. Thousands of these 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 perceptions of 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.[1]

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 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. This 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.

*Now at Washington University in St. Louis

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2505566.

[1] http://en.wikipedia.org/wiki/Wikipedia:Requests_for_comment/Paid_Editing

We then use this method to analyze the behavior of editors who successfully become administrators. We find that a significantly higher than expected fraction of users who are successful in receiving promotion to administrator status increase their CC-scores a large amount shortly after election, indicating that they are then exerting significantly more control over the portrayal of controversial topics on Wikipedia, and doing so in a topically clustered way. We would expect them to use their new powers broadly across controversial topics—administrators are expected to intervene in disputes—but a change in behavior toward editing on topically clustered controversial articles is surprising. 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.

Is it possible to identify these potentially manipulative administrators by their behavior *before* the RfA? Two intuitive tests fail. (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-history score). The prior-history score attempts to measure "admin-like" behavior on Wikipedia prior to an RfA, such as participation in maintenance tasks and dispute moderation. The prior-history 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 [5] (we refer to this as the voter-based score). This measure weights voters differently by taking into account how influential different voters who actually participated in this particular editor's RfA are, and which way they voted. Editors with very high voter-based 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 types of behavior that the editor has previously engaged in. 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.

1.1 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 [6] and very popular on the Web [22]. The dynamics of how articles become high quality and how information grows in collective media like Wikipedia have also garnered some attention [25, 4]. 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.* discuss how this can lead to a few highly ranked sites coming to dominate political discussion on the Web [9]. 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.* [17].

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.* introduce new tools for studying conflict and coordination costs in Wikipedia [12]. Vuong *et al.* characterize controversial pages using both disputes on a page and the relationships between articles and contributors [23]. 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 *et al.* [16] find that accessibility issues significantly impede this process. Welser *et al.* [24] 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.* [19] use a small number of features in a logistic regression model to detect vandalism. Smets *et al.* 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 [21]. 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 [3]. 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.

2. 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.

2.1 Controversy Score

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).

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). 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 (see 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.

2.2 Clustered Controversy Score

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 [11] to quantify how coherent an administrator's controversial edits are. 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.

2.2.1 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) [2]—on the text of Wikipedia pages. We use a procedure similar to Griffiths and Steyvers (2004); see Appendix A for details. We model articles as containing a mix of 1000 topics, which allow fine-grained comparisons while mostly 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[2]. Thus we make abstract comparisons between articles based on topics rather than concrete words or structural features.

It is worth noting that alternative approaches can be applied to the problem of assessing page similarity, especially in the context of Wikipedia. Wikipedia articles specifically have editors, categories, and links which can be used to derive a measure of similarity. While these attributes are high-dimensional, and therefore comparisons based on them may be subject to the curse of dimensionality, there are several methods for transforming metadata such as links into similarity scores while avoiding high-dimensional comparisons. We implemented a comparison methodology based on page metadata, and found that our text-based comparisons produced very similar results. Therefore, we present results based on the text, since text data is more widely available in other potential applications than rich and accurate metadata. See Appendix B for more information.

2.2.2 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 [11]. 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:

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

Then the clustering score of a page is:

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

$\text{clust}(k)$ 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[3]—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:

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

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 (say

[2]Alternatively, since we are comparing distributions, Jensen-Shannon Divergence could be employed. In limited experiments, we did not observe any qualitative changes in results when using different similarity metrics.

[3]Including the controversy and edit fraction of connected nodes, as we do through a page's impact $\imath(\cdot)$, 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.

Flamingos), 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 thread as a measure of controversy. These quantities can be entirely language independent, for example replacing text with a contributor-based similarity model [14].

2.3 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.

2.4 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 compare these measures to the percentage of support votes a candidate receives during an RfA. This support percentage determines the outcome of an RfA in practice (AUC 0.998). The distributions of all three scores are shown in Figure 1 for successful and unsuccessful candidates.

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 manages to predict success fairly well, with an AUC of 0.82.

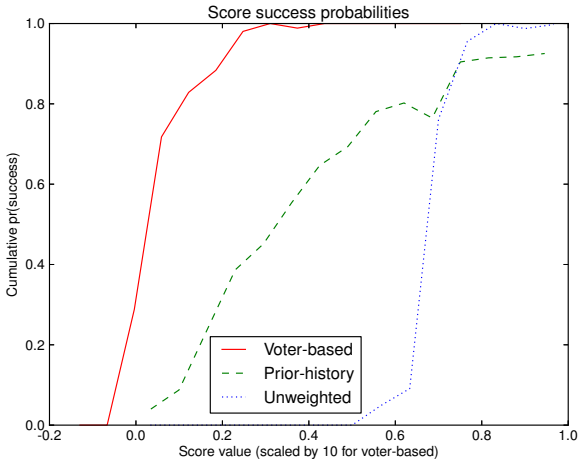

Figure 2: Probability of a successful RfA as a function of the weighted voter-based score, the prior-history score, and the unweighted vote fraction. The voter-based score is multiplied by 10 to show detail.

Voter model.

The scoring of RfAs—determining which are successful—can be contentious. Wikipedia typically eschews decisive voting in favor of consensus building, attempting to downplay the role of vote percentages in RfAs and similar decisions. 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; despite this, vote percentage is used in practice. While it is difficult to quantify consensus directly, voter history helps: by inferring the quality of voters, we can take a more principled approach to scoring RfAs.

Ghosh et al. (2011) recently introduced a technique for aggregating noisy votes in abuse detection for user-generated content. For example, 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 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 achieve good performance. We assume that the judgments of the bureaucrats who determine which RfAs are successful satisfy this requirement, and so use the outcome (success or failure) of RfAs as our "more likely than not" signal. While the algorithm implicitly determines the trustworthiness of each voter, it explicitly assigns each RfA a quality score; we use this score directly.

This method provides fresh insight into the outcomes of RfAs on Wikipedia. Figure 2 compares the distribution of success probabilities associated with the voter-based score with that of the prior-history score and raw vote fraction. The voter-based score is quite predictive of RfA success, achieving 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. Raw vote percentage is more discriminative, but despite this we show later that unweighted

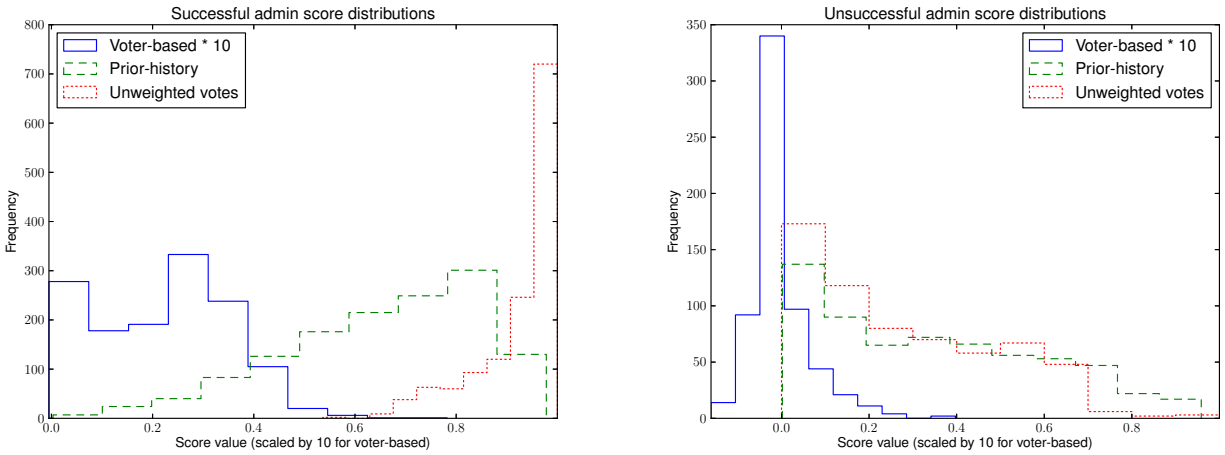

Figure 1: Distributions of the three RfA or pre-RfA scores for admin candidates. Successful candidates are shown on the left, unsuccessful on the right. The voter-based score is multiplied by a factor of 10 to show detail.

votes behave more like the predicted probability of success based on prior history described above.

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. Further, since many editors with borderline voter-based and prior activity scores did 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 in these "contentious" categories.

3. RESULTS

We evaluate our metrics in several different ways. First, to establish their validity, we examine whether the metrics 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.

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. Therefore, later in this section, we examine this hypothesis by comparing the distribution of behavior changes among administrators with those of similar groups who did not become administrators.

3.1 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.

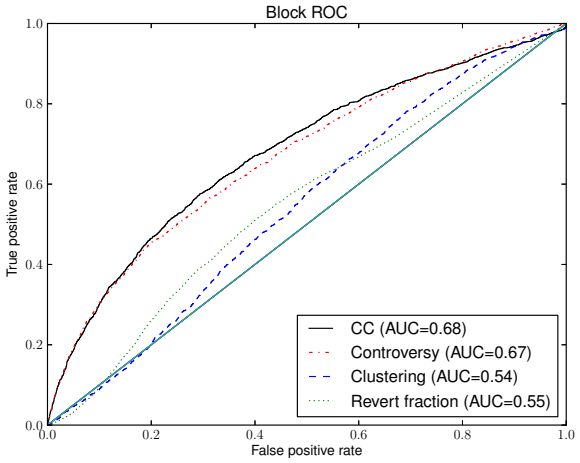

Figure 3: ROC curve for CC, Controversy, and Clustering Scores when differentiating between blocked and not-blocked users, based on 180 days of data. As a baseline, the fraction of a user's edits during this period which were reverts is also included. The CC and Controversy Scores effectively discriminate 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.

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	1.6%
Zango (company)	2.5%	controlled demolition	
Wi-Fi	2.1%	conspiracy theories	
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%	Safe sex	1.1%
Association football	3.3%	Condom	0.8%
Philosophy	3.0%	Hippie	0.7%
Irony	2.7%	Fox News Channel	0.7%
Lysergic acid diethylamide	1.9%	Planned Parenthood	0.6%
		The Beatles	0.5%
Hippie	1.3%	Masturbation	0.5%
Bill O'Reilly (political commentator)	1.3%	Lysergic acid diethylamide	0.4%
Iraq War	1.2%		

Table 1: 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.

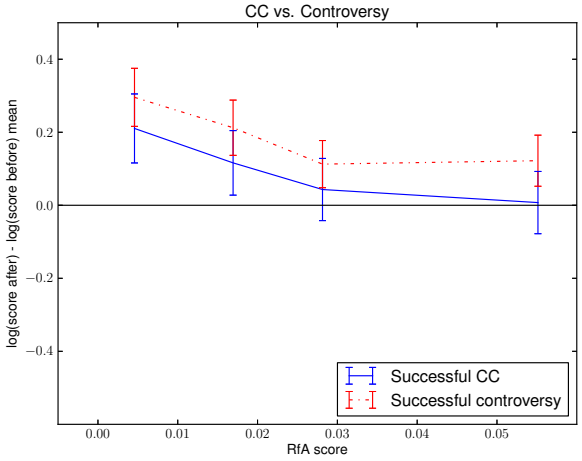

Figure 4: Behavior changes upon becoming an administrator, measured by the CC- and C-Scores for 180 days of edits before and after a successful Request for Adminship (RfA). The x axis is the vote-based RfA score, with a higher score implying a stronger consensus. The Controversy Score increases on average for both low and high scoring administrators, while only low scoring administrators increase their CC-Score.

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 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.

3.2 Administrator behavior changes

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 1 indicate that suspicious behavior changes do exist, and that the CC-Score may be useful in finding them. Figure 5 gives an anecdotal overview of the types of administrators with very high and very low CC-Scores.

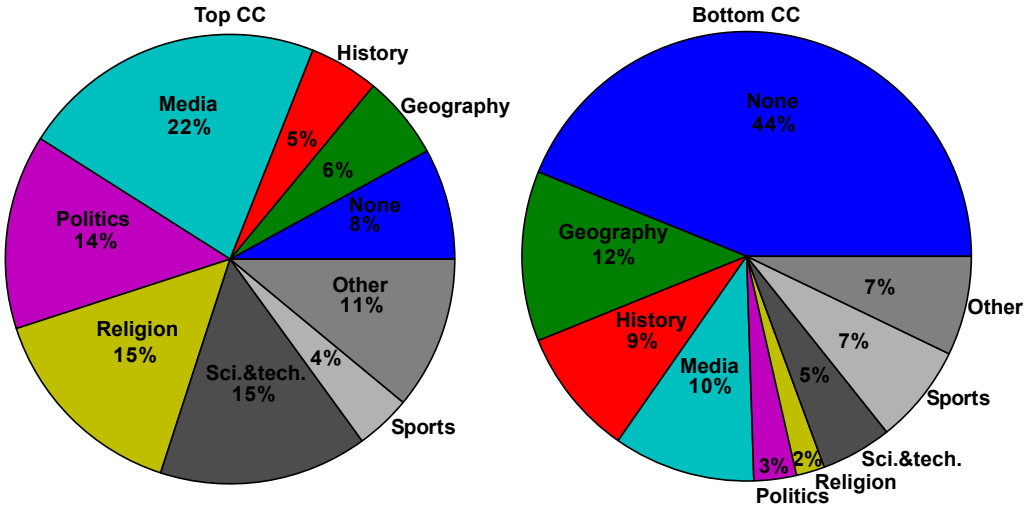

Figure 5: Blind human evaluation of the general category of edits (if any) for administrators directly after their RfA. The 100 highest and 100 lowest scoring administrators according to a previous version of the CC-Score are shown (using metadata page comparisons and a slightly different controversy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

Our analysis focuses on three groups of Wikipedia users: those who actually become administrators, those who try unsuccessfully to become administrators, and 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.

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[4] comparing the 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. Users who have never participated in an RfA serve as a baseline for what constitutes normal behavior changes over time.

3.2.1 *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.

[4]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.

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 above the 95th percentile of the distribution of users who never tried to become an administrator, while we would expect 67.5. These 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.

3.2.2 *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 voter-based 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 voter-based measure. The idea here, as in propensity

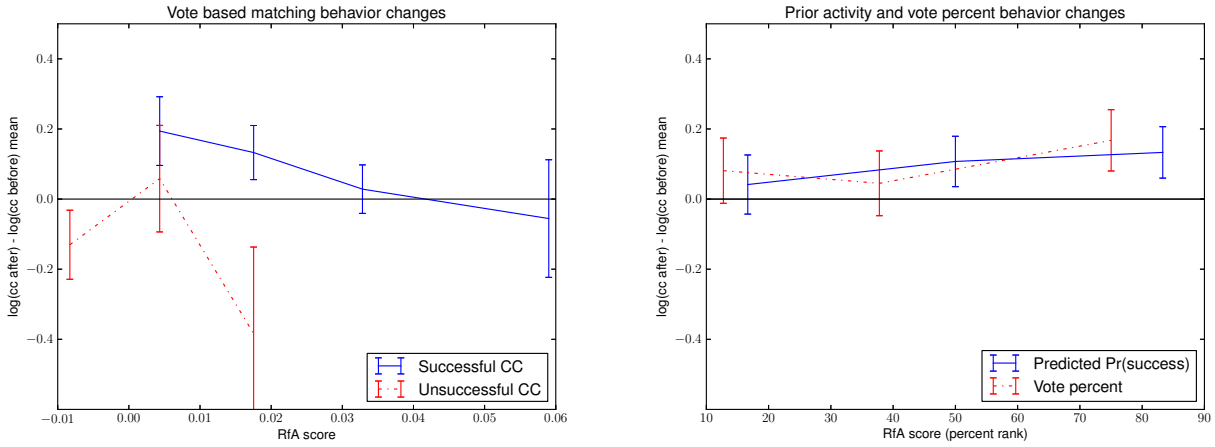

Figure 6: The vote-based score of a Request for Adminship (RfA) (left) discriminates between administrators who change their behavior significantly and those who do not; a small group with low vote-based scores skew the average for successful administrators. The activity-based score (right) does not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

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.

3.2.3 *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 voter-based scores are much less likely to exhibit large changes in their CC scores than those with moderate voter-based scores. The same is not true of simpler measures like raw vote count or the prior-history model.

First, the voter-based results. We divide administrators into groups on the basis of their voter-based scores, and find that the C-Score rises significantly after election for each group (Figure 4). 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 voter-based scores show a statistically significant increase in their CC-Score after a successful RfA, whereas administrators with very high voter-based scores show no such increase (Figure 4). For example, consider editors who succeed in their RfAs with a voter-based 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 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 voter-based 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-history 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-based model is (Figure 6, right). When we group by the prior-history 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 voter-based score and changes in the CC-Score (lower scorers change behavior more), where we find no such relationship when considering the unweighted or prior-history 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 voter-based 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 voter-based 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.

4. 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 [26], 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 [8]). 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" [18]. 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 [20].

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.

5. ACKNOWLEDGEMENTS

This research is supported in part by an NSF CAREER Award (IIS-1303350) 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

[1] Candid CAMERA. *Harper's Magazine*, July 2008.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, Mar. 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

[3] M. Burke and R. Kraut. Mopping up: Modeling Wikipedia promotion decisions. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 27–36, 2008.

[4] S. Das and M. Magdon-Ismail. Collective wisdom: Information growth in wikis and blogs. In *Proceedings of the ACM Conference on Electronic Commerce*, pages 231–240, 2010.

[5] A. Ghosh, S. Kale, and P. McAfee. Who moderates the moderators? Crowdsourcing abuse detection in user-generated content. In *Proceedings of the ACM Conference on Electronic Commerce*, pages 167–176, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0261-6. doi: 10.1145/1993574.1993599. URL http://doi.acm.org/10.1145/1993574.1993599.

[6] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005. ISSN 0028-0836.

[7] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, Apr 2004. ISSN 0027-8424. doi: 10.1073/pnas.0307752101. URL http://dx.doi.org/10.1073/pnas.0307752101.

[8] R. Hanson. Decision markets. *Entrepreneurial Economics: Bright Ideas from the Dismal Science*, page 79, 2002.

[9] M. Hindman, K. Tsioutsiouliklis, and J. Johnson. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *Annual Meeting of the Midwest Political Science Association*, volume 4, pages 1–33, 2003.

[10] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. ACM. ISBN 1-58113-567-X. doi: http://doi.acm.org/10.1145/775047.775126. URL http://doi.acm.org/10.1145/775047.775126.

[11] G. Kalna and D. J. Higham. A clustering coefficient for weighted networks, with application to gene expression data. *AI Communications*, 20: 263–271, Dec 2007. ISSN 0921-7126. URL http://dl.acm.org/citation.cfm?id=1365534.1365536.

[12] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: Conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007.

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

[14] C. Li, A. Datta, and A. Sun. Mining latent relations in peer-production environments: A case study with Wikipedia article similarity and controversy. *Social Network Analysis and Mining*, pages 1–14, 2011.

[15] Z. Liu, Y. Zhang, E. Y. Chang, and M. Sun. 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.

[16] R. Lopes and L. Carriço. On the credibility of Wikipedia: an accessibility perspective. In *Proceedings of the 2nd ACM workshop on Information credibility on the web*, WICOW '08, pages 27–34, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-259-7. URL http://doi.acm.org/10.1145/1458527.1458536.

[17] F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Googlearchy or Googlocracy? *IEEE Spectrum Online*, 2006.

[18] M. Moyer. Manipulation of the crowd. *Scientific American Magazine*, 303(1):26–28, 2010.

[19] M. Potthast, B. Stein, and R. Gerling. Automatic vandalism detection in Wikipedia. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 663–668, Berlin, Heidelberg, 2008. Springer. ISBN 3-540-78645-7, 978-3-540-78645-0. URL http://dl.acm.org/citation.cfm?id=1793274.1793363.

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

[21] K. Smets, B. Goethals, and B. Verdonk. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence*, 2008.

[22] A. Spoerri. What is popular on Wikipedia and why? *First Monday*, 12(4), April 2007.

[23] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in Wikipedia: Models and evaluation. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 171–182, 2008.

[24] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0121-3. doi: 10.1145/1940761.1940778. URL http://doi.acm.org/10.1145/1940761.1940778.

[25] D. M. Wilkinson and B. A. Huberman. Assessing the value of coooperation in Wikipedia. *First Monday*, 12(4), Feb 2007.

[26] J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

A. ARTICLE TOPIC MODELS

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 [15] 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 [15]). The log-likelihood converges well before this point.

B. PAGE SIMILARITY

There are many options for comparing Wikipedia pages. Meta-data such as links, users, and categories provide rich sources of information. In addition, it is possible to build specialized measures on top of these page features, such as SimRank[10] using links. The curse of dimensionality is an issue when making raw high-dimensional comparisons based on meta-data, and scaling derived measures such as SimRank can be problematic (requiring an approximately 4000000×4000000 matrix in this case). However, the major reason we do not use specialized Wikipedia-specific comparisons (although we have done some experiments with them) is generalization to other collective intelligence venues, including relatively unstructured environments such as email.

Instead, we focus on text-based comparisons. Although TF-IDF and similar weighting schemes can be effective, they still lead to very high-dimensional comparisons when considering whole documents. Instead, we pre-process the text with a topic model (see Appendix A) into a manageable number of topics, and base our comparisons on those topics. This type of comparison is useful wherever documents contain text, is scalable (as evidenced by our inference on all of English Wikipedia), and allows for lower-dimensional comparisons between pages without explicitly storing every similarity score (our methods only look at a small fraction of these scores which are relevant to a user's local edit graph, and storing all of the scores can be intractable). In our experiments, the choice of similarity measure (for example, the Jaccard coefficient comparing categories, users, and links) has not led to any qualitative changes in results.